METhodological RadiomICs Score (METRICS): A quality scoring tool for radiomics research

B Kocak, T Akinci d'Antonoli, N Mercaldo, A Alberich-Bavarri, B Baessler, I Ambrosini, A Andreychenko, S Bakas, R Beets-Tan, K Bressem, et al.

METhodological RadiomICs Score (METRICS): A quality scoring tool for radiomics research

Outline

METhodological RadiomICs Score (METRICS): A quality scoring tool for radiomics research endorsed by EuSoMII

Title Page

Manuscript type: Guideline

Order	First name	Surname	ORCID	Affiliations
1*	Burak	Kocak	0000-0002-7307-396X	Department of Radiology, University of Health Sciences, Basaksehir Cam and Sakura City Hospital, Basaksehir, Istanbul, Turkey
1*	Tugba	Akinci D'Antonoli	0000-0002-7237-711X	Institute of Radiology and Nuclear Medicine, Cantonal Hospital Baselland, Liestal, Switzerland
2	Nathaniel	Mercaldo	0000-0003-1658-6598	Department of Radiology, Massachusetts General Hospital, Boston, MA, USA.
3	Angel	Alberich-Bayarri	0000-0002-5932-2392	Quantitative Imaging Biomarkers in Medicine (Quibim), Valencia, Spain
4	Bettina	Baessler	0000-0002-3244-3864	University Hospital Würzburg, Department of Diagnostic and Interventional Radiology, Oberdürrbacher Str. 6, 97080 Würzburg, Germany
5	Ilaria	Ambrosini	0000-0002-0026-9101	Department of Translational Research, Academic Radiology, University of Pisa, Italy
6	Anna	Andreychenko	0000-0001-6359-0763	K-SkAI LLC, Petrozavodsk, Russia; ITMO University, St. Petersburg, Russia

7	Spyridon	Bakas	0000-0001-8734-6482	1. Center for Artificial Intelligence for Integrated Diagnostics (AI2D) and Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA 2. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA 3. Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
8	Regina	Beets-Tan	0000-0002-8533-5090	The Netherlands Cancer Institute, Department of Radiology, Amsterdam, the Netherlands.
9	Keno	Bressem	0000-0001-9249-8624	1. Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt- Universität zu Berlin, Department of Radiology, Hindenburgdamm 30, 12203 Berlin, Germany 2. Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany.
10	Irene	Buvat	0000-0002-7053-6471	Institut Curie, Inserm, PSL University, Laboratory of Translational Imaging in Oncology, Orsay, France
11	Roberto	Cannella	0000-0002-3808-0785	Section of Radiology - Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, Italy
12	Luca Alessandro	Cappellini	0000-0001-7604-5625	Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

13	Armando Ugo	Cavallo	0000-0001-8390-7721	Division of Radiology, Istituto Dermopatico dell'Immacolata (IDI) IRCCS, Rome, Italy
14	Leonid L	Chepelev	0000-0001-7010-3812	Joint Department of Medical Imaging, University Health Network, University of Toronto, Toronto, Canada
15	Linda Chi Hang	Chu	0000-0001-9729-2756	The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD 2128, USA
16	Aydin	Demircioglu	0000-0003-0349-5590	Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Germany
17	Nandita M	deSouza	0000 0003 4232 476X	1. Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, United Kingdom. 2. Department of Imaging, The Royal Marsden National Health Service (NHS) Foundation Trust, London, United Kingdom.
18	Matthias	Dietzel	0000-0001-9248-1398	Department of Radiology, University Hospital Erlangen, Maximiliansplatz 3, 91054, Erlangen, Germany.
19	Salvatore Claudio	Fanni	0000-0002-4003-3320	Department of Translational Research, Academic Radiology, University of Pisa, Italy
20	Andrey	Fedorov	0000-0003-4806-9413	Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

21	Laure S	Fournier	0000-0002-1878-0290	Université Paris Cité, AP-HP, Hôpital Européen Georges Pompidou, Department of Radiology, PARCC UMRS 970, INSERM, Paris, France
22	Valentina	Giannini	0000-0001-5052-8231	Department of Surgical Sciences, University of Turin, Turin 10126, Italy.
23	Rossano	Girometti	0000-0002-0904-5147	Institute of Radiology, Department of Medicine, University of Udine, University Hospital S. Maria della Misericordia - p.le S. Maria della Misericordia, 15 - 33100 Udine, Italy
24	Kevin B.W.	Groot Lipman	0000-0003-3651-2529	1. Netherlands Cancer Institute, Department of Radiology, Amsterdam, The Netherlands 2. Netherlands Cancer Institute, Department of Thoracic Oncology, Amsterdam, The Netherlands 3. GROW School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands.
25	Georgios	Kalarakis	0000-0002-5333-5993	1. Department of Neuroradiology, Karolinska University Hospital, 14152 Stockholm, Sweden. 2. Department of Clinical Science, Division of Radiology, Intervention and Technology (CLINTEC), Karolinska Institutet, 14152 Stockholm, Sweden. 3. Department of Radiology, Medical School, University of Crete, 71500 Heraklion, Greece.

26	Brendan	Kelly	0000-0002-3449-8017	1. Department of Radiology, St Vincent's University Hospital, Dublin, Ireland. 2. Insight Centre for Data Analytics, UCD, Dublin, Ireland. 3. School of Medicine, University College Dublin, Dublin, Ireland.
27	Michail E.	Klontzas	0000-0003-2731-933X	1. Department of Medical Imaging, University Hospital of Heraklion, Crete, Greece 2. Department of Radiology, University of Crete, Heraklion, Crete, Greece 3. Computational Biomedicine Laboratory, Institute of Computer Science, FORTH, Heraklion, Crete, Greece
28	Dow-Mu	Koh	0000-0001-7654-8011	Department of Radiology, Royal Marsden Hospital, Sutton, UK.
29	Elmar	Kotter	0000-0001-9022-6000	Department of Diagnostic and Interventional Radiology, Faculty of Medicine, Medical Center-University of Freiburg, Hugstetter Str. 55, 79106, Freiburg, Germany
30	Ho Yun	Lee	0000-0001-9960-5648	Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea.
31	Mario	Maas	0000-0001-6785-5167	Department of Radiology & Nuclear Medicine, Amsterdam UMC Location University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands.
32	Luis	Marti-Bonmati	0000-0002-8234-010X	Medical Imaging Department and Biomedical Imaging Research Group, Hospital Universitario y Politécnico La Fe and Health Research Institute, Valencia, Spain

33	Henning	Müller	0000-0001-6800-9878	1. University of Applied Sciences of Western Switzerland (HES-SO Valais), Valais, Switzerland. 2. Department of Radiology and Medical Informatics, University of Geneva (UniGe), Geneva, Switzerland.
34	Nancy	Obuchowski	0000-0003-1891-7477	Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA
35	Fanny	Orlhac	0000-0002-5588-1867	Institut Curie, Inserm, PSL University, Laboratory of Translational Imaging in Oncology, Orsay, France
36	Nikolaos	Papanikolaou	0000-0003-3298-2072	1. Computational Clinical Imaging Group, Centre for the Unknown, Champalimaud Foundation, Av. Brasília, Doca de Pedrouços, 1400-038, Lisbon, Portugal. 2. Department of Radiology, Royal Marsden Hospital and The Institute of Cancer Research, London, SM2 5NG, UK.
37	Ekaterina	Petrash	0000-0001-6572-5369	1. Radiology department, Research Institute of Pediatric Oncology and Hematology n. a. L.A. Durnov, National Medical Research Center of Oncology n. a. N.N. Blokhin Ministry of Health of Russian Federation, Moscow, Russia 2. Medical Department IRA-Labs, Moscow, Russia
38	Elisabeth	Pfaehler	0000-0002-6061-3011	Institute for advanced simulation: Machine learning and data analytics, Forschungszentrum Jülich, Jülich, Germany

39	Daniel	Pinto dos Santos	0000-0003-4785-6394	1. Department of Radiology, University Hospital of Cologne, Cologne, Germany 2. Institute for Diagnostic and Interventional Radiology, Goethe-University Frankfurt Am Main, Frankfurt, Germany
40	Andrea	Ponsiglione	0000-0002-0105-935X	University of Naples Federico II, Department of Advanced Biomedical Sciences, Naples, Italy
41	Sebastià	Sabater	0000-0003-0111-9540	Department of Radiation Oncology. Complejo Hospitalario Universitario de Albacete. Albacete. Spain
42	Francesco	Sardanelli	0000-0001-6545-9427	1. Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milan, Italy 2. Unit of Radiology, IRCCS Policlinico San Donato, San Donato Milanese, Milan, Italy
43	Philipp	Seeböck	0000-0001-5512-5810	Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria
44	Nanna M	Sijtsema	0000-0001-6644-274X	Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
45	Arnaldo	Stanzione	0000-0002-7905-5789	University of Naples Federico II, Department of Advanced Biomedical Sciences, Naples, Italy

46	Alberto	Traverso	0000-0001-6183-4429	Department of Radiotherapy, Maastro Clinic, Maastricht, The Netherlands. School of Medicine, Vita-Salute San Raffaele University,Milan, Italy
47	Lorenzo	Ugga	0000-0001-7811-4612	University of Naples Federico II, Department of Advanced Biomedical Sciences, Naples, Italy
48	Martin	Vallières	0000-0001-7639-8172	1. Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada 2. Centre de recherche du Centre hospitalier universitaire de Sherbrooke, Sherbrooke, Canada
49	Lisanne V	van Dijk	0000-0002-9515-5616	Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, NL
50	Joost J M	van Griethuysen	0000-0003-0447-0918	Department of radiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
51	Robbert W	van Hamersvelt	0000-0002-6084-0656	Department of Radiology, University Medical Center Utrecht, Utrecht University,Utrecht, the Netherlands.
52	Peter	van Ooijen	0000-0002-8995-1210	University of Groningen, University Medical Center Groningen, Dept. of Radiotherapy, Groningen, The Netherlands
53	Federica	Vernuccio	0000-0003-0350-1794	Department of Radiology, University Hospital of Padova, Padova, Italy

54	Alan	Wang	0000-0003-1610-2397	Centre for Medical Imaging & Centre for Brain Research, Faculty of Medical and Health Sciences, Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand
55	Stuart	Williams	0000-0002-6507-0531	Department of Radiology, Norfolk & Norwich University Hospital, Colney Lane, Norwich, Norfolk, UK.
56	Jan	Witowski	0000-0001-9284-4830	Department of Radiology, New York University Grossman School of Medicine, New York, NY 10016, USA.
57	Zhongyi	Zhang	0000-0002-0851-9953	School of Information and Communication Technology, Griffith University, Australia
58	Alex	Zwanenburg	0000-0002-0342-9545	1. OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany 2. German Cancer Research Center (DKFZ), Heidelberg, Germany 3. Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany 4. Helmholtz Association / Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
59	Renato	Cuocolo	0000-0002-1452-1574	Department of Medicine, Surgery and Dentistry, University of Salerno, Baronissi, Italy
*Co-first authors and contributed equally to this work.				

Corresponding author:

Tugba Akinci D'Antonoli

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

ORCID: 0000-0002-7237-711X

E-mail: tugba.akincidantonoli@unibas.ch

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material: Data generated or analyzed during this study are presented with this manuscript .

