Review Article

# Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review

**Cindy Xue[1,2], Jing Yuan[1]^, Gladys G. Lo[3], Amy T. Y. Chang[4], Darren M. C. Poon[4], Oi Lei Wong[1], Yihang Zhou[1], Winnie C. W. Chu[2]^**

[1]Medical Physics and Research Department, Hong Kong Sanatorium & Hospital, Happy Valley, Hong Kong, China; [2]Department of Imaging and Interventional Radiology, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China; [3]Department of Diagnostic & Interventional Radiology, Hong Kong Sanatorium & Hospital, Happy Valley, Hong Kong, China; [4]Comprehensive Oncology Centre, Hong Kong Sanatorium & Hospital, Happy Valley, Hong Kong, China

*Contributions:* (I) Conception and design: J Yuan, C Xue; (II) Administrative support: J Yuan, C Xue, OL Wong, Y Zhou; (III) Provision of study materials or patients: J Yuan, C Xue, OL Wong; (IV) Collection and assembly of data: J Yuan, C Xue, OL Wong, Y Zhou; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* Jing Yuan, PhD. 8/F, Li Shu Fan Block, Hong Kong Sanatorium & Hospital, 2 Village Road, Happy Valley, Hong Kong, China. Email: jyuanbwh@gmail.com.

**Abstract:** Radiomics research is rapidly growing in recent years, but more concerns on radiomics reliability are also raised. This review attempts to update and overview the current status of radiomics reliability research in the ever expanding medical literature from the perspective of a single reliability metric of intraclass correlation coefficient (ICC). To conduct this systematic review, Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed. After literature search and selection, a total of 481 radiomics studies using CT, PET, or MRI, covering a wide range of subject and disease types, were included for review. In these highly heterogeneous studies, feature reliability to image segmentation was much more investigated than reliability to other factors, such as image acquisition, reconstruction, post-processing, and feature quantification. The reported ICCs also suggested high radiomics feature reliability to image segmentation. Image acquisition was found to introduce much more feature variability than image segmentation, in particular for MRI, based on the reported ICC values. Image post-processing and feature quantification yielded different levels of radiomics reliability and might be used to mitigate image acquisition-induced variability. Some common flaws and pitfalls in ICC use were identified, and suggestions on better ICC use were given. Due to the extremely high study heterogeneities and possible risks of bias, the degree of radiomics feature reliability that has been achieved could not yet be safely synthesized or derived in this review. More future researches on radiomics reliability are warranted.

**Keywords:** Radiomics; reliability; intraclass correlation coefficient (ICC); quantitative imaging; oncology

Submitted Jan 22, 2021. Accepted for publication May 17, 2021.
doi: 10.21037/qims-21-86
**View this article at:** https://dx.doi.org/10.21037/qims-21-86

^ ORCID: Jing Yuan, 0000-0001-8112-3608; Winnie C. W. Chu, 0000-0003-4962-4132.

## Introduction

Radiomics has become one of the most popular research areas in medical imaging, in particular for clinical oncology, since its first introduction by Lambin *et al.* in 2012 (1). According to Gillies *et al.*, radiomics is defined as "the conversion of images to higher dimensional data and the subsequent mining of these data for improved decision support" (2). These higher dimension data are normally understood as the information contained in a large number of quantitative radiomics features derived from the original or transformed medical images, which are usually artificially engineered with mathematical definition and have continuous values. By utilizing the radiomics models built on the selected radiomics features, radiomics promises to increase diagnosis accuracy and precision, assessment of prognosis, and therapy response prediction for different clinical applications, bridging between medical imaging and personalized medicine (3,4). A tremendous number of papers on radiomics have been published in recent years (5). However, despite the promising results reported, the broad validity, and generality of radiomics are still much hindered by the concerns on its reliability (6-11). Variability and uncertainty of radiomics can be introduced in many procedures of its complicated workflow. These procedures include but not limited to imaging hardware configuration, patient setup, image acquisition, reconstruction, image post-processing (filtering, segmentation and registration, etc.), radiomics feature quantification (such as feature definition, calculation setting like image discretization, software implementation, calculation result harmonization, etc.), and radiomics modeling.

The term reliability is commonly used with other terms like agreement, repeatability, stability, reproducibility, accuracy/precision, and robustness in varying degrees of consistency in the medical literature. In this study, a general and mathematically expressible definition of reliability (R) is adopted to be the extent to which measurements can be replicated. It is expressed as the ratio of true (error-free) variance ($\sigma T^2$) over true variance plus error variance ($\sigma_E^2$), i.e., $R=\sigma_T^2/(\sigma_T^2 + \sigma_E^2)$. This definition of reliability is compliant with the classical definition of intraclass correlation coefficient (ICC), using the between-subject variance in the trait of interest to represent the true variance since it cannot be directly measured in reality.

$$ICC = \frac{between\ subject\ variance}{between\ subject\ variance + within\ subject\ measurement\ variance} \quad [1]$$

ICC is one of the most widely adopted reliability indexes based on the analysis of variance (ANOVA) in medical literature (12). ICC is applicable for all radiomics features that have continuous values. In addition, ICC is a ratio index ranged from 0 to 1, so it is useful in the cross-study reliability comparison. For these reasons, ICC is chosen in this study as the single statistical metric for radiomics feature reliability assessment. We adopted the ICC forms in McGraw and Wong Convention, including three components of model (one-way or two-way, random-effects or mixed-effects), type (single or multiple measurements/raters), and definition (absolute agreement or consistency), following the guideline proposed by Koo *et al.* (13).

Reliability of radiomics has to be carefully and rigorously measured and assessed prior to its real clinical deployment, but generic radiomics reliability is still not yet fully explored, so not well known. An excellent systematic review on repeatability and reproducibility of radiomics features was published in 2018 (14), in which the qualitative synthesis on 41 studies revealed the status of radiomics reliability research until April 2017. Since then, the status of radiomics reliability research has not been timely updated in the pace of an ever fast increasing number of radiomics publications. Few strongly evident consensus has been reached and well-acknowledged so far.

Thus, this review attempts to serve multi-fold purposes: (I) to have a timely updated overview on the current status of radiomics reliability research, mainly from the perspective of ICC use in the medical literature; (II) to survey what ICC was used for, and what were the major findings of radiomics reliability as revealed by the reported ICCs; (III) to critically review how ICC was used, reported and interpreted; (IV) to give some suggestions on ICC use to mitigate the flaws and pitfalls, if applicable, so as to improve radiomics reliability assessment for future studies.

## Methods

### *Systematic search strategy*

The major research question for the literature search was described as: "What are the known radiomics studies that used ICC as a radiomics feature reliability index, and reported the quantitative ICC results (as either major or secondary outcome)". Thus, a comprehensive literature search was conducted by two authors (JY and CX) to identify the relevant published studies in the database of MEDLINE/PubMed (National Center for Biotechnology Information, NCBI),

 

from 1 January 2012 to 8 December 2020 (ePub date).

A combination of the following terms and their common variations: "CT/PET/MRI", "radiomics/radiomic/texture analysis/quantitative (heterogeneity) feature", "ICC/intraclass correlation" were comprehensively used for literature search. Imaging modalities other than CT, PET, and MRI, such as ultrasound, X-ray, cone-beam CT, and Megavoltage CT, were not included in this search due to their relative minority and immaturity in the radiomics research. Image analysis based solely on the gray level histogram, i.e., histogram analysis, does not provide any voxel positional/distributive information on the images, so it was not included.

### *Study selection*

Only full-text journal or conference articles written in English were eligible and included. Conference abstracts, case reports, (systematic) reviews, editorials/commentaries, expert opinion papers, and non-English papers were excluded from selection.

After article type exclusion, all publications that involved the use of ICC for feature reliability assessment were identified through full text (and Supplementary materials if needed) examination in the searching results. If a study mentioned the ICC use in the method but reported no ICC results, it was also excluded.

Three authors (JY, CX, and OW) worked jointly on the study selection procedures as described above. Disagreements were resolved by consensus. Reasons for exclusion were documented.

### *Data extraction*

Four authors (JY, CX, OW, and YZ) jointly performed record extraction. The study information on publication date, imaging modality, study design, study subject (phantom, animal or human), organ, disease, radiomics feature type/number was extracted. In terms of ICC use and reporting, the purpose of using ICC, the sample size for ICC calculation, ICC form, ICC reporting format, and major ICC results were extracted. Despite the high heterogeneity of ICC result reporting in different studies, we attempted to extract, synthesize, or harmonize the ICC results in the form of satisfactory feature rate (SFR), i.e., the percent of features showing satisfactory (determined by excellent, good, or other ICC criteria in each individual publication) ICC in the total investigated features, as much as possible. In this way, cross-study ICC comparison might become feasible to some extent. Radiomics quality score (RQS) of the extracted studies were not individually appraised since RQS might not be applicable for many of these studies because they were not completely clinical application studies (4). The quality of ICC use and reporting was not scored either. QUADAS-2 was not applied for study appraisal either since the diagnostic accuracy was not the common purpose of the included radiomics studies (15).

### *Outcomes and prioritizations*

The primary outcome of interest in this review was radiomics feature reliability in different aspects as assessed by ICC. Quality of ICC use and reporting was the secondary outcome. Other statistic metrics used in combination with ICC were only noted but not further analyzed. The outcome was not prioritized on specific imaging modality or disease type.

### *Risk of bias analysis*

Two authors (CX and JY) jointly assessed the possible risk of bias in the included studies from the extracted study information with consensus in the following perspectives. (I) study characteristics such as the study design (retrospective or prospective), cohort, sample size and feature number; (II) appropriateness of methodology, and sufficiency of method description and disclosure, such as the details of imaging acquisition, post-processing, segmentation, as well as feature definition (standardization) and quantification; (III) the quality of ICC use and reporting, such as the ICC form selection, confidence interval reporting, threshold values, and interpretation.

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement (16) was not applied for two major reasons. First, it was not applicable for many studies because the investigation of radiomics feature reliability did not necessarily lead to the final prognosis or diagnosis performance report. Second, the clinical purposes were beyond the scope of this review, and they were also highly heterogeneous, including but not limited to prognosis or diagnosis.

