European Radiology (2023) 33:2239–2247
https://doi.org/10.1007/s00330-022-09180-w

UROGENITAL

Ovarian imaging radiomics quality score assessment: an EuSoMII radiomics auditing group initiative

Andrea Ponsiglione[1] · Arnaldo Stanzione[1] · Gaia Spadarella[1] · Agah Baran[2] · Luca Alessandro Cappellini[3] · Kevin Groot Lipman[4] · Peter Van Ooijen[5,6] · Renato Cuocolo[7,8]

Received: 20 May 2022 / Revised: 26 July 2022 / Accepted: 18 September 2022 / Published online: 27 October 2022

Abstract

Objective To evaluate the methodological rigor of radiomics-based studies using noninvasive imaging in ovarian setting.

Methods Multiple medical literature archives (PubMed, Web of Science, and Scopus) were searched to retrieve original studies focused on computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), or positron emission tomography (PET) radiomics for ovarian disorders' assessment. Two researchers in consensus evaluated each investigation using the radiomics quality score (RQS). Subgroup analyses were performed to assess whether the total RQS varied according to first author category, study aim and topic, imaging modality, and journal quartile.

Results From a total of 531 items, 63 investigations were finally included in the analysis. The studies were greatly focused (94%) on the field of oncology, with CT representing the most used imaging technique (41%). Overall, the papers achieved a median total RQS 6 (IQR, −0.5 to 11), corresponding to a percentage of 16.7% of the maximum score (IQR, 0–30.6%). The scoring was low especially due to the lack of prospective design and formal validation of the results. At subgroup analysis, the 4 studies not focused on oncological topic showed significantly lower quality scores than the others.

Conclusions The overall methodological rigor of radiomics studies in the ovarian field is still not ideal, limiting the reproducibility of results and potential translation to clinical setting. More efforts towards a standardized methodology in the workflow are needed to allow radiomics to become a viable tool for clinical decision-making.

Key Points

- *The 63 included studies using noninvasive imaging for ovarian applications were mostly focused on oncologic topic (94%).*
- *The included investigations achieved a median total RQS 6 (IQR, −0.5 to 11), indicating poor methodological rigor.*
- *The RQS was low especially due to the lack of prospective design and formal validation of the results.*

Keywords Machine learning · Ovary · Computed tomography · Magnetic resonance imaging · Positron emission tomography

Abbreviations

CT	Computed tomography	MRI	Magnetic resonance imaging
IQR	Interquartile range	PET	Positron emission tomography

✉ Arnaldo Stanzione
arnaldo.stanzione@unina.it

[1] Department of Advanced Biomedical Sciences, University of Naples Federico II, Naples, Italy

[2] Department of Diagnostic and Interventional Radiology, University of Cologne, Cologne, Germany

[3] Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

[4] Department of Radiology, Netherlands Cancer Institute, Amsterdam, Netherlands

[5] Department of Radiation Oncology, University Medical Center Groningen, Groningen, The Netherlands

[6] Machine Learning Lab, Data Science Center in Health, University Medical Center Groningen, Groningen, the Netherlands

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

[8] Augmented Reality for Health Monitoring Laboratory (ARHeMLab), Department of Electrical Engineering and Information Technology, University of Naples "Federico II", Naples, Italy

PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-analyses
PROSPERO	International Prospective Register of Systematic Reviews
RQS	Radiomics quality score
US	Ultrasound

Introduction

Radiomics represent a new comprehensive research field combining quantitative image analysis, artificial intelligence, and medical imaging [1]. This discipline allows for the extraction of information from imaging data that could not be detectable by the human eye [2, 3]. Such data may be used to create classification models able to provide diagnostic and prognostic outputs and serve as decision-support tools [4, 5]. Several studies applied radiomics to the field of ovarian imaging, being especially focused on oncologic patients [6–8]. As a matter of fact, in the last decade, there has been an increasing clinical demand for improvements in diagnostic accuracy and patient risk stratification. In this light, predictors extracted by noninvasive imaging techniques could be worthy in several clinical scenarios, such as for classifying ovarian masses or predicting their clinical outcome [9–11]. However, radiomics applications still remain confined to academic research due to the intrinsic complexity of the method and the limited reproducibility of the numerous processes involved, especially regarding image segmentation, feature extraction, and dataset analysis [12]. Therefore, a standardized assessment of the accuracy, reproducibility as well as the clinical utility of radiomics data is needed. Aiming to respond to these demands, Lambin et al proposed the radiomics quality score (RQS), a system of metrics for the overall evaluation of the methodological validity and thoroughness of radiomics-based studies [13]. This tool has been already adopted to assess the scientific rigor of radiomics-based studies in different topics, mainly focused on oncology, such as prostate, renal, and breast cancer [14–16]. In the last years, together with the increasing clinical demand for non-invasive diagnostic techniques in the ovarian field, we have been experiencing an ever-growing number of scientific research extracting features from medical images, aimed at tumor detection and characterization or to predict prognosis and response to therapy [10, 17, 18].

Therefore, the aim of our systematic review was to evaluate the methodological rigor of investigations using computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), or ultrasound (US) for ovarian assessment on which radiomics-based models for diagnostic or prognostic purposes have been explored.