Competing interests: The authors of this manuscript declare relationships with the following companies: **AAB:** CEO and shareholder of Quibim SL. Editorial Board Member of Insights into Imaging. **EK:** Speaker fees for Siemens Healthineers, Speaker fees for abbvie, member of the scientific advisory board and shareholder of contextflow GmbH, Vienna. Member of the scientific advisory board Gleamer. **RoC:** Support for attending meetings from Bracco and Bayer; research collaboration with Siemens Healthcare; co-funding by the European Union - FESR or FSE, PON Research and Innovation 2014-2020 - DM 1062/2021. Editorial Board Member of Insights into Imaging. **LSF:** General Electric Healthcare (Honoraria), Median Technologies (Honoraria), Sanofi (Honoraria), Guerbet (conference funding), Bristol-Myers Squibb (research grant). **MEK:** Meeting attendance support from Bayer. **LMB:** Editor-in-Chief of Insights into Imaging, member of the non-profit Scientific Advisory Boards of Quibim SL and the Girona Biomedical Research Institute. **DPdS:** Editorial Board Member of Insights into Imaging. Speaker fees for Bayer AG, Advisory Board for cook medical, Author fees for AMBOSS GmbH. **AP:** Editorial board member of Insights into Imaging. **FV:** None related to this study; received support to attend meetings from Bracco Imaging S.r.l., and GE Healthcare.

None of the authors related to the **Insights into Imaging** editorial team and editorial board has taken part in the review process of this article.

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.

Funding: None.

Authors' contributions: **BK, TAD, and RC** supervised the study, contributed to data collection, writing and editing the manuscript. **NM** performed the statistical analysis and edited the manuscript. All other authors participated in the Delphi consensus process. All authors read, edited if necessary, and approved the final manuscript.

Acknowledgements: This study was endorsed by the European Society of Medical Imaging Informatics (EuSoMII).

METhodological RadiomICs Score (METRICS): A quality scoring tool for radiomics research

Abstract

Purpose: To propose a new quality scoring tool, METhodological RadiomICs Score (METRICS), to assess and improve research quality of radiomics studies.

Methods: We conducted an online modified Delphi study with a group of international experts. It was performed in three consecutive stages: Stage#1, item preparation; Stage#2, panel discussion among EuSoMII Auditing Group members to identify the items to be voted; and Stage#3, four rounds of the modified Delphi exercise by panelists to determine the items eligible for the METRICS and their weights. The consensus threshold was 75%. Based on the median ranks derived from expert panel opinion and their rank-sum based conversion to importance scores, the category and item weights were calculated.

Results: In total, 59 panelists from 18 countries participated in selection and ranking of the items and categories. Final METRICS tool included 30 items within 9 categories. According to their weights, the categories were, in descending order of importance: study design, imaging data, image processing and feature extraction, metrics and comparison, testing, feature processing, preparation for modeling, segmentation, and open science. A web application and a repository were developed to streamline the calculation of the METRICS score and to collect feedback from the radiomics community.

Conclusion: In this work, we developed a scoring tool for assessing the methodological quality of the radiomics research, with a large international panel and a modified Delphi protocol. With its conditional format to cover methodological variations, it provides a well-constructed framework for the key methodological concepts to assess the quality of radiomic research papers.

Clinical relevance statement: A quality assessment tool, METhodological RadiomICs Score (METRICS), is made available by a large group of international domain experts, with transparent methodology, aiming at evaluating and improving research quality in radiomics and machine learning.

Keywords: Radiomics; Deep learning; Artificial intelligence; Machine learning; Guideline

Abbreviations:

CLEAR = CheckList for EvaluAtion of Radiomics research

ESR = European Society of Radiology

EuSoMII = European Society of Medical Imaging Informatics

IQR = Interquartile range

MAIC-10 = Must AI Criteria-10 checklist

METRICS = METhodological RadiomICs Score

RQS = Radiomics Quality Score

TRIPOD = Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

Key points:

- A methodological scoring tool, i.e.,METRICS, was developed for assessing the quality of the radiomics research, with a large international expert panel and a modified Delphi protocol.
- Proposed scoring tool presents the expert opinion-based importance weights of categories and items with a transparent methodology for the first time.
- METRICS accounts for varying use cases, from handcrafted radiomics to entirely deep learning-based pipelines.
- A web application was developed to help with the calculation of the METRICS score (https://metricsscore.github.io/metrics/METRICS.html) and a repository was created to collect feedback from the radiomics community (https://github.com/metricsscore/metrics).

Introduction

Radiomics is an evolving field of image analysis technique for extracting quantitative features from medical images with the premise of building predictive models and assisting clinical decision-making [1]. Since its introduction into medicine more than a decade ago, an exponential number of radiomics-related articles have been published yearly [2]. However, a growing translational gap exists between radiomics research and clinical practice [3, 4]. One of the main reasons for this issue is the poor quality of research methodology, including but not limited to, poor study design, inadequate description of image segmentation, feature extraction or model building methodology, lack of generalizability, lack of data, model and code sharing practices, all of which ultimately limit the reproducibility of the proposed radiomics models [3, 5–8].

In 2017, Lambin et al [9] proposed the radiomics quality score (RQS), a set of assessment criteria covering the radiomics workflow to improve the quality of radiomics research. Since then, many systematic reviews have been published applying the RQS to published research to examine the quality of radiomics studies [10]. Nevertheless, some RQS item definitions may lead to ambiguity and the applicability of the items can be limited based on different characteristics of the study design, which may negatively affect the reproducibility of the score even among experts in the field [10, 11]. In addition, as shown previously [11], a high RQS score does not always guarantee high quality of a study or lack of significant bias [12]. Furthermore, this assessment system was developed by a small group of researchers and the development process was not detailed in-depth in terms of how it deals with the relative importance of each item that contributes to overall radiomics research quality. Thus, the need for an easy-to-use and transparent evaluation system developed by an international group of experts persisted.

Recently, the CheckList for EvaluAtion of Radiomics Research (CLEAR) guideline for reporting radiomics studies that covers the entire life cycle of optimal radiomics research, was published and endorsed by the European Society of Radiology (ESR) and European Society of Medical Imaging Informatics (EuSoMII) [13]. The CLEAR reporting guideline has great potential to improve the quality of reporting in radiomics papers, which would ultimately lead to an improvement in research quality. Nevertheless, reporting guidelines are not assessment tools or instruments for measuring research quality [14, 15]. Thus, the need remains for an easy-to-use, reproducible assessment system for radiomics research. In this paper, we propose a new quality assessment tool, METhodological RadiomICs Score (METRICS), which was developed by a large group of international experts in the field and is easy to use, aimed

at improving research quality and closing the gap between research and clinical translation in radiomics and machine learning.

Material and Methods

Design and Development

As there is no guidance for developing scoring systems, the recommendations for developing reporting guidelines were followed [16]. Therefore, a steering committee (T.A.D., B.K., and R.C.) was established first to organize and coordinate the development of METRICS.

To develop the METRICS tool, an online modified Delphi study with a group of international experts was planned. The process was organized in three stages. The steering committee members conducted the first stage (Stage#1), consisting of item preparation. The second stage (Stage#2) was held with the participation of a group of panelists from the EuSoMII Radiomics Auditing Group for discussion of the items to be voted on. The third stage (Stage#3) was carried out in 4 rounds by two separate groups of panelists to determine the METRICS items and their weights. The first three rounds of Stage#3 were aimed at determining which methodological items were eligible for METRICS. The items' weights were then determined in the final round of Stage#3. Following each round, the panelists received structured feedback on the preceding round to reconcile individual opinions.

The surveys were open for at least two weeks in each round in Stage#3, and a reminder e-mail was sent one week, three days, and one day before the deadline. When necessary (e.g., when overlapping with major conferences or holidays), deadlines were extended to ensure a reasonable number of panelists was achieved.

The modified Delphi surveys were carried out using a Computer Assisted Web Interviewing (CAWI) system, i.e., Google Forms (Google LLC). For online group discussions online platforms, i.e., Google Docs (Google LLC) or WhatsApp (Meta Platforms Inc.), were used.

To simplify the calculation of the METRICS score, the development of an online calculation tool was planned. A GitHub repository was also planned for providing updates and gathering community feedback.

Anonymity

Although the panelists voted independently, the voting rounds of the modified Delphi exercise were not anonymous to track panelists' participation. Only the organizers had access to the panelists' data, and they preserved the anonymity of the votes and their respective comments during and after the voting tasks (i.e., when feedback was provided after rounds).

Informed consent

At the start of the Delphi questions, participants' informed consent was requested using the same form. Participants may have opted out of the study at any time. Those who indicated a desire to decline the survey were to be deleted from future invitations. Only while the round was active, panelists could withdraw their votes.

Consensus criteria

The vote for “strongly agree” and “agree” accounted for agreement and “strongly disagree” and “disagree” accounted for disagreement. The “neutral” votes were not included in either decision. The consensus was defined *a priori* as either agreement (agreement ≥75%) or disagreement (disagreement ≥75%) [17]. If there was no agreement or disagreement, it was referred to as "no consensus," and they were voted again. If “no consensus” items did not achieve agreement in the next voting, they were removed from the tool. The consensus items with disagreement were removed from the tool without further discussion.

Recruitment of participants

Individuals having significant experience in radiomics, machine learning, deep learning, informatics, or related editorial tasks from various countries were invited via an e-mail describing the development plan of the METRICS tool and explaining its purpose. Members of the EuSoMII Radiomics Auditing Group (Group#1 panelists) were assigned to discussion panels in Stage#2 and Round#3 of Stage#3. Other invitees (Group#2 panelists) were assigned to modified Delphi voting rounds (i.e., Round#1, Round#2, and Round#4 of Stage#3).

Modified Delphi

Stage#1 (preparation)

To identify potential items, a thorough and systematic literature review was conducted. Two members of the steering committee performed an independent literature search in PubMed using the following syntax to find the relevant checklists, guidelines, or tools: (radiomics) AND ((checklist) OR (guideline)). The search date was January 24th, 2023. All entries and related publications, if accessible by the readers, were assessed to determine the currently available tools. All eligible documents found were independently evaluated by the entire steering committee to develop the initial template of METRICS.

Participants were requested to consider the following principles: *i*, there should be no overlap between items; *ii*, an ideal study should be able to achieve a perfect score (i.e., all points available or 100%), meaning that items should not be mutually exclusive; *iii*, items must

be objectively defined, to increase reproducibility; *iv*, not only hand-crafted but also studies based on deep learning should be considered and item conditionality should be assessed accordingly; *v*, since this is a methodological scoring system, the items should be mainly related to the “materials and methods” and “results” sections of a research paper; *vi*, while items should also aim at improving the methodological reproducibility and transparency of the studies, METRICS is not a reporting checklist; and *vii*, items should point out potential bias sources and help users to avoid them.

Considering the principles defined above, an initial draft was created with three organizers of the METRICS project. For any disagreement among the organizers, the decisions were made based on a majority vote.

Stage#2 (discussion with Group#1 panelists)

The items prepared by the organizers were presented to the EuSoMII Radiomics Auditing Group with the same principles and discussed online. This stage was an open discussion and not anonymous. The panelists were free to suggest adding, removing, merging, and modifying items.

Stage#3 (modified Delphi rounds)

Round#1 (item selection)

On a 5-point Likert scale (strongly agree; agree; neutral; disagree; strongly disagree), the Group#2 panelists were asked to rate the extent to which they agreed with the inclusion of each item on the METRICS tool. With a text box, participants were further asked for suggestions on the item's name and definition. In addition, a text box was provided at the end of each section for participants to suggest additional items. After this round, the Group#2 panelists were provided with a statistical summary of each item from Round#1, along with anonymous comments.

Round#2 (continued for item selection)

The same panelists as in Round#1 were invited to participate in Round#2. Panelists who were invited but did not respond to Round#1 were also invited to participate in Round#2. Using the same structure as Round#1, panelists were also presented with items that reached no consensus as well as new item or items suggested in previous round.They were asked to use the same 5-point Likert scale to express their level of agreement with the inclusion of each item in the METRICS tool. No new item proposal was asked in this round. After Round#2, the same panelists were provided with a statistical summary of each item from Round#2, along with anonymized comments.