## Results

### *Literature search and selection*

The PubMed search yielded 2,596 records. Records were

 

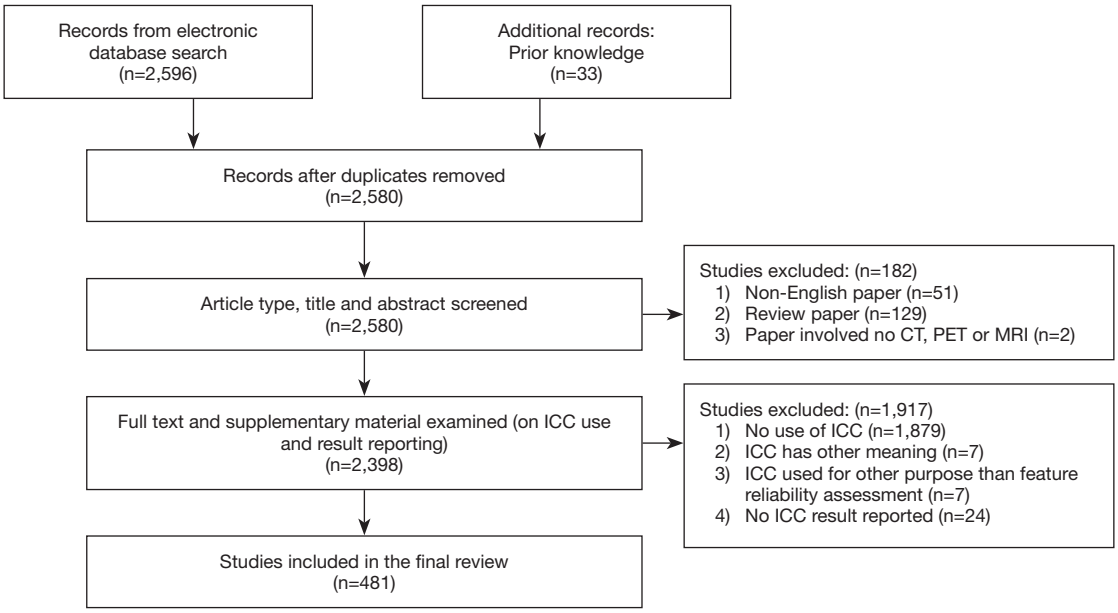


**Figure 1** Flowchart of the study selection process.

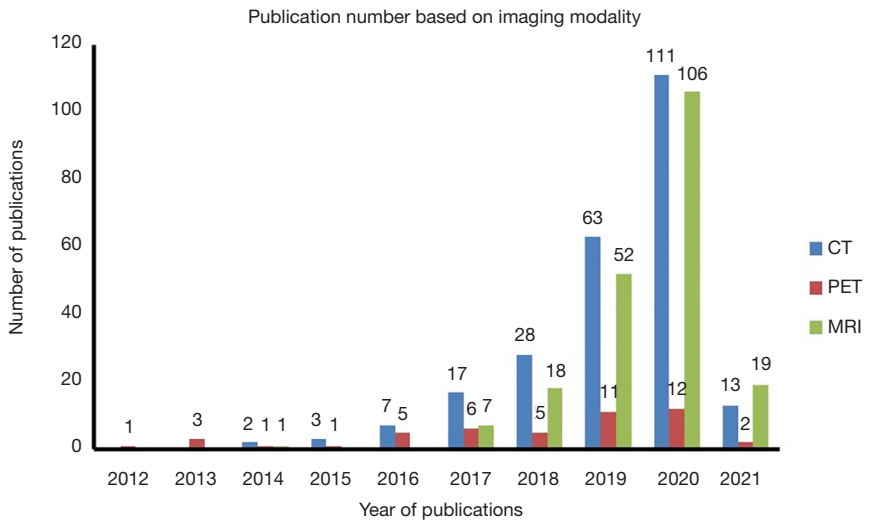


**Figure 2** Publication number based on imaging modality in recent years.

reduced to 2,580 after duplicates were removed. The subsequent article type, title, and abstract screening excluded 182 records. Then the remaining 2,398 records underwent full text (and Supplementary materials) examination, and 1,917 records were further excluded. Finally, 481 studies were included in this systematic review. The selection process is illustrated in *Figure 1*.

### *Statistics of the included publications*

*Figure 2* shows the increasing number of published radiomics studies in which ICC was used for radiomics reliability assessment from 2012 to 2021. In recent years since 2017, MRI articles show a much faster increase than CT, and PET.

In terms of publication number, CT, PET and MRI

 

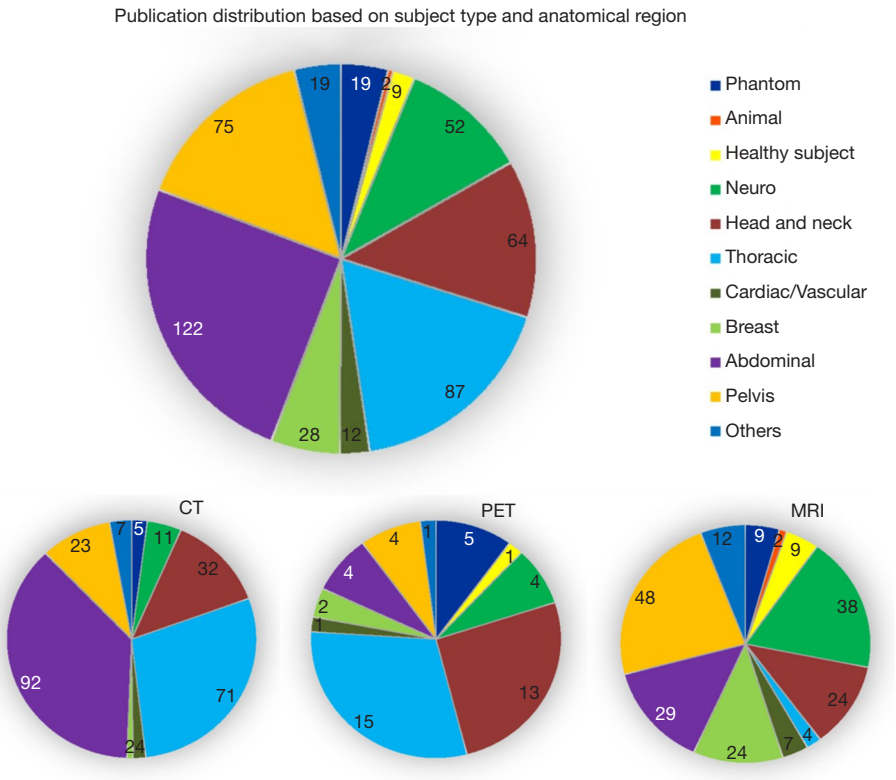


**Figure 3** Study number based on subject type and anatomical region. The study number is counted repeatedly if multiple subject types or anatomical regions were involved in a study.

account for 50.73% (n=244), 9.77% (n=47), and 42.20% (n=203) of the total publications (N=481), respectively. Note that some publications involved multi-modality radiomics, so the sum was slightly over 100%.

In the included articles, 18.09% (87/481) and 81.91% (394/481) of them assessed reliability using ICC as the major study outcome and secondary (partial) outcome, respectively. 5.82% (n=28) and 94.18% (n=453) studies were prospective and retrospective (including those studies using prospectively acquired imaging data for clinical purposes other than radiomics) in nature.

The numbers of studies based on subject type and anatomical region are illustrated in *Figure 3*.

Clinical oncology is the most common clinical application in these studies, accounting for 86.69% (417/481) of the total publications. Lung cancer is the most common type, followed by head and neck cancer and neuro-oncology. Imaging modality reflects variation in predominant type of cancer, as PET and CT studies showed a higher proportion of lung and head and neck patients, while MRI studies comprised more neuro-oncology, breast cancer, cervical cancer, and prostate cancer. The publication distribution based on cancer types, and imaging modalities is illustrated in *Figure 4*.

### *The characteristics of human radiomics studies*

The use of ICC in the included human radiomics studies was categorized by the ICC purpose. The characteristics of the studies were summarized in the following tables. If a study involved more than one ICC purpose, each purpose was separately listed in the corresponding table. SFR was directly extracted or synthesized from the original data from each study as much as possible. However, in many cases, SFR was not available or could not be clearly and reliably extracted, which was thus labeled NA in the tables. One common reason was that ICC was reported in other forms like mean ± SD. Another reason was that ICC was applied for some but not all features, e.g., for those features after correlation assessment or clustering. Meanwhile, for some other studies, very comprehensive ICC results were reported. The SFRs could not be simply extracted and

 

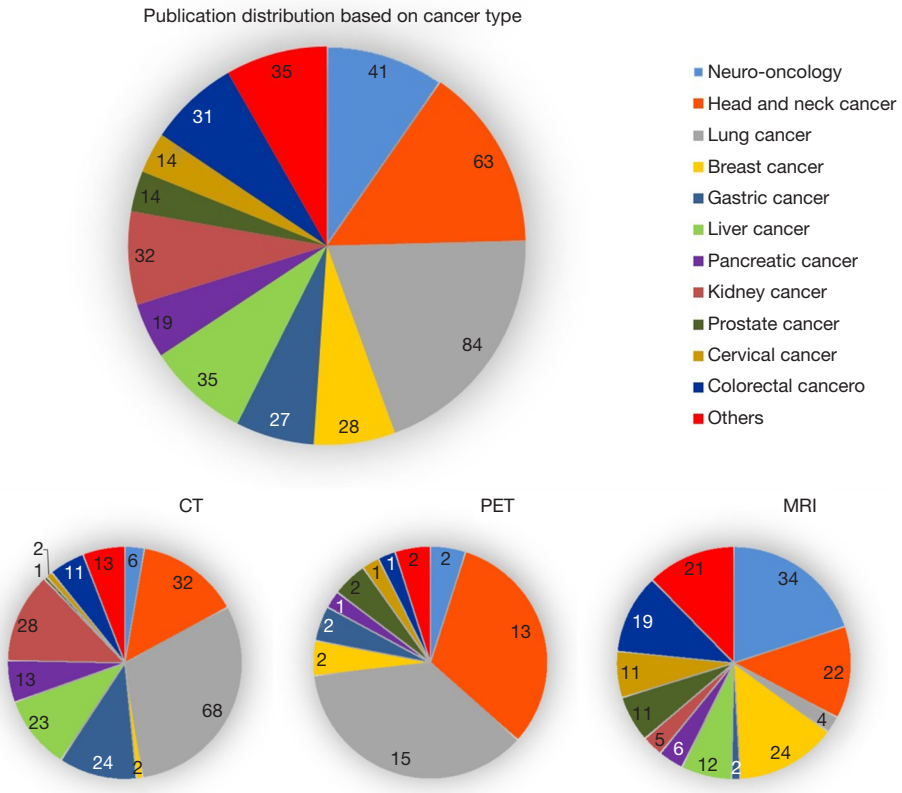


**Figure 4** Distribution of oncological patient studies. The study number is counted repeatedly for each type of cancer if a study investigated more than one cancer type.

listed. They were labeled as NE.