Methods

Protocol and registry

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [19]. The review protocol is registered on PROSPERO (CRD42021293541).

Search strategy

An English literature search was performed in consensus by two investigators (A.P. and A.S.) using the PubMed, Scopus, and Web of Science databases to identify articles published up to November 19th, 2021. The following search terms and their variations were used: "radiomics" AND "ovary" AND "computed tomography" OR "magnetic resonance" OR "positron emission tomography" OR "ultrasound". The detailed search string is available in the supplementary materials. After the removal of duplicates, all abstracts were assessed to remove papers other than original research (e.g., reviews, editorials, case reports), investigations not focused on the topic of interest, or not involving human subjects.

Data collection and study evaluation

The RQS was used to evaluate the methodological rigor of included papers [13]. It consists of 16 items regarding different steps in the workflow of radiomics. The summed total score ranges between −8 and 36, while the percentage is calculated on a 0–36 scale (Table 1). Two readers with previous experience in radiomics and the RQS (A.P. and G.S.) evaluated the papers in consensus. Disagreements were resolved by a third reviewer (R.C.), who reviewed the controversial items after reading the corresponding full text and discussed them with the other readers to reach a consensus. The full manuscripts were assessed to collect the following data: first author category (medical or other), study aim (diagnostic or prognostic), topic (oncology or other) and design (single-center or multi-center), imaging modality (CT, MRI, PET or US), journal quartile (first or other, based on Scopus data), segmentation strategy, machine learning algorithm, and total number of included patients.

Statistical analysis

The Shapiro-Wilk test was performed to evaluate the normality of distribution for continuous variables. These are

Table 1 Overview of radiomics quality score items and mode of the corresponding scores in the included papers

RQS checkpoint	RQS item number and name	Description and (points)	Mode
First	Item 1: Image protocol quality	Well-documented protocol (+1) AND/OR publicly available protocol (+1)	1
Second	Item 2: Multiple segmentation	Testing feature robustness to segmentation variability: e.g. different physicians/algorithms/software (+1)	0
	Item 3: Phantom study	Testing feature robustness to scanner variability: e.g. phantom studies/different vendors/scanners (+1)	0
	Item 4: Multiple time points	Testing feature robustness to temporal variability: e.g. organ movement/expansion/shrinkage (+1)	0
Third	Item 5: Feature reduction	Either feature reduction OR adjustment for multiple testing is implemented (+3); otherwise (-3)	3
	Item 6: Multivariable analysis	Non-radiomics feature are included in/considered for model building (+1)	0
	Item 7: Biological correlates	Detecting and discussing the correlation of biology and radiomics features (+1)	0
	Item 8: Cut-off analysis	Determining risk groups by either median, pre-defined cut-off, or continuous risk variable (+1)	0
	Item 9: Discrimination statistics	Discrimination statistics and its statistical significance are reported (+1); a resampling technique is also applied (+1)	1
	Item 10: Calibration statistics	Calibration statistics and its statistical significance are reported (+1); a resampling technique is also applied (+1)	0
	Item 11: Prospective design	Prospective validation of a radiomics signature in an appropriate trial (+7)	0
	Item 12: Validation	Validation is missing (-5) OR internal validation (+2) OR external validation on a single dataset from one institute (+3) OR external validation on two datasets from two distinct institutes (+4) OR validation of a previously published signature (+4) validation is based on three or more datasets from distinct institutes (+5)	-5
	Item 13: Comparison to "gold standard"	Evaluating model's agreement with/superiority to the current "gold standard" (+2)	2
	Item 14: Potential clinical application	Discussing model applicability in a clinical setting (+2)	0
	Item 15: Cost-effectiveness analysis	Performing the cost-effectiveness of the clinical application (+1)	0
	Item 16: Open science and data	Open-source scans (+1) AND/OR open-source segmentations (+1) AND/OR open-source code (+1) AND/OR open-source representative features and segmentations (+1)	0

RQS indicates radiomics quality score [13]

presented as median and interquartile range (IQR) whereas categorical data are as counts and percentages. Subgroup analyses were performed to establish whether the total RQS varied according to first author category, study aim, topic, imaging modality, and journal quartile, using the Mann-Whitney U test or Kruskal-Wallis rank test. When a paper belonged to more than one category it was counted for each within the subanalysis. Statistical analyses were performed with the "stats" (v4.1.3) R package (v4.1.3) [20]. A p value < 0.05 was considered statistically significant.

Results

Literature search

The study selection flowchart is shown in Fig. 1. The initial search identified 531 potentially eligible articles, 346 of which were duplicates. The reviewers, after the evaluation of the titles and abstracts of the remaining 185 papers studies removed 116 citations. Then, investigators blindly reviewed the full text of the remaining 69 articles, and 6 of these were excluded. Finally, 63 papers were included in the systematic review.