Round#3 (group discussion with EuSoMII Radiomics Auditing Group)

The purpose of Round#3 was to discuss the results of the previous rounds, modify if necessary, and finalize the items to be included in the METRICS tool. It was held on online platforms (Google Docs and WhatsApp Group). All Group#1 panelists were invited. The discussion included both agreed and unresolved topics. Any modification proposals were discussed and items were edited in consensus by the steering committee.

Round#4 (ranking of finalized items to determine the weights)

Group#2 panelists who participated in at least one of the first two rounds (Round#1 and Round#2) were invited to this round. The panelists were asked to rank the categories and then all items within each category in order of their importance in radiomics research. After Round#4, the same panelists were provided with an anonymized statistical summary of each item and category.

Pilot testing

We invited Group#1 panelists to test the usability and understandability of the online checklist. Also, the final METRICS tool was tested on studies from the literature, including a sample of different pipeline designs and aims (i.e., handcrafted radiomics, deep radiomics and end-to-end deep learning; lesion characterization and region of interest segmentation).

Statistical analysis

Descriptive statistics (i.e., median, interquartile range, percentage) were used to present the results. The ranks derived from hierarchical (i.e., multi-tiered) ranking with expert panel opinion were aggregated using their median value. Using the rank-sum method [18, 19], median ranks were first converted to importance scores with the following formula: *Score = (N+1) - Rank*, where *N* is the total number of categories or total number of items within a category. The category weights were then rescaled to 1. The final weights of each item were computed as the product of the category and item weights (e.g., [weight of Category A] x [weight of Item#1 in Category A]). The items within the respective category went through the same rescaling procedure. The final METRICS score was calculated on a percentage scale, accounting for the conditionality of items and categories.

Results

All key points are summarized with a flowchart in **Figure 1**.

Modified Delphi

In total, the 3 steering committee members invited 61 experts to participate in this study, 56 of which accepted the invitation. In detail, 14 experts from the EuSoMII Radiomics Auditing Group (Group#1) accepted the invitation to participate in panel discussions (i.e., discussions at Stage#2 and Round#3 of Stage#3), together with the steering committee members. Furthermore, 42 experts (Group#2) accepted the invitation to perform Delphi voting (i.e., rating in Round#1 and Round#2; ranking in Round#4 of Stage#3). Country data of all participants is presented in **Figure 2**.

The literature search resulted in 58 publications. After independent evaluation of the content of these publications by steering committee members, 16 relevant checklists, guidelines or quality scoring tools were identified as potentially useful for designing a new quality scoring tool [9, 20–34]. Based on the results of this literature review and previous experience, 33 items were initially drafted. These items were then reduced to 30 after discussion with the Group#1 panelists in Stage#2, as three were considered unclear or partly overlapping with other entries, with which they were merged.

The 30 items obtained after Stage#2 discussion were presented to the Group#2 panelists for the first round of the Delphi survey, which was completed by 40 of the 42 panelists. The consensus for an agreement was achieved for 26 items, while 4 items failed to achieve any consensus. No item reached the consensus threshold for disagreement. There was one new item proposal that was added to the list after discussion by the steering committee (item#17, robustness assessment of end-to-end deep learning pipelines). A summary of the votes in Delphi Round#1 is presented in **Figure 3**. The highest agreement (100%) was achieved by item#21 (i.e., consideration of uncertainty).

Following the Round#1, 4 items with no consensus and 1 newly proposed item were presented to the Group#2 panelists in Round#2 of the Delphi process. In this round, 41 of the 42 panelists participated. The consensus for an agreement was achieved for 30 items. There was no consensus on one item about prospective data collection, which was therefore removed from the list. There was no disagreement with consensus. A summary of the votes in Round#2 is presented in **Figure 3**.

All Group#1 panelists were invited to Round#3 for the panel discussion by the steering committee members. A small number of minor modifications were made to the item definitions at this time. The agreement was achieved for all 30 items within 9 categories.

The final Delphi round, Round#4, consisted of ranking of all 9 categories and the 30 items divided by category. This was performed by all 42 of the Group#2 panelists. A summary of the category and item ranks in Round#4 is presented in **Supplementary file 1 Figure S1** and **Supplementary file 1 Figure S2**, respectively.

Weights calculated for categories and items are presented in **Figure 4**. For categories, the highest and lowest weights belonged to study design and open science, respectively. According to their final weights, top 5 items with highest weights were as follows: item#3 (i.e., high-quality reference standard with a clear definition; weight, 0.092); item#27 (i.e., external testing; weight, 0.075), item#2 (i.e., eligibility criteria that describe a representative study population; weight, 0.074), item#11 (i.e., appropriate use of image preprocessing techniques with transparent description; weight, 0.062), and item#18 (i.e., proper data partitioning process; weight, 0.060). The lowest weights belonged to the three items of category "open science" and were as follows: item#28 (i.e., data availability, weight, 0.007), item#29 (i.e., code availability, weight, 0.007), and item#30 (i.e., model availability, weight, 0.007).

Anonymized individual votes and ranks obtained in the Round#1, Round#2, and Round#4 of the Stage#3 are presented in **Supplementary file 2**.

Finalized METRICS tool

The final METRICS tool included 30 items within 9 categories and is presented in **Table 1** with relative item weights. It also accounts for different study pipelines by including several conditional items. **Figures 5** and **6** present a flow diagram to exemplify their usage in practice.

A user-friendly online calculation tool was prepared to streamline the calculation of the METRICS score (https://metricsscore.github.io/metrics/METRICS.html). It also allows printing (paper and PDF) and exporting (Excel spreadsheet). **Supplementary file 3** (without explanation) and **supplementary file 4** (with explanation) allow downloading the METRICS tool in table format. However, the use of the online tool mentioned above is highly recommended, as the final METRICS percentage score is based on the maximum achievable absolute score after accounting for item conditionality. This calculation can be performed automatically by the web-based tools (both online and offline versions). **Supplementary file 5** includes evaluation examples from the literature, covering the use of METRICS on different radiomics pipeline designs.

A GitHub repository was also set up for the METRICS tool (https://github.com/metricsscore/metrics). The discussion function was also activated to receive community feedback to improve it in the future. Also, an offline version of the calculation tool can be downloaded from this repository, which requires no setup or installation but directly starts working on common web browsers such as Google Chrome (recommended; Google LLC). The online calculation tool and potential updates can also be accessed via this repository.

Total score categories

To improve the comprehensibility of the METRICS total score, we propose the use of 5 arbitrary categories as a representation of gradually increasing quality, namely total score between 0-20%, “very low”; 21-40%, “low”; 41-60%, “moderate”; 61-80%, “good”; and 81-100%, “excellent” quality. However, these categories should be validated through future systematic reviews using METRICS, and should be used as a complement of METRICS and not as a substitute for the quantitative score.

Discussion

In this work, we developed a scoring tool for assessing the methodologic quality of the radiomics research, i.e., METRICS, based on the input of a diverse and large international panel with 59 participants. Our study was conducted in 3 consecutive stages, with 4 rounds of the modified Delphi exercise in the last stage. Based on panelist ratings, 30 items within 9 categories were ultimately included in the METRICS tool. The weights of these items were then calculated using a hierarchical ranking of categories and items based on the rank-based assessment by the Delphi panelists. A web application was developed to automate the calculation of the METRICS score, and a repository was created to collect feedback from the radiomics research community.

There have been only few tools proposed to assess the methodological quality of radiomics research in the literature, e.g., the RQS [9]. Despite the fact that the RQS was published as part of a review article, it has received so much attention from the community that it became the de facto standard for evaluating radiomics methodology [10]. Although it was developed and published by leading radiomics researchers, it lacked methodological transparency in terms of how it was developed and how the scores for each item were assigned. The first and most widely used version was designed to evaluate traditional radiomics and modeling in general, and thus does not apply to deep learning workflows. Although not directly related to radiomics, the Must AI Criteria-10 (MAIC-10) checklist can be used to evaluate the quality of AI and medical imaging studies [35]. It aims to simplify the

process while overcoming some of the limitations of other published checklists in the fields of artificial intelligence and medical imaging. MAIC-10 is a very short and simple tool that covers a wide range of concepts. According to the authors of MAIC-10, unlike other checklists or quality scoring tools, it was designed to provide a quantitative, objective, and reproducible quality score with a broad scope of applications across studies on AI in medical imaging. MAIC-10 achieved a high correlation score to CLAIM [26], a widely used 42-item reporting checklist, despite being tested on a small number of publications from the journal in which it was published. It was also proposed as the most reproducible checklist in terms of intra-observer reproducibility, with CLAIM taking second place. However, the MAIC-10 scores are unweighted, namely ignoring the relative importance of each item and simply assigning a score of 1 for adherence. Such a simple scoring strategy was also used for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist as well [36]. A recent radiomics-specific reporting checklist, the CLEAR checklist, was developed by an international initiative led by a group of experts and endorsed by ESR and EuSoMII [13]. Although CLEAR was designed primarily as a reporting tool and not a methodological guide, it still provides useful information about the methodology. Furthermore, it has a shortened version called CLEAR-S that focuses solely on methodological aspects and open science, with no score or weights. There are also reporting checklists for AI and medical imaging that were not specifically designed for radiomics, such as CLAIM [26]. CLAIM is a highly cited checklist that provides guidance for reporting and methodology. However, it was created by a relatively small group of scientists with no formal methodology for determining item eligibility, such as the Delphi method. Of note, a recent article provides a comprehensive review of available guidelines that can be used in AI research and medical imaging [37].

To develop the proposed scoring system, we used a modified Delphi method with an international group of panelists and defined weights of each item to present a more nuanced way of assessment. As a result, the category “Study Design” had the highest weight, and thus the biggest effect on the final score. This result is such that adhering to all items of the category may already allow a METRICS score ranging between 20% and 25%, considering all possible conditionals. It includes three items as follows: *i*, adherence to radiomics and/or machine learning-specific checklists or guidelines; *ii*, eligibility criteria that describe a representative study population; and *iii*, high-quality reference standard with a clear definition. The first item was introduced as a new concept in comparison to the RQS [9] and MAIC-10 [35] tools. The authors of the MAIC-10 checklist included the study design as a single item and defined it as a very broad concept. While most of their 10 items were discussed in at least half of the studies evaluated as part of the MAIC-10, the study design was not defined in any of the studies evaluated. Previously, the CLEAR checklist [13] and, to a lesser extent, CLAIM [26] drew attention to some of these concepts in terms of reporting.

It may appear surprising that the category related to open science practices had the lowest weight, and thus the lowest effect on the final score. As widely known, radiomic studies suffer from significant reproducibility issues, which have been mainly attributed to the lack of data, code, and model sharing practices leading to poor generalizability [8, 38–40]. This apparent discrepancy may be attributable to the panelists' consideration that proper study design “comes first”. In other words, if the study’s methodological steps are flawed, data and model availability becomes a secondary concern as these would still lack reproducibility. It should be noted that “reproducibility” refers to the ability to implement the same methods reported in a study on the same input data and obtain identical findings. This is entirely dependent on sharing of data, code and models and represents a guarantee of the correctness in reporting study results. On the other hand, “replicability” defines the ability of obtaining consistent results in relation to the same hypothesis when using a different patient population (and even partially different methodology). In practice, replicability is a better indicator of robustness of the conclusions drawn from a study in relation to a specific hypothesis and does not require meeting the criteria of the METRICS open science items, but only appropriate reporting of the study hypothesis and methodology. Finally, “generalizability” refers to the ability of a model to be applicable on a different patient distribution compared to the original study, and represents one of the major challenges facing radiomics and its translation to the clinical setting. More information on these topics can be found in [41]. This apparently counterintuitive disconnection between open science practice and study quality can also be seen elsewhere in the recent literature. For instance, in a meta-research study about AI literature published in RSNA journals, it was reported that only 13% of the included literature shared data, 30% of the included literature shared code and only 11% of the shared code was actually reproducible [42]. Another recent literature review showed a similar trend, where the data sharing rate within randomly sampled AI publications from Q1-Q2 journals was only 1%, and proper model sharing (i.e., sharing premodeling, modeling, and post-modeling files at once) was observed in only 6% of the included studies [8]. Our belief is that papers scoring highly on METRICS will allow improved replicability and generalizability.