### Radiomics feature reliability due to image acquisition

The use of ICC in the included human radiomics studies to assess the feature reliability due to image acquisition was summarized in *Table 1*.

### Radiomics feature reliability due to image reconstruction

The use of ICC in the included human radiomics studies to assess the feature reliability due to image reconstruction was summarized in *Table 2*.

### Radiomics feature reliability due to image segmentation

There were 416 out of 481 studies reported the ICC results regarding the feature reliability influenced by image segmentation. Only those studies that substantially investigated the radiomics feature reliability to image segmentation as the primary study endpoint were listed in *Table 3*. For other studies, most simply mentioned the ICC use for intra-/inter-observer agreement, and reported a very short ICC (usually optimistic) result, in which SFR was extracted from 206 studies. The SFR distribution was provided by the histogram in Figure S1, showing no apparent differences between imaging modalities (Figure S2).

### Radiomics feature reliability due to image processing

The use of ICC in the included human radiomics studies to assess the feature reliability due to image processing was summarized in *Table 4*.

### Radiomics feature reliability due to feature quantification

The use of ICC in the included human radiomics studies to assess the feature reliability due to feature quantification was summarized in *Table 5*.

### The characteristics of phantom and animal radiomics studies

The study characteristics in the included phantom and

**Table 1** Summary of human radiomics studies using ICC for image acquisition

| First author, year | Disease | Patient number/feature number/ICC sample size | Acquisition factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| CT | | | | | |
| Aerts *et al*., 2016 (17) | Lung cancer | 47/183/31 | Test-retest | NA | NA |
| Hu *et al*., 2016 (18) | Rectal cancer | 40/775/40 | Test-retest | 64.00% | 0.8 |
| Hyunh *et al*., 2016 (19) | Lung cancer | 113/1605/31 | Test-retest | 53.27% | 0.8 |
| Hyunh *et al*., 2017 (20) | Lung cancer | 112/644/31 | Test-retest | 44.41% | 0.8 |
| Hosny *et al*., 2018 (21) | Lung cancer | 1194/NA/32 | Test-retest | NA | 0.8 |
| Soufi *et al*., 2018 (22) | Lung cancer | 162+143/432/31 | Test-retest | NE | 0.75 |
| Dou *et al*., 2018 (23) | Lung cancer | 200/2175/31 | Test-retest | 33.15–51.17% | 0.85 |
| Huang *et al*., 2019 (24) | Lung cancer | 54/203/31 | Test-retest | 65.02% | 0.9 |
| Khorrami *et al*., 2019 (25) | Lung cancer | 125/1542/31 | Test-retest | 76.00% | 0.8 |
| Khorrami *et al*., 2019 (26) | Lung cancer | 90/1542/31 | Test-retest | 67% | 0.85 |
| Osman *et al*., 2019 (27) | Prostate cancer | 342/1618/20 | Test-retest | 32.26% | 0.8 |
| Zwanenburg *et al*., 2019 (28) | Lung cancer, HNC | 31+19/4032/31+19 | Test-retest | 57.3% (lung cancer), 14.4% (HNC) | 0.9 |
| Kadoya *et al*., 2020 (29) | Lung cancer | 287/107/31 | Test-retest | 79.44% | 0.8 |
| Khorrami *et al*., 2020 (30) | Lung cancer | 350/133/31 | Test-retest | NA | 0.75 |
| Ligero *et al*., 2021 (31) | Colorectal/kidney cancer | 43 (97 liver mets)/93/43 (97 liver mets) | Acquisition voltage | 65.59–81.72% | 0.8 |
| Prayer *et al*., 2021 (32) | Lung disease | 60/86/30 | Test-retest, inter-scanner | 75.6% (test-retest), NE (inter-scanner) | 0.9 |
| Vuong *et al*., 2020 (33) | Lung cancer | 124/1404/10+11 | Motion, contrast | 41.30% (motion), 29% (contrast) | 0.9 |
| Yamashita *et al*., 2020 (34) | Pancreatic cancer | 37/266/37 | Inter-scanner, contrast agent | NE | 0.8 |
| MRI | | | | | |
| Fiset *et al*., 2019 (35) | Cervical cancer | 62/1761/8+20 | Test-retest, inter-scanner | 22.6% (test-retest), 6.2% (inter-scanner) | 0.9 |
| Li *et al*., 2019 (36) | Hippocampus | 60/55/60 | Test-retest | NE | 0.4 |
| Zinn *et al*., 2018 (37) | GBM | 79/300/79 | Inter-scanner | NA | NA |
| Bologna *et al*., 2020 (38) | NPC | 142/1072/142 | Imaging parameters, geometrical transformation | 24.72% | 0.75 |
| Jang *et al*., 2020 (39) | Healthy, cardiac referral | 10+51/1023/61 | Test-retest | 2.6-28.8% (healthy), 8.9–34.8% (patient) | 0.8 |
| | | | Test-retest (repositioning) | 0.7–3.1% | |
| Merisaari *et al*., 2020 (40) | Prostate cancer | 112/1694/34 | Test-retest | NA | 0.75 |
| Pandey *et al*., 2021 (41) | Healthy | 87+4+8/93/4+8 | Test-retest | ~71.04% | 0.5 |
| | | | Inter-scanner | 18% (GM), 21.5% (WM) | |

**Table 1** (*continued*)

 

**Table 1** (*continued*)

| First author, year | Disease | Patient number/feature number/ICC sample size | Acquisition factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| Gutmann *et al*., 2020 (42) | Diabetes | 310/684/310 | Test-retest | 26.5% | 0.85 |
| Scalco *et al*., 2020 (43) | Prostate cancer | 14/91/14 | Test-retest repositioning | 7–11% | 0.9 |
| Shiri *et al*., 2020 (44) | GBM | 17/26295192/17 | Test-retest | NE | 0.95 |
| Ta *et al*., 2020 (45) | Healthy | 6/22/6 | Test-retest, inter-scanner | NE | 0.75 |
| Zhang *et al*., 2020 (46) | Brain cancer | 1728/1595/50 | Inter-scanner | 100% | 0.75 |
| Han *et al*., 2021 (47) | GBM | 57/45/57 | Test-retest | NA | 0.75 |
| PET | | | | | |
| Leijenaar *et al*., 2013 (48) | Lung cancer | 34/106/11 | Test-retest | 71% | 0.8 |
| Willaime *et al*., 2013 (49) | Breast cancer | 15/28/9 | Test-retest, tissue | ~46.43 (test-retest), ~64.29% (across tissue) | 0.7 |
| van Velden *et al*., 2014 (50) | Colorectal carcinoma | 29/18 /29 | Test-retest | 83.33% | 0.7 |
| Cheng *et al*., 2016 (51) | Lung cancer | 56/12/56 | Different tracer | NA | 0.95 |
| van Rossum *et al*., 2016 (52) | HNC | 217/80/7 | Test-retest | NE | 0.9 |
| Carvalho *et al*., 2018 (53) | Lung cancer | 215/118/215 | Test-retest | 65.25% | 0.85 |
| Jiang *et al*., 2018 (54) | Gastric cancer | 214/80/30 | Inter-scanner | NA | 0.8 |
| Lin *et al*., 2019 (55) | Prostate cancer | 18/47/18 | Test-retest | 83% | 0.75 |
| Manabe *et al*., 2019 (56) | Cardiac disease | 89/36/33 | Inter-scanner | 47.22% | 0.8 |
| Vuong *et al*., 2019 (57) | Lung cancer | 10+9/1,355/10 | Test-retest (inhale, exhale phase) | 55.6% (shape), 39.3 (wavelet) | 0.9 |
| | | | Inter-scanner (PET/CT *vs*. PET/MRI) | 61% (shape and intensity), 28% (wavelet) | 0.9 |
| PET, CT | | | | | |
| Desseroit *et al*., 2017 (58) | Lung cancer | 73/40/73 | Test-retest | NE | NA |
| PET, MRI | | | | | |
| Jiang *et al*., 2017 (59) | Healthy | 86/NA/66 | Test-retest | NA | 0.75 |

NA, not available or not clear; NE, could not be simply extracted due to comprehensive results, see the reference for detail; HNC, head and neck cancer; liver mets, liver metastases; GBM, glioblastoma multiforme. NPC, nasopharyngeal carcinoma. GM, gray matter. WM, white matter.

**Table 2** Summary of human radiomics studies using ICC for image reconstruction

| First author, year | Disease | Patient number/feature number/ICC sample size | Reconstruction factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| CT | | | | | |
| Ahn *et al*., 2019 (60) | Liver lesion and renal cyst | 1,462/11/1,462 | Inter-method | NE | 0.75 |
| Kolossváry *et al*., 2019 (61) | Vascular disease | 60/171/60 | Inter-method | ~97.04% | 0.9 |
| Koo *et al*., 2017 (62) | Lung cancer | 194/10/194 | Reconstruction intervals | ~70% | 0.8 |
| Lee *et al*., 2019 (63) | Lung nodule | 194 patients (260 scans) /252/114 patients (180 scans) | Voxel size | 9.13% | 0.7 |
| Ligero *et al*., 2021 (31) | Colorectal/kidney disease | 43 (97 liver mets)/93/43 (97 liver mets) | Voxel size | 48.40–78.49% | 0.8 |
| | | | Slice spacing | 43.01–86.02% | 0.8 |
| | | | Slice thickness | 75.25–88.17% | 0.8 |
| | | | Convolution kernel | 55.92–97.85% | 0.8 |
| Prayer *et al*., 2021 (32) | Lung disease | 60/86/60 | Reconstruction kernel and slice thickness, | NE | NE |
| Vuong *et al*., 2020 (33) | Lung cancer | 124/1,404/23 | Convolution kernel | 17.20% | 0.9 |
| Yamashita *et al*., 2020 (34) | Pancreatic cancer | 37/266/37 | Voxel size | NE | 0.8 |
| MRI | | | | | |
| Suter *et al*., 2020 (64) | GBM | 63+76/8,327/19 | K-space subsampling | NE | 0.85 |
| PET | | | | | |
| Altazi *et al*., 2017 (65) | Cervical cancer | 88/79/8 | Inter-method | NE | NA |
| van Velden *et al*., 2016 (66) | Lung cancer | 11/105/11 | Inter-method | 63% (segmentation or reconstruction) | 0.9 |

NA, not available or not clear; NE, could not be simply extracted due to comprehensive results, see the reference for detail; liver mets, liver metastases; GBM, glioblastoma multiforme.