Study characteristics

The characteristics of the included studies are shown in the supplementary Table 1. The median population number was 105 (IQR, 67–217). Among the included papers, 49% (31/63) were published in 2021, 17% (11/63) in 2020, 13% (8/63) in 2019, 8% (5/63) in 2018, 3% (respectively 2/63 per year) in 2013 and 2017, and 2% (respectively 1/63 per year) in 2011, 2014, 2015, and 2016 (Fig. 2). The first author of most of the investigations (78%, 14/63) was a medical doctor. Radiomics analysis was conducted with diagnostic and prognostic aims respectively in 68% (43/63) and 30% (19/63) of the studies, whereas in 2% (1/63) of the investigations it was used with both intended purposes. CT was the most used imaging technique (41%, 26/63). MRI and US were respectively adopted in 34% (22/63) and 22% (14/63) of the studies, whereas in 2% (1/63) of the investigations both PET and CT were used. As

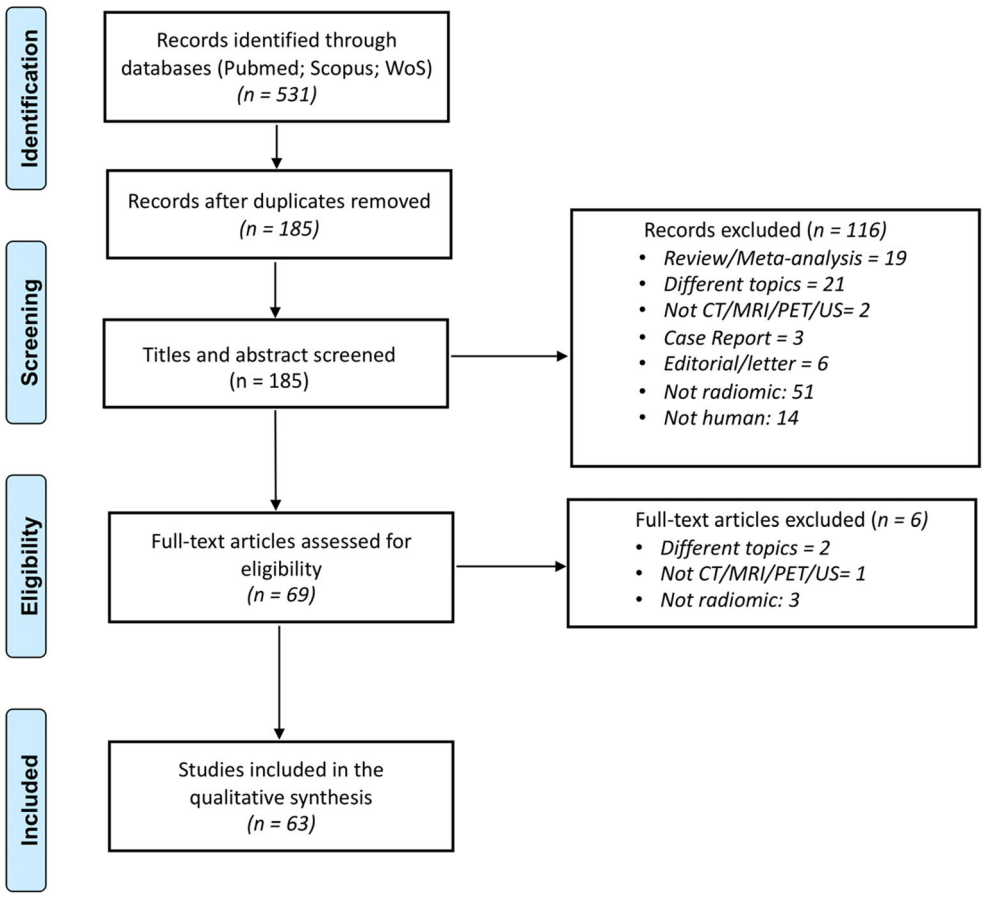

Fig. 1 Literature search and study selection flowchart

for the segmentation method, regions of interest were largely annotated manually on medical images (78%, 49/63), being three-dimensional in most cases (70%, 44/63). Finally, regarding machine learning algorithms, a high heterogeneity was found, with a minority of works adopting deep learning strategies (8%, 5/63) and the most embraced approach being overall logistic regression (40%, 25/63).

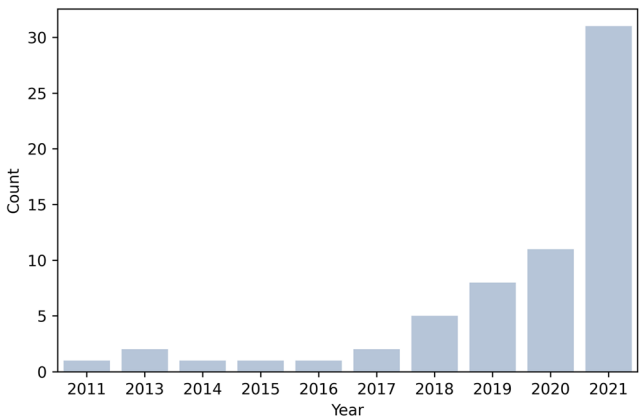

Fig. 2 Count plot showing the number of CT, MRI, PET, and US radiomics studies in ovarian setting published over the years