Even though an item focused on the role of prospective study design/data collection was initially included, the panelists were unable to reach an agreement on it, and it is not present in the final METRICS tool. The RQS, on the other hand, places a strong emphasis on prospective studies, particularly those registered in trial databases, and awards the studies with the highest score of the tool for this item [9]. Based on the feedback of the panelists during Round#1 of Stage#3, the most likely reason for this would be that radiomics research requires large data sets, which are difficult to achieve with prospective studies when compared to retrospective design and data sets. Another issue raised by panelists was the potential penalization of large retrospective data sets in comparison to prospective studies with small

data sets. Therefore, despite its undoubtedly high importance in clinical research, the role and added value of prospective data collection currently remain uncertain in radiomics and artificial intelligence research within the medical imaging domain, and could be secondary compared to other considerations on overall data labeling and management as established by the METRICS expert panel. It would be worthwhile to receive community feedback on this and other topics in the future, which may contribute to future revisions of METRICS.

Our work has several distinguishing features and strengths. First, the weights obtained in this work were not assigned arbitrarily but were the result of expert ranking. This was a primary goal of the study as there has been no previous work on radiomics quality scoring that has presented a transparent methodology for assigning item weights. Second, the METRICS tool considers not only hand-crafted radiomics but also deep learning-based radiomics. Third, both Group#1 and Group#2 had a large number of panelists. Furthermore, the panel was diverse in terms of country and domain expertise. This was necessary to reduce noise in calculations. Fourth, panelist participation in the Delphi rounds was also very high, with a minimum of 95% (40 of 42). Fifth, we created an easy-to-use web application to streamline scoring. This was crucial because METRICS contains conditional items that cover all aspects of radiomics, which may make the calculation difficult on paper. Finally, we established a living repository to discuss the METRICS tool and its content and receive feedback in order to improve them in the future.

There are however several limitations to declare. First, our modified Delphi procedure was not completely anonymous and the steering committee had access to identities, which was a significant deviation from the standard Delphi exercise. We chose this approach to ensure panelist participation. Nevertheless, we kept the votes and comments anonymous for other panelists. Second, the ranking in Round#4 of Stage#3 did not account for potential items of equal importance. An analytical hierarchy process and pairwise voting could have been an alternative approach that takes equality into account. However, by this method, the number of questions would have been doubled in Round#4, which might cause fatigue and had negative effects on the scoring process. Third, during tool development, the need for conditional items became apparent, even if their use may complicate the scoring process. In reality, radiomic research involves numerous methodological variations and nuances that could be overlooked with a fixed item list. However, the availability of online and offline automated calculation tools should help mitigate this limitation. Fourth, the conditionality of the items or categories was not taken into account when calculating weights. Dynamic weights would have necessitated calculations of all possible conditional combinations and, as a result, multiple rankings, which is impractical and of limited value as differences are expected to be small compared to the

current METRICS tool. Fifth, the number of items in each category varied. Nonetheless, the weighting process accounted for this to avoid biases in the final tool due to item number within categories. Sixth, the order of the items and categories in the Delphi rounds was fixed, which may have an influence on ranking and introduce bias. Alternatively, the order of these could have been randomized during voting, and this could have been done independently for each panelist as well. Finally, the reproducibility of the METRICS was not evaluated. Such an analysis necessitates a dedicated study design by incorporation of other tools for comparison, which should be performed in a future investigation.

In conclusion, we developed a scoring tool for a comprehensive assessment of the methodologic quality of the radiomics research, i.e., METRICS, with a large international panel of experts and by using a modified Delphi protocol. With its flexible format to cover all methodological variations, it provides a well-constructed framework for the key methodological concepts to assess the quality of the radiomic research papers. A web application was developed to help with the calculation of the METRICS score, and a repository was created to collect feedback from the radiomics community. We hope that the researchers would benefit from this tool when designing their studies, assessing the methodological quality of papers in systematic reviews, and that journals would adopt the METRICS quality scoring tool for peer review. Comments and contributions to this tool are welcome through its repository to improve it in the future.

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. Kocak B, Baessler B, Cuocolo R, et al (2023) Trends and statistics of artificial intelligence and radiomics research in Radiology, Nuclear Medicine, and Medical Imaging: bibliometric analysis. Eur Radiol. https://doi.org/10.1007/s00330-023-09772-0
3. Kocak B, Bulut E, Bayrak ON, et al (2023) NEgatiVE results in Radiomics research (NEVER): A meta-research study of publication bias in leading radiology journals. Eur J Radiol 163:110830. https://doi.org/10.1016/j.ejrad.2023.110830
4. 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
5. Papanikolaou N, Matos C, Koh DM (2020) How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 20:33. https://doi.org/10.1186/s40644-020-00311-4
6. Buvat I, Orlhac F (2019) The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. J Nucl Med Off Publ Soc Nucl Med 60:1543–1544. https://doi.org/10.2967/jnumed.119.235325
7. Vallières M, Zwanenburg A, Badic B, et al (2018) Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med Off Publ Soc Nucl Med 59:189–193. https://doi.org/10.2967/jnumed.117.200501
8. 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)00635–3. https://doi.org/10.1016/j.acra.2022.11.030
9. 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
10. Spadarella G, Stanzione A, Akinci D'Antonoli T, et al (2023) Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. Eur Radiol 33:1884–1894. https://doi.org/10.1007/s00330-022-09187-3
11. Sanduleanu S, Woodruff HC, de Jong EEC, et al (2018) Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother Oncol J Eur Soc Ther Radiol Oncol 127:349–360. https://doi.org/10.1016/j.radonc.2018.03.033
12. Welch ML, McIntosh C, Haibe-Kains B, et al (2019) Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol J Eur Soc Ther Radiol Oncol 130:2–9. https://doi.org/10.1016/j.radonc.2018.10.027
13. 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
14. Caulley L, Catalá-López F, Whelan J, et al (2020) Reporting guidelines of health research studies are frequently used inappropriately. J Clin Epidemiol 122:87–94. https://doi.org/10.1016/j.jclinepi.2020.03.006
15. Logullo P, MacCarthy A, Kirtley S, Collins GS (2020) Reporting guideline checklists are not quality evaluation forms: they are guidance for writing. Health Sci Rep 3:e165. https://doi.org/10.1002/hsr2.165
16. Moher D, Schulz KF, Simera I, Altman DG (2010) Guidance for Developers of Health Research Reporting Guidelines. PLOS Med 7:e1000217. https://doi.org/10.1371/journal.pmed.1000217
17. Diamond IR, Grant RC, Feldman BM, et al (2014) Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 67:401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002
18. Roszkowska E (2013) Rank Ordering Criteria Weighting Methods – a Comparative

Overview. Optim Stud Ekon 14–33
19. Stillwell WG, Seaver DA, Edwards W (1981) A comparison of weight approximation techniques in multiattribute utility decision making. Organ Behav Hum Perform 28:62–77. https://doi.org/10.1016/0030-5073(81)90015-5
20. Whiting PF, Rutjes AWS, Westwood ME, et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
21. Bossuyt PM, Reitsma JB, Bruns DE, et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 351:h5527. https://doi.org/10.1136/bmj.h5527
22. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 13:1. https://doi.org/10.1186/s12916-014-0241-z
23. Luo W, Phung D, Tran T, et al (2016) Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res 18:e323. https://doi.org/10.2196/jmir.5870
24. Martin J (2017) © Joanna Briggs Institute 2017 Critical Appraisal Checklist for Analytical Cross Sectional Studies
25. Vallières M, Zwanenburg A, Badic B, et al (2018) Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med Off Publ Soc Nucl Med 59:189–193. https://doi.org/10.2967/jnumed.117.200501
26. 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
27. 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
28. Orlhac F, Nioche C, Klyuzhin I, et al (2021) Radiomics in PET Imaging:: A Practical Guide for Newcomers. PET Clin 16:597–612. https://doi.org/10.1016/j.cpet.2021.06.007
29. 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
30. Shur JD, Doran SJ, Kumar S, et al (2021) Radiomics in Oncology: A Practical Guide. Radiogr Rev Publ Radiol Soc N Am Inc 41:1717–1732. https://doi.org/10.1148/rg.2021210037
31. Sollini M, Cozzi L, Ninatti G, et al (2021) PET/CT radiomics in breast cancer: Mind the step. Methods San Diego Calif 188:122–132. https://doi.org/10.1016/j.ymeth.2020.01.007
32. Volpe S, Pepa M, Zaffaroni M, et al (2021) Machine Learning for Head and Neck Cancer: A Safe Bet?-A Clinically Oriented Systematic Review for the Radiation Oncologist. Front Oncol 11:772663. https://doi.org/10.3389/fonc.2021.772663
33. Jha AK, Bradshaw TJ, Buvat I, et al (2022) Nuclear Medicine and Artificial Intelligence: Best Practices for Evaluation (the RELAINCE Guidelines). J Nucl Med Off Publ Soc Nucl Med 63:1288–1299. https://doi.org/10.2967/jnumed.121.263239
34. 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
35. 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
36. Heus P, Damen JAAG, Pajouheshnia R, et al (2019) Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness

of reporting of prediction model studies. BMJ Open 9:e025611. https://doi.org/10.1136/bmjopen-2018-025611
37. Klontzas ME, Gatti AA, Tejani AS, Kahn CE (2023) AI Reporting Guidelines: How to Select the Best One for Your Research. Radiol Artif Intell 5:e230055. https://doi.org/10.1148/ryai.230055
38. Gidwani M, Chang K, Patel JB, et al (2023) Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models. Radiology 307:e220715. https://doi.org/10.1148/radiol.220715
39. 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
40. Park JE, Park SY, Kim HJ, Kim HS (2019) Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol 20:1124–1137. https://doi.org/10.3348/kjr.2018.0070
41. National Academies of Sciences Engineering, Medicine (2019) Reproducibility and Replicability in Science. The National Academies Press, Washington, DC
42. Venkatesh K, Santomartino SM, Sulam J, Yi PH (2022) Code and Data Sharing Practices in the Radiology Artificial Intelligence Literature: A Meta-Research Study. Radiol Artif Intell 4:e220081. https://doi.org/10.1148/ryai.220081

Tables

Table 1: METRICS tool.