**Table 3** Summary of human radiomics studies using ICC for image segmentation

| First author, year | Disease | Patient number/ feature number/ICC sample size | Segmentation factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| CT | | | | | |
| Parmar *et al*., 2014 (67) | Lung cancer | 20/56/20 | Intra-/inter-observer | 51.8% (manual), | 0.8 |
| | | | Software | 73.2% (software) | |
| Echegaray *et al*., 2015 (68) | Liver cancer | 29/745/29 | Inter-observer | 78.39% | 0.8 |
| Echegaray *et al*., 2016 (69) | Lung cancer | 100/94/100 | Semi-automatic | 89% | 0.7 |
| Qiu *et al*., 2017 (70) | Liver cancer | 15/71/15 | Intra-/inter-observer, semi-automatic | 81% (GrowCut), 77% (GraphCut), 73% (manual) | 0.75 |
| Owens *et al*., 2018 (71) | Lung cancer | 10/83/10 | Intra-/inter-observer, software | NE | 0.75 |

**Table 3** (*continued*)

 

**Table 3** (*continued*)

| First author, year | Disease | Patient number/ feature number/ICC sample size | Segmentation factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| Pavic *et al*., 2018 (72) | Lung cancer, HNC | 11+11+11/1, 404/11+11+11 | Inter-observer | 90% (NSCLC); 59% (HNC); 36% (MPM) | 0.8 |
| Kocak *et al*., 2019 (73) | Kidney cancer | 47/828/25 | Inter-method | 86.2% (contour-focused); 93.2% (margin shrinkage) | 0.9 |
| Kocak *et al*., 2019 (74) | Kidney cancer | 30/744/30 | Intra- observer | 84.4–92.2% (CT); 85.5–93.1% (CECT) | 0.75 |
| | | | Inter-observer | 76.7% (CT); 84.9% (CECT) | |
| Moltz *et al*., 2019 (75) | Liver cancer | 13/110/13 | Inter-observer | 13.64% | 0.9 |
| Mori *et al*., 2019 (76) | Pancreatic cancer | 31/69/31 | Inter-observer | 94.20% | 0.8 |
| Qiu *et al*., 2019 (77) | Liver cancer | 106/71/26 | Semi-automatic | 79% (GrowCut); 73% (GraphCut); 69% (manual) | 0.75 |
| Sung *et al*., 2019 (78) | Liver disease | 58/5/58 | Intra- inter- observer | NE | 0.75 |
| Uthoff *et al*., 2019 (79) | Lung cancer | 71/101/71 | Intra- inter- observer | NA | NA |
| Caballo *et al*., 2020 (80) | Breast cancer | 69/672/35 | Automatic | 90% | 0.75 |
| Haarburger *et al*., 2020 (81) | Lung, kidney, liver tumors | 1,536/92/1,536 | Automatic | 84% (PHiSeg); 88% (expert) | 0.8 |
| Kakino *et al*., 2020 (82) | Lung cancer | 256/93/31 | Inter-observer | 64.52% | 0.5 |
| Kulkarni *et al*., 2021 (83) | Pancreatic cancer | 128/14/128 | Inter-method | NA | 0.9 |
| Liu *et al*., 2020 (84) | HNC | 436/109/436 | Inter-observer, inter-method | NE | 0.9 |
| Nguyen *et al*., 2020 (85) | Kidney cancer | 165/25/165 | Inter-observer | NE | 0.8 |
| Ren *et al*., 2020 (86) | HNC | 47/1032/47 | Intra-observer | 21.6% (2D) 30.4% (3D) | 0.9 |
| | | | Inter-observer | 14.1% (2D) 31.4% (3D) | |
| MRI | | | | | |
| Adduru *et al*., 2017 (87) | Healthy | 38/NA/38 | Software | NE | 0.56 |
| Lee *et al*., 2017 (88) | GBM | 45/180/45 | Inter-observer, software | ~91.1% (tumor Prism 3D), 90% (3D slicer) | 0.8 |
| Bologna *et al*., 2018 (89) | Sarcoma, HNC | 18+18/69/18+18 | Inter-method | ~78.26% (HNC), 85.5% (sarcoma) | 0.78 |
| Saha *et al*., 2018 (90) | Breast cancer | 50/529/50 | Inter-observer | NE | 0.9 |
| Duron *et al*., 2019 (91) | Lachrymal gland tumors, breast cancer | 74+30/69/74+30 | Intra-/inter- observer | NE | 0.8 |
| Fiset *et al*., 2019 (35) | Cervical cancer | 62/1,761/62 | Inter-observer | 95.2% | 0.9 |
| Koçak, 2019 (92) | GBM | 25/1,116/25 | Inter-method | ~25% | 0.9 |
| Lecler *et al*., 2019 (93) | Breast cancer | 37/510/37 | Intra-/inter- observer | 53% | 0.8 |

**Table 3** (*continued*)

**Table 3** (*continued*)

| First author, year | Disease | Patient number/ feature number/ICC sample size | Segmentation factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| Tixier *et al*., 2019 (94) | GBM | 90/108/90 | Semi-automatic | NE | 0.9 |
| Traverso *et al*., 2019 (95) | Cervical cancer | 81/552/81 | Inter-observer | NE | 0.9 |
| Alis *et al*., 2021 (96) | Healthy cardiac | 59/352/59 | Inter-observer | 92% | 0.8 |
| Chen *et al*., 2021 (97) | Cervical cancer | 20/105/20 | Intra- inter- observer | NE | 0.8 |
| Granzier *et al*., 2020 (98) | Breast cancer | 129 (tumors)/1,328 (Radiomix), 833 (pyradiomics)/129 | Inter-observer, software | 41.6% (RadiomiX); 32.8% (Pyradiomics); 41.1% (unfiltered RadiomiX); 16.2% (unfiltered Pyradiomics) | 0.9 |
| Jang *et al*., 2020 (39) | Healthy, cardiac referral | 10+51/1,023/15 | Intra-observer | 61–73% | 0.8 |
| | | | Inter-observer | 32–47% | |
| Lin *et al*., 2020 (99) | Cervical cancer | 169/51/169 | Automatic | NA | NA |
| Gutmann *et al*., 2020 (42) | Diabetes | 310/684/310 | Inter-observer | 82.9% | 0.85 |
| Pati *et al*., 2020 (100) | GBM | 31/11,700/31 | Inter-observer | NE | 0.8 |
| Scalco *et al*., 2020 (43) | Prostate cancer | 14/91/14 | Inter-observer | NE | 0.9 |
| Suter *et al*., 2020 (64) | GBM | 63+76/8,327/19 | Inter-observer | NE | 0.85 |
| PET | | | | | |
| Leijenaar *et al*., 2013 (48) | Lung cancer | 34/106/23 | Inter-observer | 91% | 0.8 |
| Lu *et al*., 2016 (101) | NPC | 40/88/40 | Inter-observer, inter-method | 50% ($^{18}$F-FDG), 62% (11C-choline) | 0.8 |
| van Velden *et al*., 2016 (66) | Lung cancer | 11/105/11 | Inter-method | 63% (segmentation or reconstruction) | 0.9 |
| Bashir *et al*., 2017 (102) | Lung cancer | 53/83/53 | Inter-observer, inter-method | NE | 0.85 |
| Belli *et al*., 2018 (103) | HNC, pancreatic cancer | 25+25/72/25+25 | (Semi)-automatic | 19% (HN-N); 19% (HN-T); 47% (pancreas) | 0.8 |
| Manabe *et al*., 2019 (56) | Cardiac disease | 89/36/33 | Inter-observer | 77.78% | 0.8 |
| PET, MRI | | | | | |
| Yang *et al*., 2020 (104) | NPC | 21/540/21 | Inter-method | 85.74% (PET), 84.81% (T2), 89.07% (DWI) | 0.95 |

NA, not available or not clear; NE, could not be simply extracted due to comprehensive results, see the reference for detail; HNC, head and neck cancer; NSCLC, non-small-cell lung carcinoma; MPM, malignant pleural mesothelioma; CECT, contrast-enhanced CT; GBM, glioblastoma multiforme; HN-N, head and neck cancer-positive lymph node; HN-T, head and neck cancer-tumor; NPC, nasopharyngeal carcinoma.