Study evaluation

Results are detailed in Table 2. Overall, the 63 included investigations obtained a median total RQS of 6 (IQR, −0.5 to 11), corresponding to a percentage of 16.7% (IQR, 0–30.6%) (Fig. 3). Median RQS distribution over the years is shown in Fig. 4. In regard of the first RQS checkpoint, the Authors included comprehensive information of their imaging protocol in 71% (45/63) of the corresponding investigations. In the second RQS checkpoint, features robustness to segmentation variability was assessed in 29% of the papers (18/63), while only one study (2%) performed a phantom experiment. Concerning the third RQS checkpoint, 76% (48/63) of the studies used reduction techniques to avoid feature overfitting, while less than half of the investigations (29/63) included non-radiomics features for model building. Discrimination statistics were usually performed (86%, 59/63), while only 6% (4/63) of the investigations had a prospective design. Validation, either internal or external, of the results was missing in about half of the included studies (51%, 32/63). A direct comparison between radiomics and the current gold standard was performed in 52% (33/63) of the investigations, whereas 24% (15/63) of the papers presented a formal assessment of radiomics models' clinical utility. Finally, only one study

Table 2 Radiomics quality scores for all included studies

Author (year)	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	RQS (total)	RQS (%)
Acharya (2013)	1	0	0	0	-3	0	0	0	0	0	0	2	0	2	0	0	2	5,6
Acharya (2014)	0	0	0	0	3	0	0	0	1	0	0	-5	2	0	0	0	1	2,8
Ai (2021)	1	0	0	0	3	1	0	1	2	0	0	2	2	0	0	0	12	33,3
Al-Karawi (2021)	0	0	0	0	-3	0	0	0	2	0	0	-5	2	2	0	0	-2	0
An (2021)	1	1	0	1	3	1	0	0	1	0	0	-5	2	0	0	0	5	13,9
Beer (2020)	1	1	0	0	-3	1	1	1	1	0	0	-5	0	0	0	1	-1	0
Chen (2021)	1	1	0	1	3	1	0	0	2	1	0	2	2	2	0	0	16	44,4
Chen (2021)	1	1	0	0	3	0	0	0	2	1	0	2	2	2	0	0	14	38,9
Chiappa (2021)	0	1	0	1	3	0	0	0	0	0	0	2	2	0	0	0	9	25
Chiappa (2021)	0	0	0	0	-3	1	0	0	1	0	0	-5	2	2	0	1	-1	0
Danala (2017)	1	0	0	0	3	0	0	0	1	0	0	-5	2	0	0	0	2	5,6
Faschingbauer (2013)	0	0	0	0	-3	0	0	0	2	0	0	-5	0	0	0	0	-6	0
Fathi Kazerooni (2018)	1	0	0	0	3	0	0	0	0	0	7	-5	0	0	0	0	6	16,7
He (2020)	1	1	0	0	-3	0	0	0	1	0	0	-5	0	0	0	0	-5	0
Himoto (2019)	1	0	0	0	-3	1	0	0	2	0	0	-5	0	0	0	1	-3	0
Hu (2021)	1	0	0	0	3	1	0	0	2	1	0	2	0	0	0	0	10	27,8
Jian (2021)	0	0	0	0	3	0	1	1	0	0	0	5	0	0	0	0	10	27,8
Khazendar (2015)	0	0	0	0	-3	0	0	0	2	0	0	-5	2	0	0	0	-4	0
Kiruthika (2018)	0	0	0	0	3	0	0	0	1	0	0	-5	0	0	0	0	-1	0
Kyriazi (2011)	1	1	0	1	-3	1	1	0	1	0	7	-5	2	0	0	0	7	19,4
Lee (2021)	0	0	0	0	-3	0	0	0	0	0	0	-5	0	0	0	0	-8	0
Li H (2021)	0	0	0	0	3	1	0	1	1	0	0	2	0	0	0	0	8	22,2
Li HM (2019)	1	0	0	0	-3	0	1	0	1	0	0	-5	0	0	0	1	-4	0
Li HM (2020)	1	1	0	0	-3	0	1	0	1	0	0	-5	0	0	0	0	-4	0
Li HM (2021)	1	0	0	0	3	1	0	0	1	0	0	-5	2	0	0	0	3	8,3
Li MR (2021)	1	0	0	0	3	0	0	0	1	0	0	2	2	0	0	0	9	25
Li NY (2021)	1	1	0	0	3	0	0	0	1	0	0	-5	2	0	0	1	4	11,1
Li S (2021)	1	0	0	0	3	1	0	0	1	1	0	3	2	2	0	0	14	38,9
Li YA (2020)	1	1	0	0	3	0	0	0	1	0	0	4	2	0	0	0	12	33,3
Lu H (2019)	0	0	0	0	3	1	1	0	0	0	0	3	2	0	0	1	11	30,6
Lu J (2021)	1	0	0	1	-3	0	0	1	1	0	0	-5	2	0	0	0	-2	0
Lupean (2020)	0	0	0	0	3	1	0	0	1	0	0	-5	0	0	0	0	0	0
Lupean (2020)	1	0	0	0	-3	0	0	1	1	0	0	-5	0	0	0	0	-5	0
Meier (2019)	1	0	0	0	3	0	0	0	0	0	0	-5	0	0	0	0	-1	0
Mimura (2016)	1	0	0	0	-3	0	0	0	1	0	0	-5	2	0	0	0	-4	0
Nero C (2020)	0	0	0	0	3	0	0	0	2	0	0	2	2	0	0	0	9	25
Pan (2020)	1	0	0	0	3	1	0	0	1	1	0	3	0	2	0	0	12	33,3
Park H (2021)	1	0	0	0	3	1	0	0	1	0	0	-5	0	0	0	0	1	2,8
Qi (2021)	0	0	0	0	3	1	0	0	1	1	0	2	2	2	0	0	12	33,3
Qian (2020)	1	1	0	0	3	1	1	0	2	2	0	2	2	2	0	0	17	47,2
Rizzo (2018)	1	0	1	0	3	1	0	0	1	0	0	-5	2	0	0	0	4	11,1
Seo (2021)	1	1	0	0	3	0	0	0	0	0	0	-5	0	0	0	0	0	0
Song (2021)	1	0	0	0	3	1	0	0	2	1	7	2	0	2	0	0	19	52,8
Song (2021)	1	1	0	0	3	0	0	0	1	1	7	2	2	2	0	1	21	58,3
Stefan (2021)	1	1	0	0	3	0	0	1	1	0	0	-5	0	0	0	0	2	5,6
Ștefan (2021)	1	1	0	1	3	0	1	1	1	0	0	-5	2	0	0	0	6	16,7