Categories		No.	Items	Weights	Score[6]
Study Design		#1	Adherence to radiomics and/or machine learning-specific checklists or guidelines	0.037	
		#2	Eligibility criteria that describe a representative study population	0.074	
		#3	High-quality reference standard with a clear definition	0.092	
Imaging Data		#4	Multi-center	0.044	
		#5	Clinical translatability of the imaging data source for radiomics analysis	0.029	
		#6	Imaging protocol with acquisition parameters	0.044	
		#7	The interval between imaging used and reference standard	0.029	
Segmentation[1]		#8	Transparent description of segmentation methodology	0.034	
		#9	Formal evaluation of fully automated segmentation[2]	0.022	
		#10	Test set segmentation masks produced by a single reader or automated tool	0.011	
Image Processing and Feature Extraction		#11	Appropriate use of image preprocessing techniques with transparent description	0.062	
		#12	Use of standardized feature extraction software[3]	0.031	
		#13	Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement	0.041	
Feature Processing		#14	Removal of non-robust features[4]	0.020	
		#15	Removal of redundant features[4]	0.020	
		#16	Appropriateness of dimensionality compared to data size[4]	0.030	
		#17	Robustness assessment of end-to-end deep learning pipelines[5]	0.020	
Preparation for Modeling		#18	Proper data partitioning process	0.060	
		#19	Handling of confounding factors	0.030	
Metrics and Comparison		#20	Use of appropriate performance evaluation metrics for task	0.035	
		#21	Consideration of uncertainty	0.023	
		#22	Calibration assessment	0.018	
		#23	Use of uni-parametric imaging or proof of its inferiority	0.012	
		#24	Comparison with a non-radiomic approach or proof of added clinical value	0.029	
		#25	Comparison with simple or classical statistical models	0.018	
Testing		#26	Internal testing	0.037	
		#27	External testing	0.075	
Open Science		#28	Data availability	0.007	
		#29	Code availability	0.007	
		#30	Model availability	0.007	
Total METRICS score (should be given as percentage)					

[1] Conditional for studies including region/volume of interest labeling. [2] Conditional for studies using fully automated segmentation. [3] Conditional for the hand-crafted radiomics. [4] Conditional for tabular data use. [5] Conditional on the use of end-to-end deep learning. [6] Score is simply the weight if present and 0 otherwise. Proposed total score categories: 0-20% = very low, 21-40% = low, 41-60% = moderate, 61-80% = good, and 81-100% = excellent.

Figures

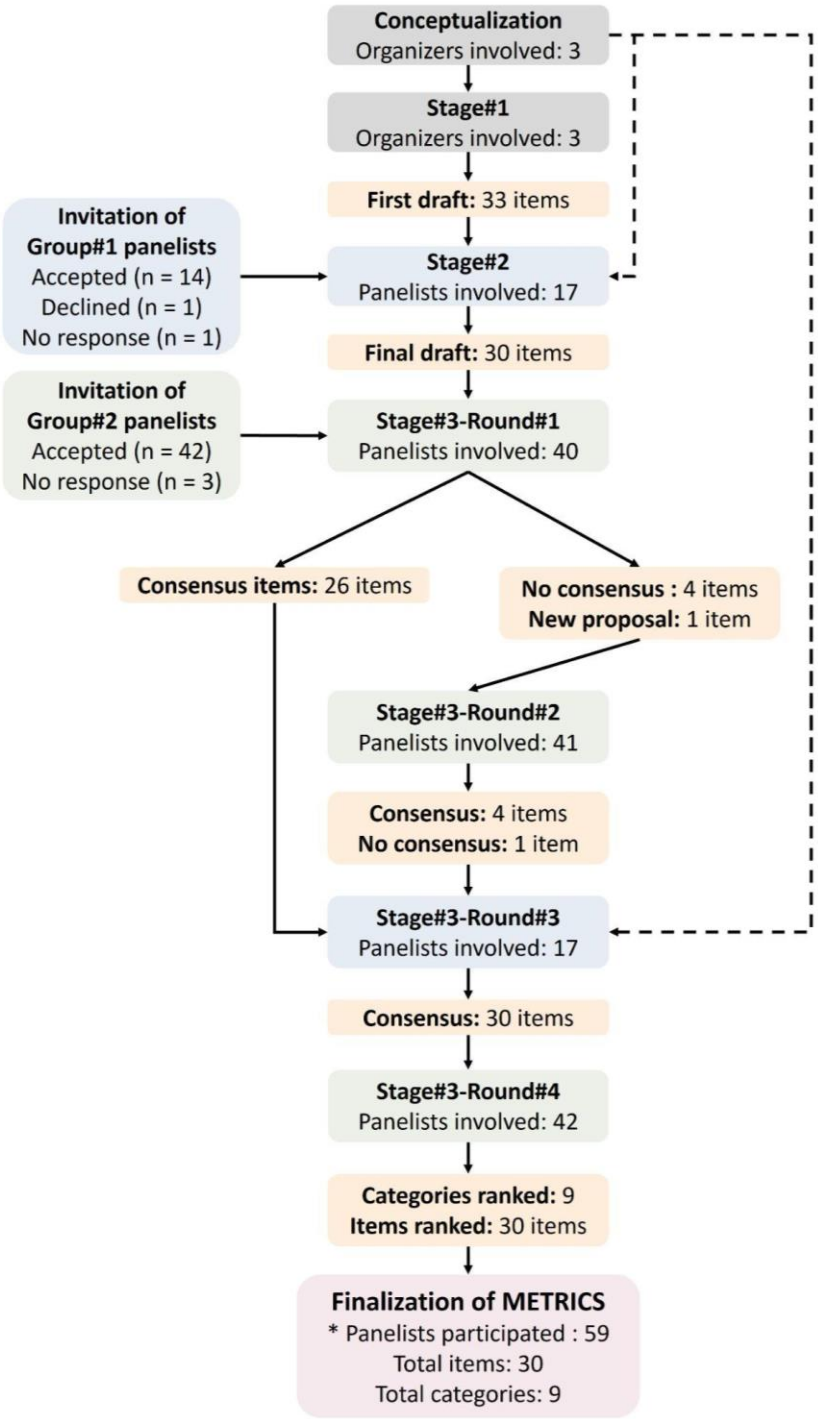

Figure 1: Key steps in the development of METRICS. Boxes related to stages and rounds are color-coded based on the main group of panelists involved. Dotted lines indicate the participation of organizers in the discussions in the relevant rounds as panelists. *Including organizers (i.e., steering committee members).

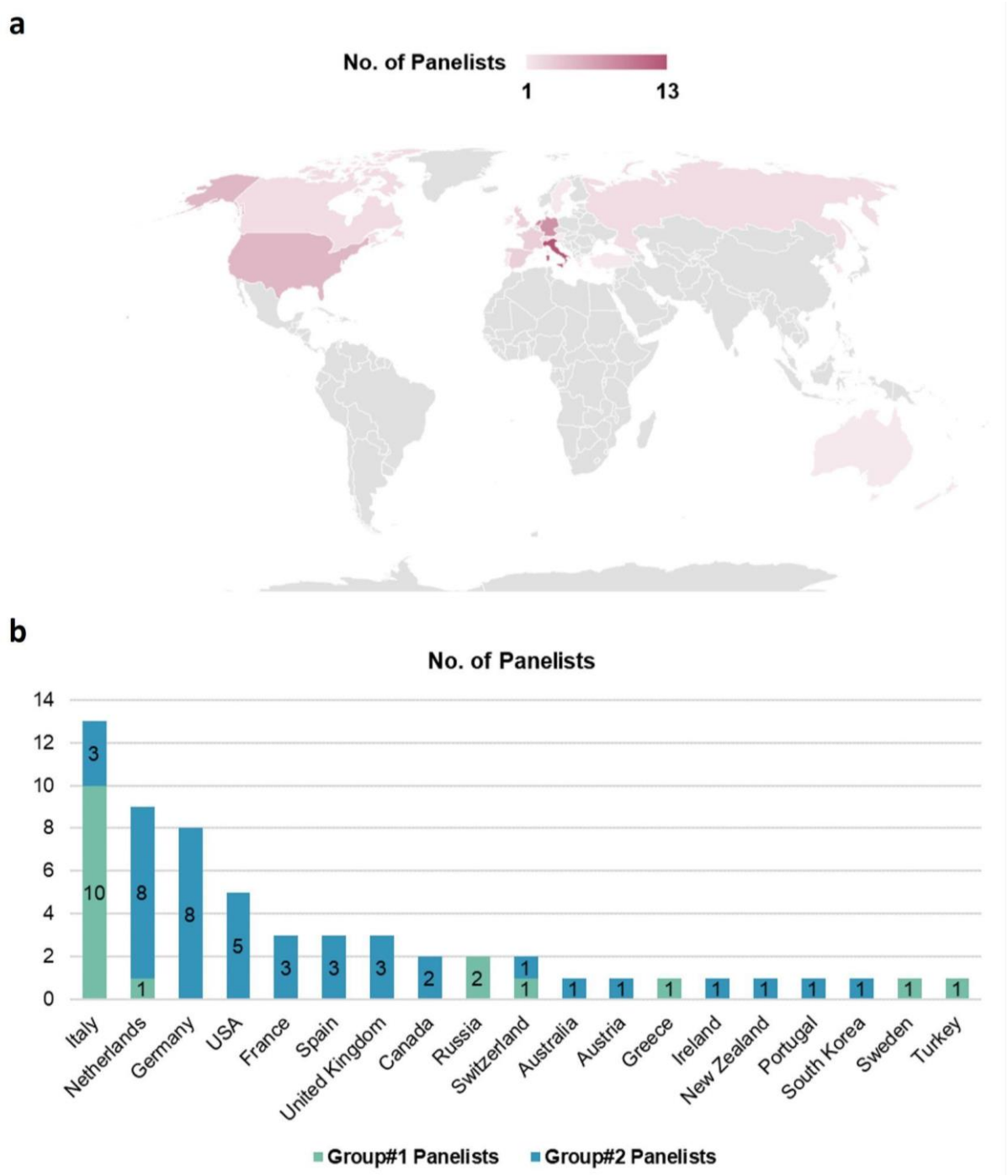

Figure 2: Country of panelists. ***a***, World map for distribution of 59 panelists including three organizers by country. ***b***, Countries by groups. *Group#1*, EuSoMII auditing group including three organizers; *Group#2*, voters participated in Round#1, Round#2, and Round#4 of Stage#3. In case of multiple countries, the country of the first affiliation was considered.