 

**Table 4** Summary of human radiomics studies using ICC for image processing

| First author, year | Disease | Patient number/feature number/ICC sample size | Image processing factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| CT | | | | | |
| Bogowicz *et al.*, 2016 (105) | Lung cancer, HNC | 11+11/315/11+11 | HU threshold | ~35% | 0.9 |
| | | | Voxel size resampling | ~5% | |
| | | | Temporal resolution | ~30% | |
| | | | Artery contouring | ~45% | |
| | | | Noise threshold | ~52.5% | |
| Ger *et al.*, 2018 (106) | Lung cancer HNC | 20+30/49/20+30 | Voxel size resampling | 71–79.6% | 0.9 |
| Shafiq-Ul-Hassan *et al.*, 2018 (107) | Lung cancer | 18/24/18 | Gray level normalization | 100.00% | 0.8 |
| Lee *et al.*, 2019 (63) | Lung nodule | 194 patients (260 scans)/252/114 patients (180 scans) | Voxel size resampling | 15% | 0.7 |
| Zwanenburg *et al.*, 2019 (28) | Lung cancer HNC | 31+19/4,032/31+19 | Noise | 95.0–97.4% | 0.9 |
| | | | Rotation | ~75–80% | |
| | | | Translation, volume adaptation, contrast | 16.6–32.9% | |
| | | | Rotation, volume adaptation, contrast | 16.8–33.3% | |
| | | | Noise, translation, volume adaptation, contrast | 16.7–33.7% | |
| | | | Rotation, noise, volume adaptation, contrast | 16.7–34.2% | |
| | | | Other | NE | |
| Defeudis *et al.*, 2020 (108) | Colorectal cancer | 14/35/14 | Standardization | ~36–60% | 0.9 |
| Park *et al.*, 2020 (109) | Bladder cancer | 83/55/83 | SNR, and outlier inclusion | NE | 0.75 |
| MRI | | | | | |
| Kim *et al.*, 2019 (110) | GBM | 167/356/167 | Pertubation | 77.20% | 0.75 |
| Li *et al.*, 2019 (36) | Hippocampus | 60/55/60 | Normalization | ~36.36–56.36% | 0.4 |
| Schwier *et al.*, 2019 (111) | Prostate cancer | 15/1,120/15 | Normalization | NE | 0.85 |
| Traverso *et al.*, 2020 (95) | Cervical cancer | 81/552/81 | Normalization | NE | 0.9 |
| Fan *et al.*, 2020 (112) | Breast cancer | 322/107/322 | Voxel size resampling | NE | 0.7 |
| Moradmand *et al.*, 2020 (113) | GBM | 65/107/65 | Noise | 21.4% | 0.9 |
| | | | Noise + bias field | 20.4% | |
| | | | Bias field | 23.2% | |
| | | | Bias field + noise | 22.5% | |
| Scalco *et al.*, 2020 (43) | Prostate cancer | 14/91/14 | Normalization | ~12–14% | 0.9 |
| Shiri *et al.*, 2020 (44) | GBM | 17/26,295,192/17 | Transformation, bias field removal | NE | 0.95 |

**Table 4** (*continued*)

**Table 4** (*continued*)

| First author, year | Disease | Patient number/feature number/ICC sample size | Image processing factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| Suter *et al.*, 2020 (64) | GBM | 63+76/8,327/19 | Perturbation | 42.5% | 0.85 |
| PET | | | | | |
| Branchini *et al.*, 2019 (114) | Pediatric | 21/106/21 | Activity reduction simulation | NE | 0.9 |
| Whybra *et al.*, 2019 (115) | HNC | 441/141/441 | Voxel size resampling | 66% | 0.9 |
| PET, MRI | | | | | |
| Yang *et al.*, 2020 (104) | NPC | 21/540/21 | Pixel size resampling | 55.74% (T2), 60.37% (DWI), 58.33% (PET) | 0.95 |
| | | | Slice thickness | 24.07% (T2 and DWI), 23.89% (PET) | |

NA, not available or not clear; NE, could not be simply extracted due to comprehensive results, see the reference for detail; HNC, head and neck cancer; HU, Hounsfield unit; SNR, signal-to-noise ratio; GBM, glioblastoma multiforme; NPC, nasopharyngeal carcinoma; DWI, diffusion-weighted imaging.

animal radiomics studies were summarized in *Table 6*.

### *Quality of ICC use and reporting*

Generally speaking, the quality of ICC use and reporting was found unsatisfactory in many publications, associating with various flaws and pitfalls. Only 63 studies (13.10%) explicitly and precisely reported the selected ICC form, in which the ICC definition of absolute agreement rather than consistency on feature values predominated. The rationale of ICC form selection was seldom explained. In the remaining 418 articles, the ICC form was either unavailable, implicit (e.g., giving the general ICC formula not specific to a certain ICC form), or incomplete. The available ICC forms described in the studies, either completely or not, were summarized in the Table S1. Very few studies tested the normal distribution of data prior to ICC use (as ICC was based on ANOVA). The adopted reliability criteria/level and the corresponding threshold ICC values could be found in most studies but were heterogeneous. The reliability levels could be binary (low/high, acceptable/unacceptable, stable/unstable, repeatable/unrepeatable, etc.), three (e.g., poor/moderate/good), four (e.g., poor/moderate/good/excellent), five (e.g., poor/fair/moderate/good/excellent) and even six. The thresholds of >0.7, >0.75, >0.8, and >0.9 were frequently used to determine the highest reliability level. The thresholds of <0.2, <0.4, and <0.5 were frequently adopted to determine the lowest reliability level. The reported ICC values were normally presented in the mean (± SD), median, range, or interquartile range (IQR). Confidence interval (usually 95% CI) was reported along with ICC only in 64 studies. Many studies seemed to interpret reliability levels based on the estimated ICC values without giving or referring to the reported ICC confidence interval.

### *Notable findings of radiomics feature reliability as revealed by ICC*

The reported ICC results were highly heterogeneous, varying by imaging modality, ICC purpose, disease, lesion type, sample size, as well as feature types. Meanwhile, they were also frequently reported with pitfalls. Therefore, it was impractical to conduct quantitative data synthesis and meta-analysis based on the reported ICCs to reliably estimate the achievable absolute radiomics reliability levels for different modalities, purposes, or diseases. But there were still a few notable consistent findings observed on the report ICCs even in the presence of high study heterogeneities.

#### High satisfactory feature rates were reported for most intra/inter-observer segmentation studies, indicating the high robustness of many radiomics features to intra/inter-observer segmentation variability

In the hundreds of articles using ICC for assessing intra/inter-observer segmentation reliability of radiomics features, only a small number reported relatively negative

 

**Table 5** Summary of human radiomics studies using ICC for feature quantification

| First author, year | Disease | Patient number/feature number/ICC sample size | Quantification factor | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| CT | | | | | |
| Bogowicz *et al*., 2016 (105) | Lung cancer, HNC | 11+11/315/11+11 | Discretization | ~40% | 0.9 |
| Foy *et al*., 2018 (116) | HNC | 39/12/39 | Radiomics calculation | 33% | 0.9 |
| Soufi *et al*., 2018 (22) | Lung cancer | 162+143/432/305 | Discretization | NE | 0.75 |
| Lee *et al*., 2019 (63) | Lung nodule | 194 patients (260 scans) /252/114 patients (180 scans) | Discretization | ~9.5–11.5% | 0.7 |
| Zwanenburg *et al*., 2019 (28) | Lung cancer, HNC | 31+19/4,032/31+19 | Discretization | ~10–65% (FBW); 10–40%(FBN) | 0.9 |
| Park *et al*., 2020 (109) | Bladder cancer | 83/55/83 | Discretization | NE | 0.75 |
| MRI | | | | | |
| Duron *et al*., 2019 (91) | Lachrymal gland tumors, breast cancer | 74+30/69/74+30 | Discretization | NE | 0.8 |
| Schwier *et al*., 2019 (111) | Prostate cancer | 15/1,120/15 | Discretization | NE | 0.85 |
| Traverso *et al*., 2020 (95) | Cervical cancer | 81/552/81 | Discretization | NE | 0.9 |
| Pandey *et al*., 2021 (41) | Healthy | 87+4+8/93/4+8 | Harmonization | 60.33% (GM); 62% (WM) | 0.5 |
| Shiri *et al*., 2020 (44) | GBM | 17/26,295,192/17 | Discretization | NE | 0.95 |
| Suter *et al*., 2020 (64) | GBM | 63+76/8,327/19 | Discretization | NE | 0.85 |
| PET | | | | | |
| Tixier *et al*., 2012 (117) | HNC | 16/25/16 | Discretization | NE | NA |
| Leijenaar *et al.*, 2015 (118) | Lung cancer | 35/44/35 | Discretization | NE | NA |
| Lu *et al*., 2016 (101) | NPC | 40/88/40 | Discretization | 23% (18F-FDG), 21% (11C-choline) | 0.8 |
| Altazi *et al*., 2017 (65) | Cervical cancer | 88/79/80 | Discretization | 18% (GLCM and GLRLM) | 0.9 |
| Bogowicz *et al*., 2017 (119) | HNC | 128+50/649/178 | Feature implementation | 12% | 0.8 |
| Lv *et al*., 2018 (120) | NPC | 106/57/106 | Averaging, symmetry, distance | NE | 0.8 |
| Branchini *et al*., 2019 (114) | Pediatric | 21/106/21 | Discretization | NE | 0.9 |
| PET, MRI | | | | | |
| Yang *et al*., 2020 (104) | NPC | 21/540/21 | Discretization | 12.96% (PET and T2), 11.30% (DWI) | 0.95 |

NA, not available or not clear; NE, could not be simply extracted due to comprehensive results, see the reference for detail; HNC, head and neck cancer; FBW, fixed bin width; FBN, fixed bin number; GM, gray matter; WM, white matter; NPC, nasopharyngeal carcinoma; GBM, glioblastoma multiforme; GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; DWI, diffusion-weighted imaging.

 

**Table 6** Summary of phantom and animal studies

| First author, year | Phantom or animal | ICC purpose | Factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| CT | | | | | |
| Panth *et al*., 2015 (121) | Animal (mice) | Acquisition, segmentation | Test-retest, inter-observer | NE | NA |
| Berenguer *et al*., 2018 (122) | Anthropomorphic pelvic phantom, multi-material phantom | Acquisition | Test-retest | 93.80% | 0.9 |
| | | | Imaging parameters | 69.89% | 0.9 |
| Ger *et al*., 2018 (106) | Credence Cartridge Radiomics phantom | Processing | Resampling | 71.43% | 0.9 |
| Mannil *et al*., 2018 (123) | Human urinary stones | Segmentation | Intra-/inter-observer | 29.87% | 0.8 |
| Li *et al*., 2020 (124) | Anthropomorphic thoracic phantom | Acquisition, reconstruction, segmentation, quantification | Effective dose and pitch | 86.90% | 0.8 |
| | | | Slice thickness and filter | 89.75% | |
| | | | Intra-observer | 89.80% | |
| | | | Inter-observer | 92.00% | |
| | | | Gray-level range (HU) | 86.40% | |
| | | | Bin size (HU) | 58.00% | |
| Nardone *et al*., 2020 (125) | Commercial phantom (CIRS model 467) | Acquisition | Inter-scanner | 0% | 0.75 |
| | | | Inter-scanner (delta) | 45.90% | |
| MRI | | | | | |
| Song *et al*., 2014 (126) | Animal (rats) | Acquisition, segmentation | Test-retest, Intra-observer | NE | 0.75 |
| Baeßler *et al*., 2019 (127) | Fruits | Segmentation | Intra-observer | 100.00% | 0.75 |
| | | | Inter-observer | 62.22% | |
| Bologna *et al*., 2019 (128) | Phantom (virtual) (brain-web) | Acquisition, processing | Varying TE and TR | 76.00% | 0.75 |
| | | | Normalization | 81.00% | 0.75 |
| Bologna *et al*., 2019 (129) | Virtual phantom (brainweb) | Acquisition, processing | Varying TE and TR, denoising, voxel size resampling, bias field correction (sequence dependent T1 & T2) | 59.80% | 0.75 |
| Cattell *et al*., 2019 (130) | Fruits | Segmentation, processing | Erosion and dilation (normalized mean ±3SD) | 75.89% | 0.9 |
| | | | Erosion and dilation (normalized 0-max) | 90.18% | |
| | | | Denoising (normalized mean ± 3SD) | 34.78% | |
| | | | Denoising (normalized 0-max) | 47.82% | |
| | | | Voxel size resampling (normalized mean ± 3SD) | 64.28% | |
| | | | Voxel size resampling (normalized 0-max) | 83.92% | |
| Bianchini *et al*., 2020 (131) | Customised female pelvis phantom | Acquisition | Test-retest (T1) | 74.76% | 0.9 |
| | | | Test-retest (T2) | 95.26% | |