Table 2 (continued)

Author (year)	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	RQS (total)	RQS (%)
Vargas (2017)	0	0	0	0	3	0	0	0	0	0	0	-5	0	0	0	0	-2	0
Veeraraghavan (2020)	0	0	0	0	3	0	1	0	1	0	0	-5	2	0	0	0	2	5,6
Wang R (2021)	1	1	0	0	3	1	0	0	2	0	0	2	0	0	0	1	11	30,6
Wang S (2019)	1	0	0	0	3	1	0	0	1	0	0	3	2	0	0	0	11	30,6
Wang X (2021)	1	0	0	0	3	1	0	0	2	1	0	2	0	0	0	0	10	27,8
Wei C (2020)	1	0	0	0	3	1	0	0	2	0	0	-5	2	0	0	0	4	11,1
Wei W (2018)	0	0	0	0	3	0	0	0	1	0	0	2	0	0	0	0	6	16,7
Wei W (2019)	1	0	0	0	3	1	0	0	2	1	0	3	2	0	0	0	13	36,1
Yao (2021)	1	1	0	0	3	1	0	1	1	0	0	2	2	0	0	0	12	33,3
Ye (2021)	1	0	0	0	3	1	1	0	1	0	0	2	2	0	0	0	11	30,6
Yi (2021)	1	0	0	0	3	1	0	0	1	1	0	2	0	2	0	0	11	30,6
Xu XP (2021)	1	1	0	0	3	0	0	0	1	0	0	2	0	0	0	0	8	22,2
Yu XY (2021)	1	0	0	0	3	1	0	0	1	1	0	2	0	2	0	0	11	30,6
Zargari (2018)	1	0	0	0	3	0	0	0	2	0	0	-5	0	0	0	0	1	2,8
Zhang H (2019)	1	0	0	1	3	0	0	1	1	0	0	2	2	2	0	0	13	36,1
Zhang L (2019)	0	0	0	0	3	0	0	0	1	0	0	2	0	0	1	0	7	19,4
Zhu (2021)	1	0	0	0	3	1	0	1	2	2	0	2	2	2	0	0	16	44,4

The total score ranges from −8 to 36 and the percentage was based on the maximum value of 36

RQS indicates radiomics quality score [13]

(2%) performed a cost-effectiveness analysis and 8 studies (13%) made their code and/or data publicly available.

Subgroup analysis

Table 3 shows the results of the subgroup analysis according to first author category, study aim and topic, imaging modality, and journal quartile. The 4 studies not based on the oncologic topic received significantly lower scores than the others ($p = 0.01$). Conversely, no statistically significant differences were found between papers according to first author category ($p = 0.75$), study aim ($p = 0.9$), and imaging modality ($p = 0.48$). Moreover, in studies published in first quartiles journals, the total RQS percentage was higher than that of investigations published in lower quartiles journals (median 19.4 vs. 8.3), but this difference was not statistically significant ($p = 0.09$).

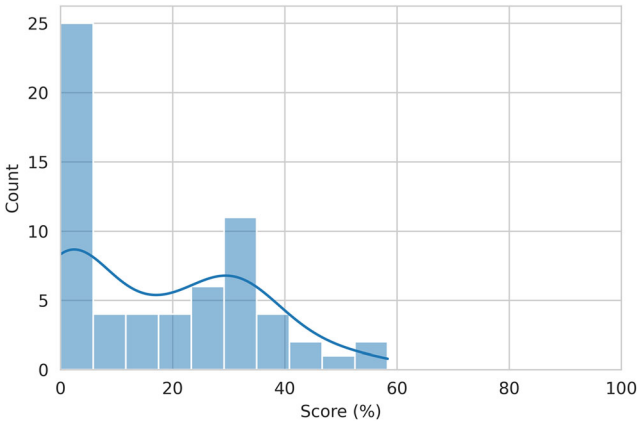

Fig. 3 Distribution of median total RQS percentage score of investigations included in our review. This is presented both as a histogram (bars) and its corresponding density function (line)