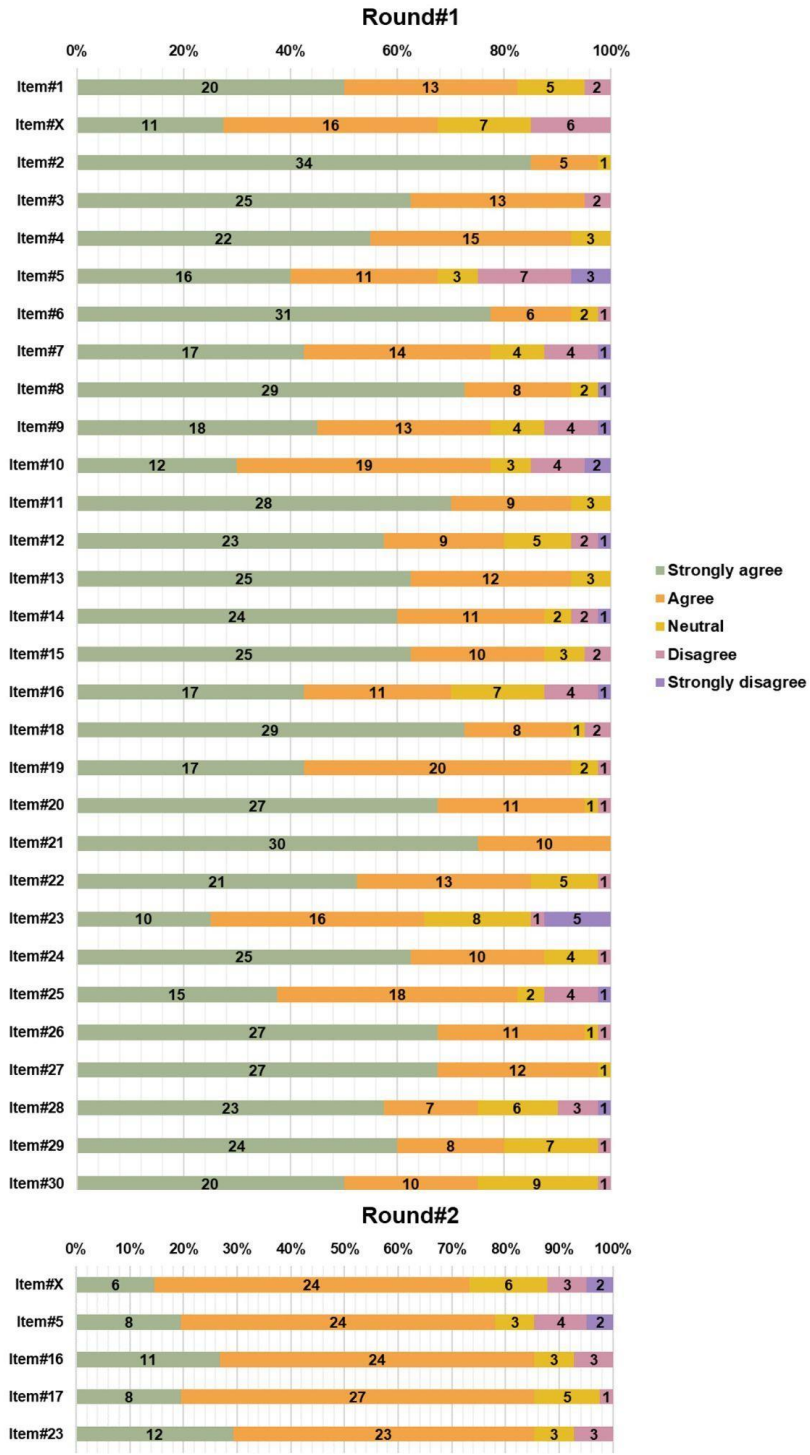

Figure 3: Rates from modified Delphi Round#1 and Round#2 of Stage#3. The number of the items matches those of the final METRICS tool. Item#X, i.e., prospective data collection, stands for the excluded item from the final METRICS tool. Please note Item#17 is missing in Round#1, which is the proposed item in Round#1 to be voted in Round#2.

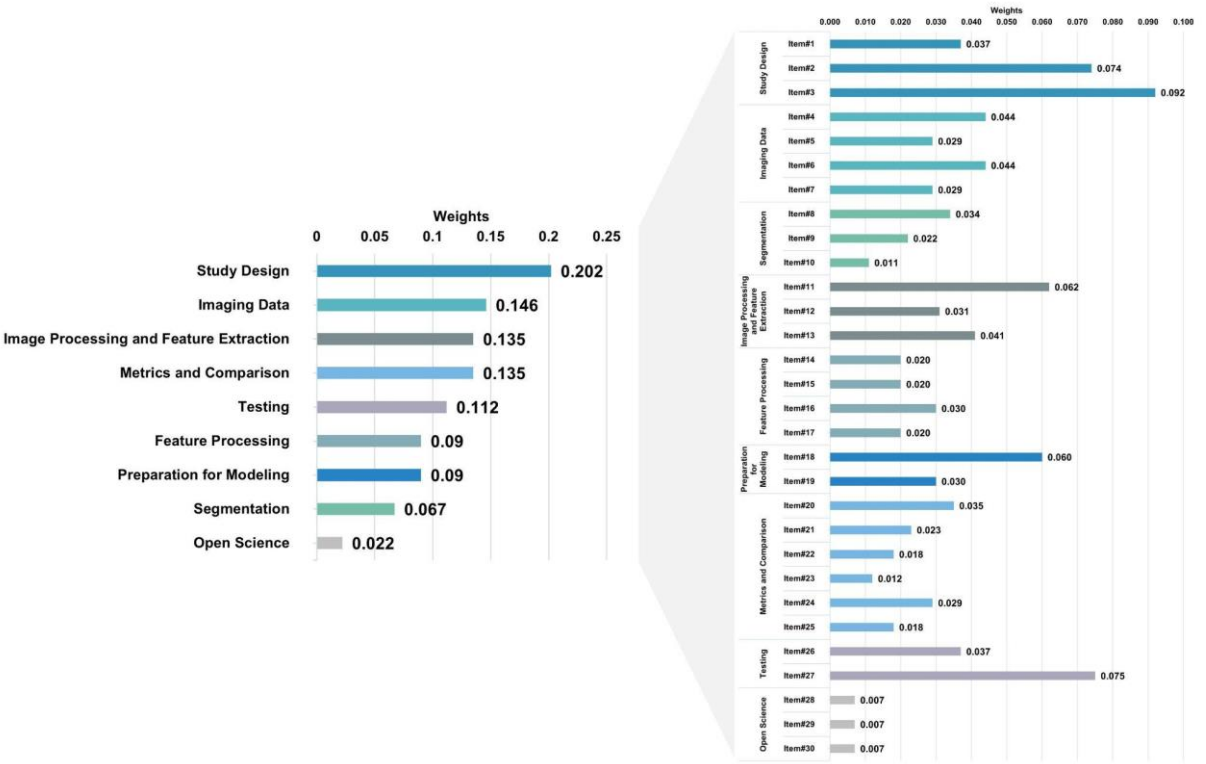

Figure 4: Weights of METRICS categories and items.

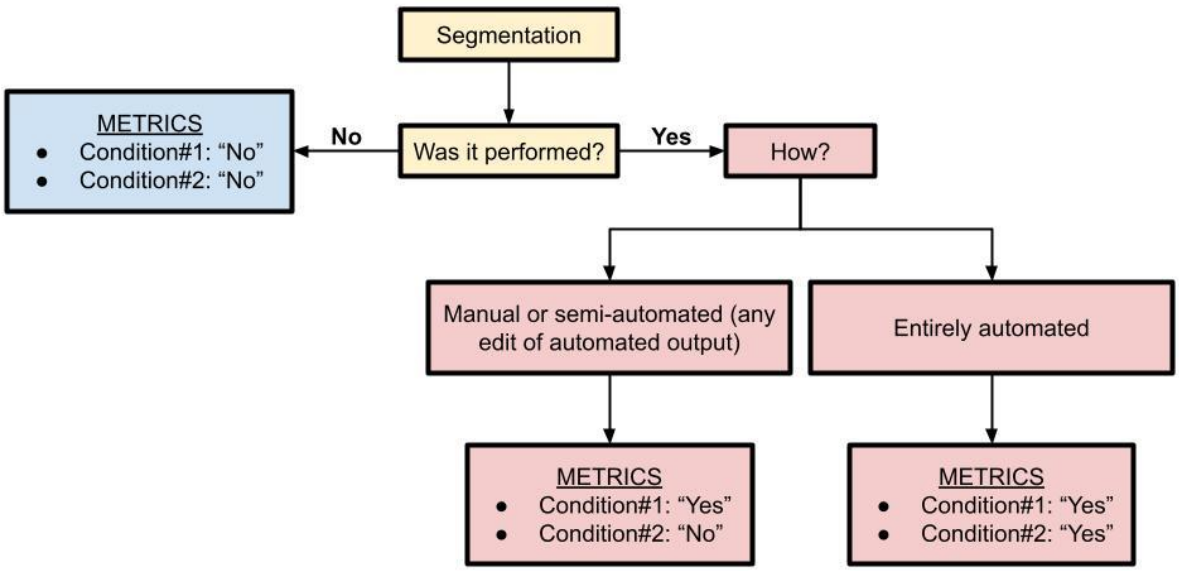

Figure 5: Use of conditions for the “Segmentation” section. Please note, the term “segmentation” is referred to either fine (e.g., semantic or pixel-based) or rough (e.g., cropping or bounding box) delineation of a region or volume of interest within an image or image stack for model training or evaluation. Studies can also be performed without such annotations, for example, using class labels that are assigned either to the entire image, volume, exam or patient or with unsupervised approaches that require no labeling at all (e.g., clustering models).

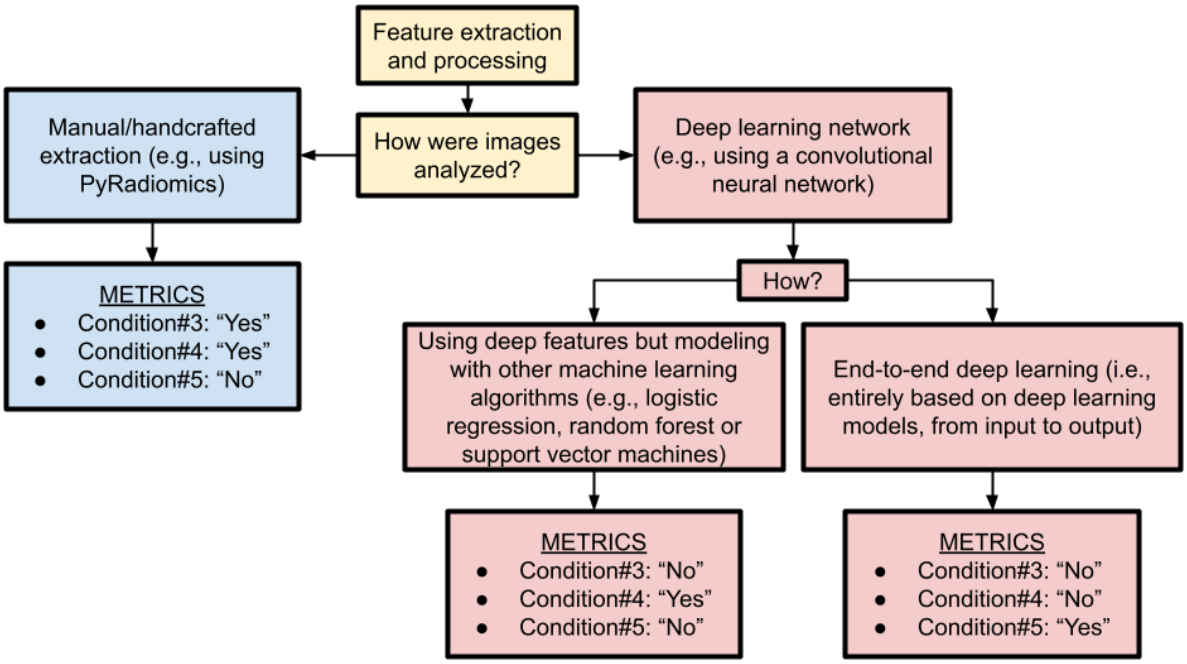

Figure 6: Use of conditions related to the sections "Image Processing and Feature Extraction'' and "Feature Processing". Please note the flowchart assumes a single pipeline is used in a given study. However, different techniques might coexist in a single study. For instance, a study might include both hand-crafted feature extraction and end-to-end deep learning for comparison purposes, in such a case, all conditions can be selected as "Yes".

Electronic Supplementary Materials

Supplementary File 1: Rank statistics.