**Table 6** (*continued*)

**Table 6** (*continued*)

| First author, year | Phantom or animal | ICC purpose | Factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| Dreher C *et al*., 2020 (132) | DWI phantom and fruits | Acquisition, segmentation | Test-retest | 84.78% | 0.9 |
| | | | Intra-observer | 97.80% | |
| | | | Inter-observer | 95.65% | |
| Eresen *et al*., 2020 (133) | Animal (mice) | Segmentation | Intra-observer | ~58.33% (T2), ~61.74% (T1) | 0.75 |
| Jang *et al*., 2020 (39) | Fruits and vegetables | Acquisition | Test-retest | ~43.44% | 0.8 |
| | | | Test-retest re-positioning | ~9.82% | |
| Rai *et al*., 2020 (134) | Customized phantom printed by a Connex 260 polyjet printer | Acquisition | Inter-scanner | 53.62% | 0.8 |
| | | | Phantom dependence | 55.07% | |
| Bianchini *et al*., 2021 (135) | Pelvic-shaped container (NEMA IEC Body Phantom Set: Spectrum Corporation) | Acquisition | Test-retest | 82.90% | 0.9 |
| | | | Test-retest re-positioning | 58.33% | |
| | | | Inter-scanner | 3.30% | |
| PET | | | | | |
| Gallivanone *et al*., 2018 (136) | Anthropomorphic Alderson Thorax phantom | Acquisition | Test-retest | 53.45% | 0.6 |
| Ger *et al*., 2019 (137) | 3-dimensional Hoffman brain phantom | Processing, reconstruction | Pixel size resampling | ~97.00% | 0.9 |
| | | | Filter cutoff | ~82.30% | |
| | | | Effective iteration | ~86.16% | |
| | | | Time per bed position alteration | ~92.00% | |
| | | | Q.clear | 92.00% | |
| Pfaehler *et al*., 2019 (138) | NEMA NU 2-2012 IQ phantom | Acquisition, reconstruction, quantification | Scan duration (noise) | NE | 0.8 |
| | | | High uptake | 32.00% | |
| | | | Low uptake | 21.00% | |
| | | | Point-spread-function (PSF) | 53.00% | |
| | | | Osem | 30.00% | |
| | | | Time-of-flight | 32.00% | |
| | | | Discretization (FBW) | 89.00% | |
| | | | Discretization (FBN) | 35.00% | |
| Pfaehler *et al*., 2020 (139) | 3D printed phantom inserts in NEMA NU 2-2012 IQ phantom | Acquisition, quantification | Test-retest (statistically equal replicates) (FBN) | 90.70% | 0.9 |
| | | | Test-retest (statistically equal replicates) (FBW) | 97.30% | |
| | | | Test-retest (FBN) | 79.13% | 0.6 |
| | | | Test-retest (FBW) | 89.96% | |
| | | | Multicenter (FBN) | 45.60% | |
| | | | Multicenter (FBW) | 62.97% | |

**Table 6** (*continued*)

**Table 6** (*continued*)

| First author, year | Phantom or animal | ICC purpose | Factors | Satisfactory feature rate | ICC threshold |
|---|---|---|---|---|---|
| Yang *et al*., 2020 (140) | Digital phantom (Zubal anthropomorphic phantom) | Segmentation, quantification | Inter-observer (bin number 32 or 64) | 84.00% | 0.9 |
| | | | Inter-observer (bin number 128 or 256) | 88.00% | |

NA, not available or not clear; NE, could not be simply extracted due to comprehensive results, see the reference for detail; HU, Hounsfield unit; TE, echo time; TR, repetition time; DWI, diffusion-weighted imaging; OSEM, ordered subset expectation maximization; FBW, fixed bin width; FBN, fixed bin number.

reliability results. For instance, Jang *et al.* (39) showed that in the inter-observer segmentation reproducibility study in cardiac patients, only 32.1%, 46.7%, and 35.5% of MRI radiomics features were reproducible with the cine bSSFP, T1 mapping, and T2 mapping, worse than the corresponding 73.1%, 66.8%, and 61.1% reproducible features in the intra-observer segmentation. Liu *et al.* (84) investigated 109 radiomics features on 436 contrast-enhanced CT images of oropharyngeal cancer patients and found that "most radiomic features in this study varied a lot when the ROIs were not well segmented. For both the representation agreement and predictive agreement, the ICC and CCC were below 0.5 for all the features." Uthoff *et al.* (79) reported that "observers had perfect intra-repeatability (ICC =1.0)" but "demonstrated fair inter-reader variability (ICC =0.52)" for 4 observers (2 radiologists, 2 pulmonologists) in 100 cases of non-small cell lung cancer CT scans. Many other studies generally reported high SFRs, implying excellent robustness to intra/inter-observer segmentation disagreement, independent of modalities and diseases, although different ICC thresholds were applied. Meanwhile, radiomics feature robustness to inter-observer segmentation seemed not notably inferior to intra-observer segmentation.

**Comparable or better radiomics feature reliability was reported for (semi-)automated segmentation than manual segmentation with much shorter segmentation time**

Manual image segmentation in radiomics analysis involved intensive labor work of clinicians and was time-consuming, and also suffered from intra-/inter-observer segmentation disagreement, leading to the low cost-effectiveness/efficiency of radiomics so greatly hampering its wide application in clinical practice. Thus, lots of efforts were taken to develop (semi-)automated segmentation as a potential alternative in radiomics research. Moreover, (semi-)automatic segmentation was frequently reported useful to further reduce the intra-/inter-observer radiomics feature variability induced by manual segmentation in the included studies in addition to its advantage in segmentation time, suggesting the future role of (semi-)automated segmentation in more reliable and cost-effective/efficient radiomics analysis (67,69,70,77,80,99,103,141,142).

**Acquisition had substantial impacts on radiomics feature values, and their impact on feature reliability was larger than the impact by intra-/inter-observer segmentation**

Based on the reported ICC results, image acquisition had substantial impacts on the radiomics feature values for all imaging modalities and acquisition protocols. Regarding modality dependence, the reported SFRs seemed to be highest in PET and lowest in MRI [excluding the outlier of 100% inter-scanner SFR reported by Zhang J *et al.* (46)], which might be partially explained by the relatively smooth low-resolution PET image and the multi-contrast high-resolution MRI images with considerable anatomical details acquired by different sequences. The simple intra-scanner test-retest could introduce considerable feature value variations(17,18,24,28,32,36,40,44,139). The inter-scanner (or inter-center) acquisition (with similar imaging protocols or imaging parameter changes) induced even more radiomics feature variability than the intra-scanner test-retest (32,35,45,135,137,139). In the studies investigating both acquisition and segmentation, acquisition was consistently reported to have much larger impacts on feature variability (always smaller ICCs) than segmentation (35,39,42,44,48,56,126,132) for all modalities.

**Feature reliability and ICCs were heterogeneous for post-processing and feature quantification; optimized post-processing and feature quantification could be used to mitigate acquisition-induced radiomics variability**

Image post-processing and feature quantification were

 

usually used to explore the robustness of radiomics features and improve feature reliability by optimized or standardized approaches. Among these approaches, image intensity discretization and normalization were most frequently investigated. Actually, there could be tremendous types of image post-processing and feature quantification methods, algorithms, and tools that were applicable to the acquired original images and thus had remarkable influences on feature values. In many studies, various post-processing and feature quantification approaches were conducted and optimized to mitigate the possible radiomics feature variability introduced in the acquisition procedure (28,31,36,41,44,111,114,129). The results suggested that comprehensive image perturbation and quantification might be helpful to improve radiomics reliability, in particular for those retrospective radiomics studies in which existing imaging data were used without control on imaging acquisition protocol. For example, in a study by Zwanenburg *et al.* (28), image perturbation chains were proposed to be used as an alternative to test-retest imaging to assess feature robustness. Most robust features in acquisition test-retest were successfully identified by comprehensive image perturbations. In another study by Suter *et al.* (64), single-center MRI data was perturbed to simulate unseen multi-center MRI data with greater variabilities, which generated and conducted over 16 million tests of typical perturbations and to identify robust radiomics features for multi-center radiomics study. In contrast, post-processing and feature quantification were seldom proposed for mitigation or compensation for the radiomics feature variability affected by segmentation.

**Shape and first-order (FO) radiomics features were frequently reported to be more robust to various variability factors than texture features in the original image domain**

Different types of radiomics features could subject to different levels of variability influenced by different factors. Among the heterogeneous reported results, it was noticed that shape or first-order (FO, or named histogram) features in the original image domain were often reported to be more robust than texture (also named second-order or higher-order) features to different variability factors of acquisition (18,36,44,57,134), post-processing and quantification (36,115,116,128,130,132), and segmentation (94,99,104,142-145). In different types of texture features, GLCM (gray-level co-occurrence matrix) features were observed to be more robust than other texture features in a few studies (18,44,94,116,128,143,145). On the other hand, opposite or deviant results on the low reliability of shape features were also occasionally reported. For instance, Rai *et al.* reported that none of the shape features exhibited high inter-(MRI) scanner stability (ICC >0.8), the lowest among all feature types (134). Tixier *et al.* showed that shape features in MRI (ICC =0.74) were among the most impacted feature types by the choice of segmentation method, with poorer reliability than first-order and GLCM features (ICC >0.96) (94). Beyond the radiomics features in the original image domain, radiomics features in the transformed domains, most frequently in the Laplacian of Gaussian (LoG) filtered domain and wavelet domain, were also investigated in many studies. No uniform robustness of these transformed features compared to those original features could be derived from the included studies.