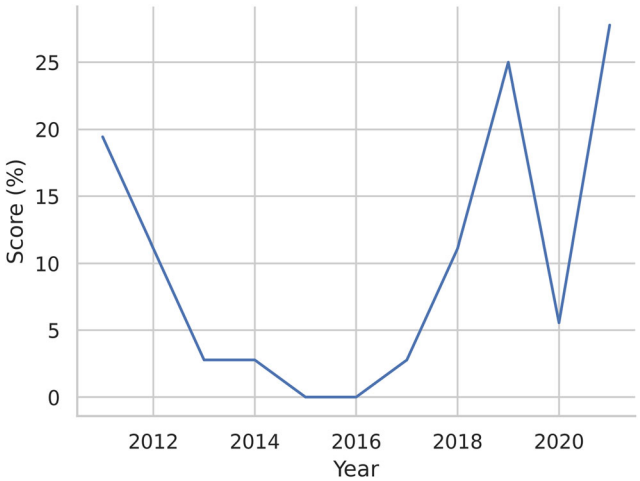

Fig. 4 Line plot of median total RQS percentage in relation to the publication year

Table 3 Subgroup analysis according to first author category, study aim, topic and design, imaging modality, and journal quartile

Group	Studies (*n*)	RQS percentage	*p* value
First author category			0.75
Medical	49	19.4 (0–30.6)	
Other	14	13.9 (3.5–29.9)	
Study aim			0.9
Diagnostic	44	18 (0–31.2)	
Prognostic	20	13.9 (2.1–30.6)	
Study topic			0.01
Oncology	59	19.4 (2.8–30.6)	
Other	4	0 (0–0)	
Study design			0.08
Single-center	52	12.5 (0–30.6)	
Multi-center	11	30.6 (15.3–33.3)	
Imaging modality			0.48
CT	27	22.2 (2.8–31.9)	
MRI	22	18.1 (0–30.6)	
PET	1	27.8	
US	14	5.6 (0–23.6)	
Journal quartile			0.09
First	31	19.4 (5.6–30.6)	
Others	29	8.3 (0–28.5)	

Values are expressed as number or median (interquartile range)

RQS indicates radiomics quality score [13]

Discussion

Several radiomics-based investigations have been performed either with a diagnostic or prognostic aim for various ovarian pathologies, being especially focused on oncologic topics [9, 10, 17, 21–23]. Of note, CT was the most used diagnostic technique, even if it does not represent the imaging of choice in clinical routine.

However, despite the promising results, their translation to clinical routine still appears as a distant goal. This is particularly due to the complexity of the method and the low reproducibility of the several processes involved [3, 24, 25].

In our systematic review, the overall methodological rigor of ovarian radiomics investigations either with CT, MRI, PET, or US resulted to be unsatisfactory, with a median RQS total score of 6, corresponding to 16.7% of the maximum possible rating. Moreover, our results do not represent an exception. Indeed, previous studies highlighted that the overall methodological quality of radiomics studies is heterogenous and lower than desirable in various fields of medical imaging [14–16, 26–29]. In particular, Granzier et al for breast cancer, Ugga et al for meningioma, and Ursprung et al for renal cell carcinoma reported in their systematic reviews low average or median total RQS percentage, respectively of 11.8%, 19%, and 9.4% [15, 16, 27]. The trend of RQS over the years is fairly inconsistent, even though the increase in 2021 could represent a positive sign for the future. Considering that almost half of the investigations (31/63) were published in 2021, and that "how to" guides have been recently published aiming to standardize practice in radiomics, we could be cautiously optimistic that the tendency will be towards an overall improvement [2, 30]. A greater focus on this issue by journals and editors could also assist in improving the quality and diagnostic efficacy level of these types of investigations, in turn facilitating their introduction into clinical practice [21, 31].

Our systematic review has pointed out several issues in the included studies that will necessarily have to be solved in future radiomics-based research in the field of ovarian imaging. In detail, a comprehensive documentation of the imaging protocol is still lacking in some investigations; however, the corresponding item seems to have been better scored compared to the studies focused on different topics [14, 15]. Another major issue is represented by the overall lack of testing features robustness either to segmentation, scanner, or temporal variability. This could be at least partly due to the predominant retrospective design of the included investigations, which also represents a significant limitation. Segmentation definitely represents a crucial step in radiomics workflow as data are extracted from the segmented regions of interest. Of note, in the included papers, regions of interest were mostly annotated manually on medical images. However, the "ideal" segmentation strategy is still debated [32]. Some authors employ manual segmentation by expert readers as the ground truth, but this method can be highly time-consuming [33]. Automatic segmentation of the whole volume of interest could overcome this issue, but intensive user correction might be necessary for inhomogeneous lesions [34].

Moreover, as patient numbers are limited and countless radiomics features can be extracted, it is fundamental to reduce feature number, especially removing those poorly reproducible that could affect algorithm performance [3, 25, 35]. On a positive note, 76% of the reviewed papers performed feature reduction, thus lowering the risk of overfitting. Furthermore, even if the need of validating radiomics has been extensively discussed [36], less than half of the included investigations conducted a validation, either internal or external, of their results. However, the scores of this specific item obtained in the ovarian field are slightly better than those reported for prostate as well as breast cancer radiomics-based research [14, 15].