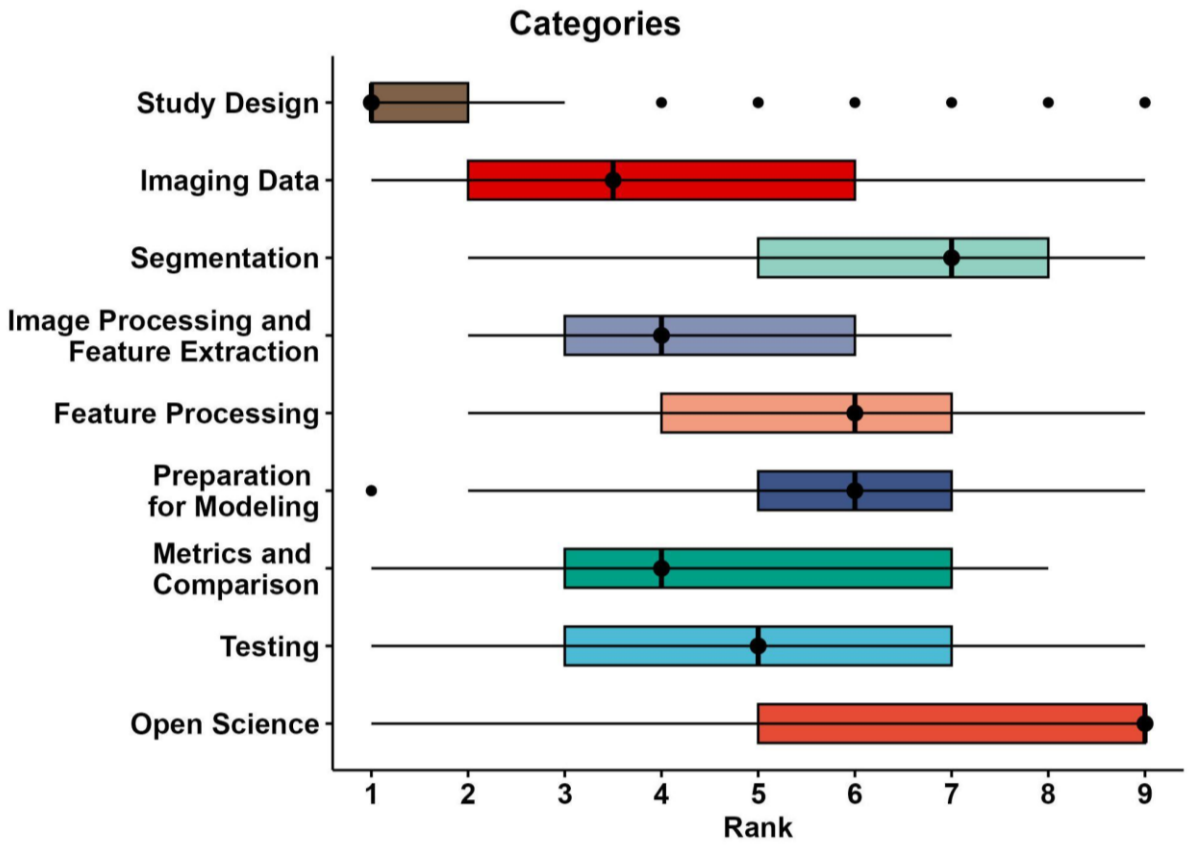

Figure S1: Box plots for rank statistics of categories. The closer a rank is to 1, the greater its importance.

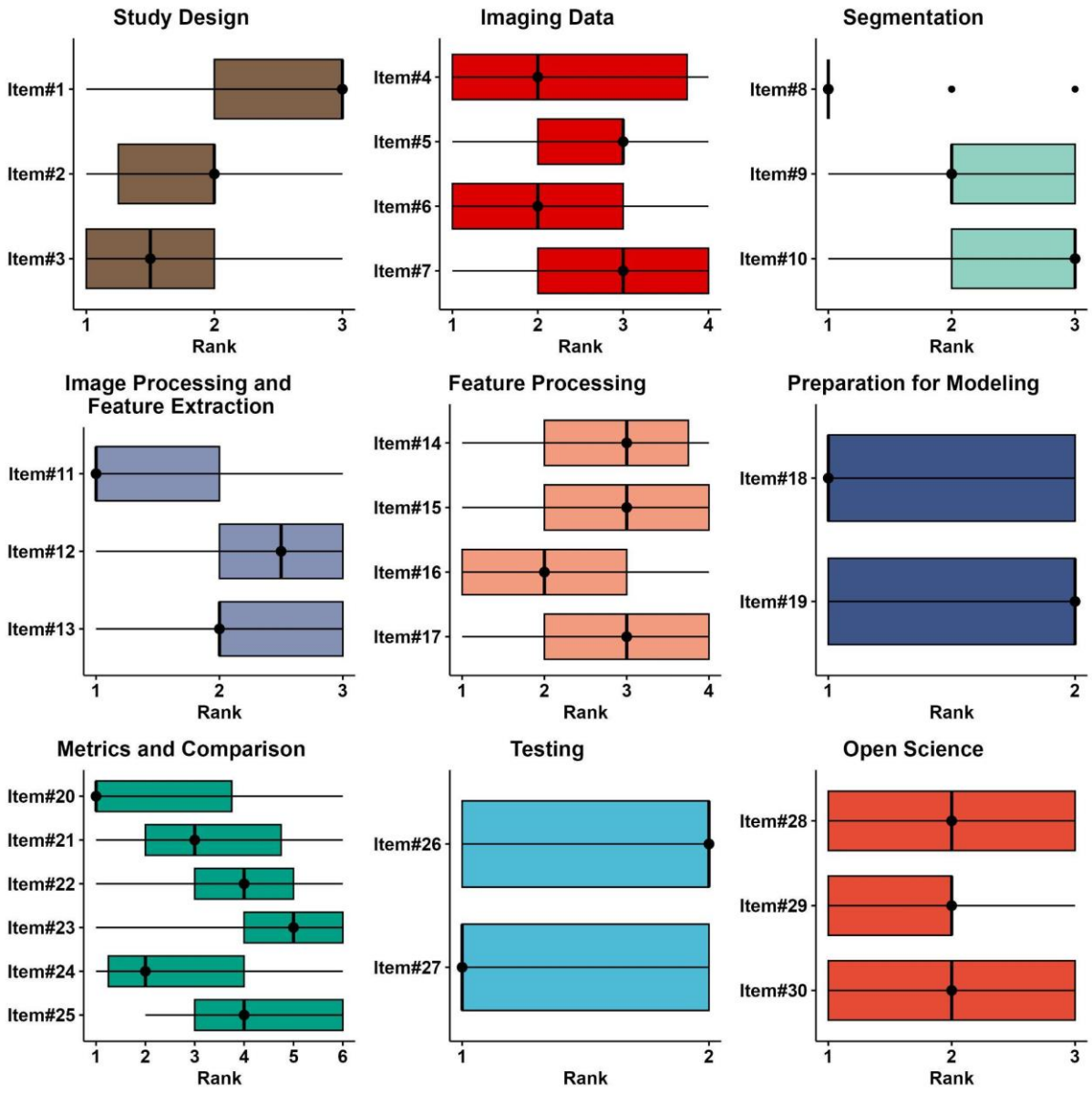

Figure S2: Box plots for rank statistics of items. The closer a rank is to 1, the greater its importance.

Supplementary file 2: Votes and ranks in Round#1, Round#2, and Round#4 of Stage#3.

Supplementary file 3: METRICS tool without explanations.

Categories	No.	Items	Weights	Score
Study Design	#1	Adherence to radiomics and/or machine learning-specific checklists or guidelines	0.037	
	#2	Eligibility criteria that describe a representative study population	0.074	
	#3	High-quality reference standard with a clear definition	0.092	
Imaging Data	#4	Multi-center	0.044	
	#5	Clinical translatability of the imaging data source for radiomics analysis	0.029	
	#6	Imaging protocol with acquisition parameters	0.044	
	#7	The interval between imaging used and reference standard	0.029	
Segmentation[1]	#8	Transparent description of segmentation methodology	0.034	
	#9	Formal evaluation of fully automated segmentation[2]	0.022	
	#10	Test set segmentation masks produced by a single reader or automated tool	0.011	
Image Processing and Feature Extraction	#11	Appropriate use of image preprocessing techniques with transparent description	0.062	
	#12	Use of standardized feature extraction software[3]	0.031	
	#13	Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement	0.041	
Feature Processing	#14	Removal of non-robust features[4]	0.020	
	#15	Removal of redundant features[4]	0.020	
	#16	Appropriateness of dimensionality compared to data size[4]	0.030	
	#17	Robustness assessment of end-to-end deep learning pipelines[5]	0.020	
Preparation for Modeling	#18	Proper data partitioning process	0.060	
	#19	Handling of confounding factors	0.030	
Metrics and Comparison	#20	Use of appropriate performance evaluation metrics for task	0.035	
	#21	Consideration of uncertainty	0.023	
	#22	Calibration assessment	0.018	
	#23	Use of uni-parametric imaging or proof of its inferiority	0.012	
	#24	Comparison with a non-radiomic approach or proof of added clinical value	0.029	
	#25	Comparison with simple or classical statistical models	0.018	
Testing	#26	Internal testing	0.037	
	#27	External testing	0.075	
Open Science	#28	Data availability	0.007	
	#29	Code availability	0.007	
	#30	Model availability	0.007	
Total METRICS score (should be given as percentage)				

[1] Conditional for studies including region/volume of interest labeling. [2] Conditional for studies using fully automated segmentation. [3] Conditional for the hand-crafted radiomics. [4] Conditional for tabular data use. [5] Conditional on the use of end-to-end deep learning. Proposed total score categories: 0-20% = very low, 21-40% = low, 41-60% = moderate, 61-80% = good, and 81-100% = excellent.

Supplementary file 4: METRICS tool with explanations.

Categories	No.	Items	Weights	Score[6]
Study Design	#1	Adherence to radiomics and/or machine learning-specific checklists or guidelines >>> Whether any guideline or checklist, e.g., CLEAR checklist, is used in designing and reporting, as appropriate for the study design (e.g., handcrafted radiomics or deep learning pipeline).	0.037	
	#2	Eligibility criteria that describe a representative study population >>> Whether inclusion and exclusion criteria are explicitly defined. These should lead to a representative study sample that matches the general population of interest for the study aim.	0.074	
	#3	High-quality reference standard with a clear definition >>> Whether the reference standard or outcome measure is representative of the current clinical practice and robust. Examples of high-quality reference standards are preferably histopathology, well-established clinical and genomic markers, the latest version of the prognostic tools, guideline-based follow-up or consensus-based expert opinions. Examples of poor quality reference standards are those based on qualitative image evaluation, images that are later used for feature extraction, or outdated versions of prognostic tools.	0.092	
Imaging Data	#4	Multi-center >>> Whether more than one institution is involved as a diagnostic imaging data source for radiomics analysis.	0.044	
	#5	Clinical translatability of the imaging data source for radiomics analysis >>> Whether the source of the radiomics data is an imaging technique that reflects established standardization approaches, such as acquisition protocol guidelines (e.g., PI-RADS specifications).	0.029	
	#6	Imaging protocol with acquisition parameters >>> Whether the image acquisition protocol is clearly reported to ensure the replicability of the method.	0.044	
	#7	The interval between imaging used and reference standard >>> Whether the time interval between the diagnostic imaging exams (used as an input for the radiomics analysis) and the outcome measure/reference standard acquisition is appropriate to validate the presence or absence of target conditions of the radiomics analysis at the moment of the diagnostic imaging exams.	0.029	
Segmentation[1]	#8	Transparent description of segmentation methodology >>> Whether the rules or the method of the segmentation are defined (e.g., margin shrinkage, peri-tumoral sampling, details of segmentation regardless of whether manual, semi-automated or automated methods are used). In the case of DL-based radiomics, the segmentation can refer to the rough delineation with bounding boxes or cropping the image around a region of interest.	0.034	
	#9	Formal evaluation of fully automated segmentation[2] >>> If a segmentation technique that does not require any sort of human intervention is used, examples of the results should be presented and a formal assessment of its accuracy compared to domain expert annotations included in the study (e.g., DICE score or Jaccard index compared with a radiologist's semantic annotation). Any intervention to the annotation in terms of volume or area should be considered as the use of a semi-automated segmentation technique. This item also applies to the use of segmentation models previously validated on other datasets.	0.022	
	#10	Test set segmentation masks produced by a single reader or automated tool >>> Whether final segmentation in the test set is produced by a single reader (manually or with a semi-automated tool) or an entirely automated tool, to better reflect clinical practice.	0.011	
Image Processing and Feature Extraction	#11	Appropriate use of image preprocessing techniques with transparent description >>> Whether preprocessing of the images is appropriately performed considering the imaging modality (e.g., gray level normalization for MRI, image registration in case of multiple contrasts or modalities) and feature extraction techniques(i.e., 2D or 3D) that are used. For	0.062	