### *Other statistical metrics in conjunction with ICC*

A variety of statistical metrics were used in conjunction with ICC for different purposes. For segmentation purposes, dice similarity coefficient (DSC) was often reported. Bland-Altman analysis was conducted in some studies involving paired comparison of two observers/acquisitions/measurements. Other types of statistical metrics such as concordance correlation coefficient (CCC), coefficient of variation (CV), Pearson/Spearman correlation coefficient, false discovery rate (FDR), (normalized) dynamic range, Krippendorff's alpha, percentage difference, and between-class distance (BD) were also used in combination with ICC.

### *Risk of bias*

There were different levels and aspects of potential risks of bias in the included studies for many reasons. Many clinical studies were limited in their retrospective study nature and were usually conducted without phantom validation and control on acquisition protocol. Many technical studies utilized public imaging data, and the heterogeneity in these data might not be well understood or compensated. Very few studies described the imaging protocol sufficiently to the desired level of detail, as suggested in (4). Similarly, details in the intra-/inter-observer segmentation process were normally insufficiently described. For many clinical studies aiming for radiomics diagnosis or prognosis performance, the possible publication bias on the very high feature reliability to intra/inter-observer segmentation might not be neglected in that much lower ICC and SFRs

 

were reported in the studies that substantially assessed feature robustness to segmentation as the primary study endpoint. The statistical power of the calculated ICC might not be strong enough due to the limited sample size and observer/acquisition/measurement number. The investigated radiomics features in many studies might have different definitions even with the same or similar name. They might not have been well standardized due to the different implementations in a variety of software and in-house built programs, in particular in the studies before the proposal of feature standardization by the image biomarker standardization initiative (IBSI) (146). Besides, risks of bias could also be induced by the flaws and pitfalls of using and reporting ICC as identified in the articles.

## Discussion

Radiomics research is experiencing an increasing explosive rate both in publication volume and diversity in recent years (5). However, along with the soaring publication numbers of radiomics, more concerns, questions, and/or criticisms on radiomics reliability are also increasingly raised in the last few years (11,147,148). Some recent systematic review papers (149-153) also showed that many radiomics publications had suboptimal or poor study quality, as revealed by the low RQS.

Reliability is highly correlated to RQS criteria. For instance, image protocol quality (+1 point), phantom study on all scanners (+1 point) and imaging at multiple time points (+1 point) are all related to acquisition reproducibility/replicability/reliability; Multiple segmentations (+1 point) is related to intra-/inter-observer agreement analysis; Feature reduction or adjustment for multiple testing (+3 point if implemented or −3 if not implemented) is related to feature correlation and redundancy analysis. ICC could be used to fulfill these criteria on RQS to improve radiomics study quality.

The increasing publication number with time reflects the fact that much more efforts have been taken to investigate radiomics reliability in recent years, in particular for MRI. Clinical oncology is still the major arena of radiomics research as revealed in this review, consistent with a recent bibliometric review (5), while different imaging modalities have substantially different roles in different cancers, as reflected in the proportions of the publications.

Regarding the ICC purpose, it is within the expectation that ICC was most frequently used for segmentation reliability assessment, particularly for intra-/inter-observer agreement. This could be mainly explained by the large fractions of retrospective studies. It also reflects the common interests and concerns on radiomics reliability from clinicians. It is interesting to notice that many intra-/inter-observer agreement studies reported high ICC values or SFRs, which might suggest that many radiomics features are quite robust to (manual) lesion segmentation. In other words, radiomics reliability to lesion segmentation might not be much concerned.

Although segmentation dominated the ICC use, the importance of radiomics reliability in other aspects could not be overlooked or underestimated. Image acquisition and reconstruction are at the very front-end of the complex radiomics workflow, and greatly impact the quality of the original imaging data for radiomics reliability assessment. Indeed, image acquisition and reconstruction were strongly suggested to impact much more on radiomics reliability than segmentation. However, the influence of image acquisition and reconstruction on radiomics feature reliability was still much underexplored relative to segmentation studies. There are still many unknowns about how image acquisition and reconstruction affect radiomics reliability. Much research work is warranted in the future, in particular for MRI, due to its semi-quantitative image intensity nature, various image contrasts, and much greater variability in image acquisition and reconstruction compared to CT and PET.

ICC was also frequently used for reliability assessment attributed to post-acquisition image processing and feature quantification. The reported ICCs were heterogeneous in these studies, much dependent on the various processing types and different implementations. In theory, there could be infinite types of post-processing methods applicable to original images, so potentially lead to even bigger radiomics feature variability than acquisition and reconstruction. But, in practice, post-processing and feature quantification are investigated and utilized to mitigate acquisition and reconstruction-related feature variability by taking advantage of comprehensive and powerful computation capability for robust feature selection, without the need for prior knowledge on image acquisition variability. However, the evidence is so far not strong enough. More rigorous validation and evaluation are definitely warranted.

Some common pitfalls of using and reporting ICC were frequently identified in the articles. First, the information on ICC form and its selection was missing, ambiguous or incomplete in a large number of articles. Meanwhile, relevant information like scanner/observer/measurement

 

numbers sometimes was not clarified to facilitate ICC form selection. When ICC is used as a reliability metric, it is important for researchers to carefully select the most appropriate ICC form. The inappropriate selection of ICC form might mathematically yield similar ICC values but could lead to substantially different and even misleading interpretations. No article conducted sample size estimation for ICC calculation, which could be helpful, although might not be necessary, for ICC precision estimation. In terms of ICC reporting, ICC values were often reported without the confidence interval. Without reporting the confidence interval, the precision of ICC could not be known. For instance, a very high ICC value but associating with a very wide width of confidence interval (large uncertainty and low precision) could not guarantee the high reliability. Heterogeneous ICC thresholds were used for reliability assessment, also hampering rigorous data synthesis for cross-study comparison. Occasionally, ICC threshold values were only implicitly indicated or unavailable. The reliability levels of poor (ICC <0.5), moderate (ICC: 0.50–0.75), good (ICC: 0.75–0.90), and excellent (ICC >0.90) as suggested in (13) were frequently adopted. But, on the other hand, the conditions under which the criteria were suggested were usually neglected, i.e., "*As a rule of thumb, researchers should try to obtain at least 30 heterogeneous samples and involve at least 3 raters whenever possible when conducting a reliability study*." (13). The reliability levels were also seemed to be inappropriately determined on the basis of the ICC value itself rather than its confidence interval. ICC was normally interpreted without further quantifying the underlying true variance ($\sigma_T^2$) and error variance ($\sigma_E^2$). Actually, a high ICC value might mainly reflect the high between-subject heterogeneity (such as malignant tumors) in the sampled population but does not guarantee the accuracy or precision of radiomics feature quantification. Vice versa, a low ICC might probably be resulted from the high homogeneity (such as normal tissues) in the subjects, even with high measurement accuracy and precision.

Some suggestions could be given to mitigate the identified pitfalls for future radiomics studies. Overall, if radiomics reliability itself is the major purpose of a radiomics study, guidelines for reporting reliability and agreement studies (GRRAS) should be helpful in the study planning (154). In order to facilitate ICC form selection, the model, type, and definition of the ICC form should be justified or explained. Relevant information like scanner/observer/measurement numbers need to be sufficiently disclosed. The guideline proposed by Koo *et al.* is an excellent reference and is easy to follow (13). It would be very helpful to conduct sample size estimation for ICC calculation in order to assure that the study could have an adequate chance of achieving the desired ICC precision (155,156). After ICC form selection, the tool used for ICC calculation should be reported with software name, version, and setting. The ICC calculation results should be reported along with the confidence interval. Meanwhile, the criteria for ICC appraisal should be clearly described. It should also be kept in mind that it is the estimated CI forms the basis to evaluate the reliability level, but not the ICC value itself. Along with ICC, the joint use of other statistical metrics could strengthen the study quality and statistical power. For instance, if paired observers/acquisitions/measurements were involved, Bland-Altman analysis is anticipated and beneficial. For segmentation reliability assessment, dice similarity coefficient (DSC) is desirable. Last but not least, the acceptability of ICC should be determined on the requirements by each specific study and clinical application, rather than simply on the calculated values from the specific sample populations and pre-defined thresholds.

There are some limitations to this study. First, the literature search in a single database was one limitation, although partially compensated by the prior knowledge on additional papers. Meanwhile, even in a single database, there are tremendous numbers of publications relevant to radiomics, but it is not uncommon that a variety of terms are used instead, which makes the precise localization of these publications even more difficult. So, there might still be potentially eligible studies missed for analysis. The ICC use and its result reporting had to be recognized and extracted through full text (and even Supplementary materials) examination rather than title and abstract screening. This procedure involved tremendous work and might slightly affect the inclusion and exclusion of papers. Nonetheless, the large sample size of 481 studies should not considerably weaken or bias the statistics in this review. Second, this review concentrated on a single metric of reliability, i.e., ICC, which tackles only a very narrow topic on general radiomics reliability. ICC is only applicable to continuous variables, so the radiomics reliability revealed by ICC is usually on the level of radiomics feature values. The role of ICC is relatively minor in the reliability aspects of radiomics feature reduction and modeling, as well as model outcome/performance assessment. It is acknowledged that many other statistical metrics could be applicable or more suitable in radiomics reliability assessment in various scenarios, providing complementary or additional information on radiomics

 

reliability beyond ICC. Thus, the current status of radiomics reliability could only be partially reflected in the included papers. This study by no means formed a systematic review and meta-analysis on the diagnostic accuracy of radiomics, so study quality was not individually assessed in each article by following QUADAS-2 (15), TRIPOD (16) or RQS (4), but PRISMA statement was followed (157). Third, there were great difficulties in study quality normalization, data synthesis, and harmonization on the highly heterogeneous study characteristics along with the pitfalls in ICC use and reporting. It was of great difficulty to conduct quantitative analysis on cross-study ICC assessment. The use of SFR slightly mitigated this issue, but SFR itself also had pitfalls such as different ICC thresholds. Therefore, the consensus on the degree of radiomics reliability that has been achieved, or could be achievable in radiomics research could not be safely derived. Fourth, radiomics feature reliability has been suggested to be dependent on imaging modality, organ, disease, and other factors, which was also noticed in some included individual studies (72,89,91,103,104,158). But these dependencies could not be further generalized in this review. Our study collected, analyzed and presented data in a modality-neutral and disease-neutral way. Moreover, we also recognized that it was still an extremely difficult task for this dependency investigation in the presence of high heterogeneities of study characteristics even though hundreds of studies had been included. But, on the other hand, it should be cautioned that there might be a potential risk of bias by trying to present modality-neutral or disease-neutral common findings in the study. The validity of these findings might be violated if applied to some fewer common diseases or other modalities. Therefore, future research efforts on disease-specific and modality-specific feature reliability are desirable. Fifth, some flaws and pitfalls in selecting, reporting, and interpreting ICC were identified in many radiomics studies, so some suggestions were given. But we did not intend to specifically propose a standardized form of ICC use for future radiomics studies. The standardization of QIB metrology (159), the IBSI radiomics feature standardization (146,160), the guidelines for reporting reliability and agreement studies (GRRAS) (154), the general guideline of selecting and reporting ICCs (13), and statistical methods for clinical reliability in different aspects (121,161-164), have been well established in the medical literature. They could act as excellent guidelines or references for radiomics study planning. But, consensus toward the standardized radiomics reliability assessment and reporting is yet to be reached by the whole community.