Open science remains a major issue also in ovarian setting, with 87% of the included papers not sharing their data and/or code. Publicly available datasets, such as the Cancer Genome Atlas Program and National Cancer Institute Imaging Data Commons, may represent a possible solution, helping to increase knowledge regarding the impact of varying factors in radiomics [37–39]. Of note, none of the included studies used public image protocols.

Subgroup analyses pointed out that the papers focused on the oncologic topic showed significantly higher RQS total scores than the non-oncologic ones. However, it should be taken into account that most of the studies (94%) aimed to assess radiomics performance in the field of oncology. Moreover, even if not reaching statistical significance, papers published in first quartile journals showed higher median RQS percentage than those published in the other quartile ones, possibly due to the greater demand of the high-ranking journals in terms of methodological rigor, especially regarding validation of the results.

Of note, the RQS may not represent the perfect tool to evaluate the methodological quality of a radiomics study. For example, due to the nature of its items, the RQS might penalize studies using deep learning algorithms, that are at risk of getting lower scores for lacking feature selection or multiple segmentations (which are not necessarily limitations in deep learning studies) [40]. Furthermore, the relative weight of some items might be unbalanced and penalize those preliminary, exploratory studies that were retrospectively designed but needed as a first ground on which stronger evidence must be built. Finally, it should be considered that generalizability is one of the key issues for the clinical translation of radiomics models but needs external independent validation that was rare in this experience (11%, 7/63). To increase the scientific merit and methodological robustness of radiomics studies, researchers might want to focus on validating previously published radiomics signatures using their datasets as independent validation cohorts rather than building new models. However, open science represents a necessary prerequisite to achieve this goal.

Our study suffers from some limitations that should be acknowledged. First of all, inter-reader agreement of RQS assessment was not explored; however, the two readers evaluating the papers had previous experience with this system of metrics [14, 28]. Second, since the field of radiomics is constantly evolving, even in terms of nomenclature, potential eligible investigations could have been missed. Finally, some included studies were published before the introduction of the RQS.

In conclusion, the overall scientific rigor of ovarian radiomics studies was unsatisfactory, resulting particularly lacking in terms of features reproducibility and formal validation of the results. More efforts towards a standardized methodology in the pipeline are needed to allow radiomics to become a viable tool for clinical decision-making.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09180-w.

Acknowledgements The European Society of Medical Imaging Informatics supports the Radiomics Auditing Group Initiative with the aim to evaluate and improve the quality of radiomics studies.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement. The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Prof. Renato Cuocolo, MD, PhD.

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

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was not required for this study because it is a systematic literature review.

Ethical approval Institutional Review Board approval was not required because it is a systematic literature review.

Methodology
- systematic review
- performed at multiple institutions

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. van Timmeren JE, Cester D, Tanadini-Lang S et al (2020) Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 11:91. https://doi.org/10.1186/s13244-020-00887-2
3. Rogers W, Thulasi Seetha S, Refaee TAG et al (2020) Radiomics: from qualitative to quantitative imaging. Br J Radiol 93:20190948. https://doi.org/10.1259/bjr.20190948
4. Chu H, Liu Z, Liang W et al (2021) Radiomics using CT images for preoperative prediction of futile resection in intrahepatic cholangiocarcinoma. Eur Radiol 31:2368–2376. https://doi.org/10.1007/s00330-020-07250-5
5. Sollini M, Cozzi L, Chiti A, Kirienko M (2018) Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: where do we stand? Eur J Radiol 99:1–8. https://doi.org/10.1016/j.ejrad.2017.12.004