		instance, in the case of large slice thickness (e.g., ≥5 mm), extreme upsampling (e.g., 1 x 1 x 1 mm3) of the volume might be inappropriate. In such a case, 2D feature extraction could be preferable, ensuring in-plane isotropy of the pixels. On the other hand, achieving isotropic voxel values should be targeted in 3D feature extraction, to allow for texture feature rotational invariance. Also, whether gray level discretization parameters (bin width, along with resulting gray level range, or bin count) are described in full detail. Description of different image types used (e.g., original, filtered) should also be included (e.g., high and low pass filter combinations for wavelet decomposition, sigma values for Laplacian of Gaussian edge enhancement filtering). If the image window is fixed, it should be clarified.		
	#12	Use of standardized feature extraction software[3] >>> Whether a standardized software (e.g., compliant with IBSI) was used for feature extraction, including information on the version number.	0.031	
	#13	Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement >>> Whether feature types (e.g., hand-crafted, deep features) and feature classes (for hand-crafted) are described. Also, if a default configuration statement is provided for the remaining feature extraction parameters. A file presenting the complete configuration of these settings should be included in the study materials (e.g., parameter file such as in YAML format, screenshot if a dedicated file for this is not available for the software). In the case of DL, neural network architecture along with all image operations should be described.	0.041	
Feature Processing	#14	Removal of non-robust features[4] >>> Whether unstable features are removed via test-retest, reproducibility analysis by analysis of different segmentations, or stability analysis [i.e., image perturbations]. Instability may be due to random noise introduced by manual or even automated image segmentation or exposed in a scan-rescan setting. The specific methods used should be clearly presented, with specific results for each component in multi-step feature removal pipelines.	0.020	
	#15	Removal of redundant features[4] >>> Whether dimensionality is reduced by selecting the more informative features such as with algorithm-based feature selection (e.g., LASSO coefficients, Random Forest feature importance), univariate correlation, collinearity, or variance analysis. The specific methods used should be clearly presented, with specific results for each component in multi-step feature removal pipelines.	0.020	
	#16	Appropriateness of dimensionality compared to data size[4] >>> Whether the number of instances and features in the final training data set is appropriate according to the research question and modeling algorithm. This should be demonstrated by statistical means, empirically through consistency of performance in validation and testing, or based on previous evidence in the literature.	0.030	
	#17	Robustness assessment of end-to-end deep learning pipelines[5] >>> Whether DL pipeline consistency of performance has been assessed in a test-retest setting, for example by a scan-rescan approach, use of segmentations by different readers, or stability analysis [i.e., image perturbations]. The specific methods used should be clearly presented.	0.020	
Preparation for Modeling	#18	Proper data partitioning process >>> Whether the training-validation-test data split is done at the very beginning of the analysis pipeline, prior to any processing step. Data split should be random but reproducible (e.g., fixed random seed), preferably without altering outcome variable distribution in the test set (e.g., using a stratified data split). Moreover, the data split should be on the patient level, not the scan level (i.e., different scans of the same patient should be in the same set). Proper data partitioning should guarantee that all data processing (e.g., scaling, missing value imputation, oversampling or undersampling) is done blinded to the test set data. These techniques should be exclusively fitted on training (or development) data sets and then used to transform test data at the time of inference. If a single training-validation data split is not done	0.060	

		and a resampling technique (e.g., cross-validation) is used instead, test data should always be handled separately from this.		
	#19	Handling of confounding factors >>> Whether potential confounding factors were analyzed, identified if present, and removed if necessary (e.g., if it has a strong influence on generalizability). These may include different distributions of patient characteristics (e.g., gender, lesion stage or grade) across sites or scanners.	0.030	
Metrics and Comparison	#20	Use of appropriate performance evaluation metrics for task >>> Whether appropriate accuracy metrics are reported, such as AUC for Receiver Operating Characteristics (ROC) or Precision-Recall (PRC) curves and confusion matrix-derived accuracy metrics (e.g., specificity, sensitivity, precision, F1 score) for classification tasks; MSE, RMSE, and MAE for regression tasks. For classification tasks, the confusion matrix should always be included, to allow the calculation of additional metrics. If training a DL network, loss curves should be presented.	0.035	
	#21	Consideration of uncertainty >>> Whether uncertainty measures are included in the analysis, such as 95% confidence interval (CI), standard deviation (SD), or standard error (SE). Report on methodology to derive that distribution (ie. bootstrapping with replacement, etc).	0.023	
	#22	Calibration assessment >>> Whether the final model's calibration is assessed.	0.018	
	#23	Use of uni-parametric imaging or proof of its inferiority >>> Use of a single imaging set (such as a single MRI sequence rather than multiple, or a single phase in a dynamic contrast-enhanced scan) should be preferred, as multi-parametric imaging may unnecessarily increase data dimensionality and risk of overfitting. Therefore, in the case of multi-parametric studies, uni-parametric evaluations should also be performed to justify the need for a multi-parametric approach by formally comparing their performance (e.g., DeLong's or McNemar's tests). This item is also intended to reward studies that experimentally justify the use of more complex models compared to simpler alternatives, in regard to input data type.	0.012	
	#24	Comparison with a non-radiomic approach or proof of added clinical value >>> Whether a non-radiomic method that is representative of the clinical practice is included in the analysis for comparison purposes. Non-radiomic methods might include semantic features, RADS or RECIST scoring, and simple volume or size evaluations. If no non-radiomics method is available, proof of improved diagnostic accuracy (e.g., improved performance of a radiologist assisted by the model's output) or patient outcome (e.g., decision analysis, overall survival) should be provided. In any case, the comparison should be done with an appropriate statistical method to evaluate the added practical and clinical value of the model (e.g., DeLong's test for AUC comparison, decision curve analysis for net benefit comparison, Net Reclassification Index). Furthermore, in case of multiple comparisons, multiple testing correction methods (e.g., Bonferroni) should be considered in order to reduce the false discovery rate provided that the statistical comparison is done with a frequentist approach (rather than Bayesian).	0.029	
	#25	Comparison with simple or classical statistical models >>> Whether a comparison with a simple baseline reference model (such as a Zero Rules/No Information Rate classifier) was performed. Use of machine learning methods should be justified by proof of increased performance. In any case, the comparison should be done with an appropriate statistical method (e.g., DeLong's test for AUC comparison, Net Reclassification Index). Furthermore, in case of multiple comparisons, multiple testing correction methods (e.g., Bonferroni, Benjamini–Hochberg, or Tukey) should be considered in order to reduce the false discovery rate provided that the statistical comparison is done with a frequentist approach (rather than Bayesian).	0.018	
Testing	#26	Internal testing >>> Whether the model is tested on an independent data set that is sampled from the same source as the training and/or validation sets.	0.037	

	#27	External testing >>> Whether the model is tested with independent data from other institution(s). This also applies to the studies validating the previously published models trained at another institution.	0.075	
Open Science	#28	Data availability >>> Whether any imaging, segmentation, clinical, or radiomics analysis data is shared with the public.	0.007	
	#29	Code availability >>> Whether all scripts related to automatic segmentation and/or modeling are shared with the public. These should include clear instructions for their implementation (e.g., accompanying documentation, tutorials).	0.007	
	#30	Model availability >>> Whether the final model is shared in the form of a raw model file or as a ready-to-use system. If automated segmentation was employed, the corresponding trained model should also be made available to allow replication. These should include clear instructions for their usage (e.g., accompanying documentation, tutorials).	0.007	
Total METRICS score (should be given as percentage)				

[1] Conditional for studies including region/volume of interest labeling. [2] Conditional for studies using fully automated segmentation. [3] Conditional for the hand-crafted radiomics. [4] Conditional for tabular data use. [5] Conditional on the use of end-to-end deep learning. [6] Score is simply the weight if present and 0 otherwise. Proposed total score categories: 0-20% = very low, 21-40% = low, 41-60% = moderate, 61-80% = good, and 81-100% = excellent.

Supplementary file 5: Evaluation examples for the demonstration of how to use METRICS. The table reports whether each of the example studies fulfill ("yes") or not ("no") the item requirements or if the item is not applicable to the study design ("n/a"), as appropriate. Please note that conditional item weights do not influence the maximum obtainable total score in case of non-applicability.

Items/Conditions	Weights	Cuocolo (2021)[1]	Gitto (2021)[2]	Kobayashi (2021)[3]
Study Design				
Item#1	0.037	no	no	no
Item#2	0.074	yes	yes	yes
Item#3	0.092	yes	yes	yes
Imaging Data				
Item#4	0.044	no	yes	yes
Item#5	0.029	no	yes	yes
Item#6	0.044	yes	yes	yes
Item#7	0.029	yes	yes	yes
Segmentation				
Condition#1		yes	yes	yes
Condition#2		yes	no	yes
Item#8	0.034	yes	yes	yes
Item#9	0.022	yes	n/a	yes
Item#10	0.011	yes	yes	yes
Image Processing and Feature Extraction				
Condition#3		no	yes	no
Item#11	0.062	yes	yes	yes
Item#12	0.031	n/a	yes	n/a
Item#13	0.041	no	yes	yes
Feature Processing				
Condition#4		no	yes	yes
Condition#5		yes	no	no
Item#14	0.02	n/a	yes	no
Item#15	0.02	n/a	yes	no
Item#16	0.03	n/a	yes	no

Item#17	0.02	no	n/a	n/a
Preparation for Modeling				
Item#18	0.06	yes	yes	yes
Item#19	0.03	no	no	yes
Metrics and Comparison				
Item#20	0.035	yes	yes	yes
Item#21	0.023	yes	yes	yes
Item#22	0.018	no	yes	no
Item#23	0.012	yes	yes	no
Item#24	0.029	yes	yes	no
Item#25	0.018	no	no	no
Testing				
Item#26	0.037	yes	yes	no
Item#27	0.075	no	yes	no
Open Science				
Item#28	0.007	yes	yes	yes
Item#29	0.007	no	yes	yes
Item#30	0.007	no	yes	no
METRICS score		63.70%	91.10%	68.00%
METRICS score category		**Good**	**Excellent**	**Good**

[1]Cuocolo R, Comelli A, Stefano A, Benfante V, Dahiya N, Stanzione A, Castaldo A, De Lucia DR, Yezzi A, Imbriaco M. Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset. J Magn Reson Imaging. 2021 Aug;54(2):452-459. doi: 10.1002/jmri.27585. Epub 2021 Feb 26. PMID: 33634932.

[2]Gitto S, Cuocolo R, van Langevelde K, van de Sande MAJ, Parafioriti A, Luzzati A, Imbriaco M, Sconfienza LM, Bloem JL. MRI radiomics-based machine learning classification of atypical cartilaginous tumour and grade II chondrosarcoma of long bones. EBioMedicine. 2022 Jan;75:103757. doi: 10.1016/j.ebiom.2021.103757. Epub 2021 Dec 18. PMID: 34933178; PMCID: PMC8688587.

[3]Kobayashi K, Miyake M, Takahashi M, Hamamoto R. Observing deep radiomics for the classification of glioma grades. Sci Rep. 2021 May 25;11(1):10942. doi: 10.1038/s41598-021-90555-2. PMID: 34035410; PMCID: PMC8149679.