6. Wang T, Wang H, Wang Y et al (2022) MR-based radiomics-clinical nomogram in epithelial ovarian tumor prognosis prediction: tumor body texture analysis across various acquisition protocols. J Ovarian Res 15:6. https://doi.org/10.1186/s13048-021-00941-7
7. Arezzo F, Loizzi V, La Forgia D et al (2021) Radiomics analysis in ovarian cancer: a narrative review. Appl Sci 11:7833. https://doi.org/10.3390/app11177833
8. Song X, Ren J-L, Zhao D et al (2021) Radiomics derived from dynamic contrast-enhanced MRI pharmacokinetic protocol features: the value of precision diagnosis ovarian neoplasms. Eur Radiol 31:368–378. https://doi.org/10.1007/s00330-020-07112-0
9. Zhang H, Mao Y, Chen X et al (2019) Magnetic resonance imaging radiomics in categorizing ovarian masses and predicting clinical outcome: a preliminary study. Eur Radiol 29:3358–3371. https://doi.org/10.1007/s00330-019-06124-9
10. An H, Wang Y, Wong EMF et al (2021) CT texture analysis in histological classification of epithelial ovarian carcinoma. Eur Radiol 31:5050–5058. https://doi.org/10.1007/s00330-020-07565-3
11. Beer L, Sahin H, Bateman NW et al (2020) Integration of proteomics with CT-based qualitative and radiomic features in high-grade serous ovarian cancer patients: an exploratory analysis. Eur Radiol 30:4306–4316. https://doi.org/10.1007/s00330-020-06755-3
12. Recht MP, Dewey M, Dreyer K et al (2020) Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol 30:3576–3584. https://doi.org/10.1007/s00330-020-06672-5
13. 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
14. Stanzione A, Gambardella M, Cuocolo R et al (2020) Prostate MRI radiomics: a systematic review and radiomic quality score assessment. Eur J Radiol 129:109095. https://doi.org/10.1016/j.ejrad.2020.109095
15. Ursprung S, Beer L, Bruining A et al (2020) Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma—a systematic review and meta-analysis. Eur Radiol 30:3558–3566. https://doi.org/10.1007/s00330-020-06666-3
16. Granzier RWY, van Nijnatten TJA, Woodruff HC et al (2019) Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: a systematic review. Eur J Radiol 121:108736. https://doi.org/10.1016/j.ejrad.2019.108736
17. Song X-L, Ren J-L, Yao T-Y et al (2021) Radiomics based on multisequence magnetic resonance imaging for the preoperative prediction of peritoneal metastasis in ovarian cancer. Eur Radiol 31:8438–8446. https://doi.org/10.1007/s00330-021-08004-7
18. Yao F, Ding J, Hu Z et al (2021) Ultrasound-based radiomics score: a potential biomarker for the prediction of progression-free survival in ovarian epithelial cancer. Abdom Radiol (NY) 46:4936–4945. https://doi.org/10.1007/s00261-021-03163-z
19. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339:b2535–b2535. https://doi.org/10.1136/bmj.b2535
20. R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
21. 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
22. Ştefan R-A, Ştefan P-A, Mihu CM et al (2021) Ultrasonography in the differentiation of endometriomas from hemorrhagic ovarian cysts: the role of texture analysis. J Pers Med 11:611. https://doi.org/10.3390/jpm11070611
23. Seo M, Choi MH, Lee YJ et al (2021) Evaluating the added benefit of CT texture analysis on conventional CT analysis to differentiate benign ovarian cysts. Diagnostic Interv Radiol 27:460–468. https://doi.org/10.5152/dir.2021.20225
24. Sollini M, Antunovic L, Chiti A, Kirienko M (2019) Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging 46:2656–2672. https://doi.org/10.1007/s00259-019-04372-x
25. Lennartz S, O'Shea A, Parakh A et al (2022) Robustness of dual-energy CT-derived radiomic features across three different scanner types. Eur Radiol 32:1959–1970. https://doi.org/10.1007/s00330-021-08249-2
26. Park JE, Kim D, Kim HS et al (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 30:523–536. https://doi.org/10.1007/s00330-019-06360-z
27. Ugga L, Perillo T, Cuocolo R et al (2021) Meningioma MRI radiomics and machine learning: systematic review, quality score assessment, and meta-analysis. Neuroradiology. https://doi.org/10.1007/s00234-021-02668-0
28. Spadarella G, Calareso G, Garanzini E et al (2021) MRI based radiomics in nasopharyngeal cancer: systematic review and perspectives using radiomic quality score (RQS) assessment. Eur J Radiol 140:109744. https://doi.org/10.1016/j.ejrad.2021.109744
29. Ponsiglione A, Stanzione A, Cuocolo R et al (2022) Cardiac CT and MRI radiomics: systematic review of the literature and radiomics quality score assessment. Eur Radiol 32:2629–2638. https://doi.org/10.1007/s00330-021-08375-x
30. Shur JD, Doran SJ, Kumar S et al (2021) Radiomics in oncology: a practical guide. Radiographics 41:1717–1732. https://doi.org/10.1148/rg.2021210037
31. Pinto dos Santos D (2022) Radiomics in endometrial cancer and beyond - a perspective from the editors of the EJR. Eur J Radiol 150:110266. https://doi.org/10.1016/j.ejrad.2022.110266
32. Rizzo S, Botta F, Raimondi S et al (2018) Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp 2:36. https://doi.org/10.1186/s41747-018-0068-z
33. Kumar V, Gu Y, Basu S et al (2012) Radiomics: the process and the challenges. Magn Reson Imaging 30:1234–1248. https://doi.org/10.1016/j.mri.2012.06.010
34. Sofka M, Wetzl J, Birkbeck N et al (2011) Multi-stage learning for robust lung segmentation in challenging CT volumes. Med Image Comput Comput Assist Interv 14:667–674. https://doi.org/10.1007/978-3-642-23626-6_82
35. Gitto S, Cuocolo R, Emili I et al (2021) Effects of interobserver variability on 2D and 3D CT- and MRI-based texture feature reproducibility of cartilaginous bone tumors. J Digit Imaging 34:820–832. https://doi.org/10.1007/s10278-021-00498-3
36. European Society of Radiology (ESR) (2020) ESR statement on the validation of imaging biomarkers. Insights Imaging 11:76. https://doi.org/10.1186/s13244-020-00872-9
37. Oakden-Rayner L (2019) Exploring large scale public medical image datasets. https://doi.org/10.48550/arXiv.1907.12720
38. The Cancer Genome Atlas Program - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
39. Fedorov A, Longabaugh WJR, Pot D et al (2021) NCI imaging data commons. Cancer Res 81:4188–4193. https://doi.org/10.1158/0008-5472.CAN-21-0950
40. Guiot J, Vaidyanathan A, Deprez L et al (2022) A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev 42:426–440. https://doi.org/10.1002/med.21846