## Conclusions

This study attempted to have an updated overview on the current status of radiomics reliability research from the perspective of using and reporting ICC in the ever-fast-expanding radiomics literature. The 481 eligible CT, PET, and MRI radiomics studies yielded from the literature search partially revealed the fact that much more efforts have been taken to rigorously assess radiomics reliability for clinical use, in particular in the recent two years. ICC was used for assessing different aspects of radiomics feature reliability in these studies, but feature reliability with respect to image segmentation was much more reported than reliability to other factors such as image acquisition, reconstruction, post-processing, and feature quantification. As indicated by the reported satisfactory ICCs in intra/inter-observer segmentation agreement, manual segmentation seems to be the least influential factor on radiomics reliability, but the risk of bias might be cautioned. The (semi-)automated segmentation may further increase segmentation agreement to further increase radiomics feature reliability with better cost-effectiveness/efficiency in the future. Image acquisition could introduce much more feature variability than image segmentation. More research on radiomics reliability with respect to image acquisition and reconstruction is desired. Comprehensive image post-processing and feature quantification techniques could be applied for radiomics analysis and yield different levels of radiomics reliability. Optimized comprehensive image post-processing and feature quantification could be used to mitigate image acquisition-induced variability and thus improve reliability. There were some common flaws and pitfalls in ICC use, as identified in many studies. Thus, some suggestions were given in order to mitigate them and to improve radiomics reliability research quality for future studies. Unfortunately, it was also recognized that the included studies were highly heterogeneous in characteristics and quality, greatly hampering the reliable data synthesis for further meta-analysis. Therefore, no consensus on the degree of radiomics reliability that has been achieved or could be achievable in radiomics research could be safely derived and reached by this review. More research works are warranted in the future.

## Acknowledgments

*Funding:* This study was supported by hospital research project REC-2019-09. The authors have no relevant

 

conflicts of interest to disclose.

## Footnote

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/qims-21-86). The authors have no conflicts of interest to declare. Dr. JY serves as an unpaid Associate Editor of *Quantitative Imaging in Medicine and Surgery*. The authors have no other conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.



## References

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-6.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563-77.
3. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: The promise of radiomics. Phys Med 2017;38:122-39.
4. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749-62.
5. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. Eur J Radiol 2020;127:108991.
6. Reuzé S, Schernberg A, Orlhac F, Sun R, Chargari C, Dercle L, Deutsch E, Buvat I, Robert C. Radiomics in Nuclear Medicine Applied to Radiation Therapy: Methods, Pitfalls, and Challenges. Int J Radiat Oncol Biol Phys 2018;102:1117-42.
7. Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol 2016;61:R150-66.
8. Miles K. Radiomics for personalised medicine: the long road ahead. Br J Cancer 2020;122:929-30.
9. Fornacon-Wood I, Faivre-Finn C, O'Connor JPB, Price GJ. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. Lung Cancer 2020;146:197-208.
10. Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 2021;31:1-4.
11. Hatt M, Le Rest CC, Tixier F, Badic B, Schick U, Visvikis D. Radiomics: Data Are Also Images. J Nucl Med 2019;60:38S-44S.
12. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
13. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155-63.
14. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 2018;102:1143-58.
15. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63.
17. Aerts HJ, Grossmann P, Tan Y, Oxnard GR, Rizvi N, Schwartz LH, Zhao B. Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC. Sci Rep 2016;6:33860.
18. Hu P, Wang J, Zhong H, Zhou Z, Shen L, Hu W, Zhang Z. Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget 2016;7:71440-6.
19. Huynh E, Coroller TP, Narayan V, Agrawal V, Hou Y, Romano J, Franco I, Mak RH, Aerts HJ. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. Radiother Oncol 2016;120:258-66.
20. Huynh E, Coroller TP, Narayan V, Agrawal V, Romano J, Franco I, Parmar C, Hou Y, Mak RH, Aerts HJ. Associations of Radiomic Data Extracted from Static and

 

Respiratory-Gated CT Scans with Disease Recurrence in Lung Cancer Patients Treated with SBRT. PLoS One 2017;12:e0169172.
21. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, Bussink J, Gillies RJ, Mak RH, Aerts HJWL. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 2018;15:e1002711.
22. Soufi M, Arimura H, Nagami N. Identification of optimal mother wavelets in survival prediction of lung cancer patients using wavelet decomposition-based radiomic features. Med Phys 2018;45:5116-28.
23. Dou TH, Coroller TP, van Griethuysen JJM, Mak RH, Aerts H. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. PLoS One 2018;13:e0206108.
24. Huang L, Chen J, Hu W, Xu X, Liu D, Wen J, Lu J, Cao J, Zhang J, Gu Y, Wang J, Fan M. Assessment of a Radiomic Signature Developed in a General NSCLC Cohort for Predicting Overall Survival of ALK-Positive Patients With Different Treatment Types. Clin Lung Cancer 2019;20:e638-51.
25. Khorrami M, Khunger M, Zagouras A, Patil P, Thawani R, Bera K, Rajiah P, Fu P, Velcheti V, Madabhushi A. Combination of Peri- and Intratumoral Radiomic Features on Baseline CT Scans Predicts Response to Chemotherapy in Lung Adenocarcinoma. Radiol Artif Intell 2019;1:e180012.
26. Khorrami M, Jain P, Bera K, Alilou M, Thawani R, Patil P, Ahmad U, Murthy S, Stephans K, Fu P, Velcheti V, Madabhushi A. Predicting pathologic response to neoadjuvant chemoradiation in resectable stage III non-small cell lung cancer patients using computed tomography radiomic features. Lung Cancer 2019;135:1-9.
27. Osman SOS, Leijenaar RTH, Cole AJ, Lyons CA, Hounsell AR, Prise KM, O'Sullivan JM, Lambin P, McGarry CK, Jain S. Computed Tomography-based Radiomics for Risk Stratification in Prostate Cancer. Int J Radiat Oncol Biol Phys 2019;105:448-56.
28. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, Löck S. Assessing robustness of radiomic features by image perturbation. Sci Rep 2019;9:614.
29. Kadoya N, Tanaka S, Kajikawa T, Tanabe S, Abe K, Nakajima Y, Yamamoto T, Takahashi N, Takeda K, Dobashi S, Takeda K, Nakane K, Jingu K. Homology-based radiomic features for prediction of the prognosis of lung cancer based on CT-based radiomics. Med Phys 2020;47:2197-205.
30. Khorrami M, Bera K, Leo P, Vaidya P, Patil P, Thawani R, Velu P, Rajiah P, Alilou M, Choi H, Feldman MD, Gilkeson RC, Linden P, Fu P, Pass H, Velcheti V, Madabhushi A. Stable and discriminating radiomic predictor of recurrence in early stage non-small cell lung cancer: Multi-site study. Lung Cancer 2020;142:90-7.
31. Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, Mast R, Suarez C, Sala-Llonch R, Calvo N, Escobar M, Navarro-Martin A, Villacampa G, Dienstmann R, Perez-Lopez R. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2021;31:1460-70.
32. Prayer F, Hofmanninger J, Weber M, Kifjak D, Willenpart A, Pan J, Röhrich S, Langs G, Prosch H. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. Methods 2021;188:98-104.
33. Vuong D, Bogowicz M, Denzler S, Oliveira C, Foerster R, Amstutz F, Gabryś HS, Unkelbach J, Hillinger S, Thierstein S, Xyrafas A, Peters S, Pless M, Guckenberger M, Tanadini-Lang S. Comparison of robust to standardized CT radiomics models to predict overall survival for non-small cell lung cancer patients. Med Phys 2020;47:4045-53.
34. Yamashita R, Perrin T, Chakraborty J, Chou JF, Horvat N, Koszalka MA, Midya A, Gonen M, Allen P, Jarnagin WR, Simpson AL, Do RKG. Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation. Eur Radiol 2020;30:195-205.
35. Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, Fyles A, Traverso A, Jaffray D, Metser U, Xie J, Han K. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. Radiother Oncol 2019;135:107-14.
36. Li Z, Duan H, Zhao K, Ding Y. Stability of MRI Radiomics Features of Hippocampus: An Integrated Analysis of Test-Retest and Inter-Observer Variability. IEEE Access 2019;7:97106-16.
37. Zinn PO, Singh SK, Kotrotsou A, Hassan I, Thomas G, Luedi MM, et al. A Coclinical Radiogenomic Validation Study: Conserved Magnetic Resonance Radiomic Appearance of Periostin-Expressing Glioblastoma in Patients and Xenograft Models. Clin Cancer Res 2018;24:6288-99.
38. Bologna M, Corino V, Tenconi C, Facchinetti N, Calareso G, Iacovelli N, Cavallo A, Alfieri S, Cavalieri S, Fallai C, Valdagni R, Rancati T, Trama A, Licitra L, Orlandi

 

E, Mainardi L. Methodology and technology for the development of a prognostic MRI-based radiomic model for the outcome of head and neck cancer patients. Annu Int Conf IEEE Eng Med Biol Soc 2020;2020:1152-5.
39. Jang J, Ngo LH, Mancio J, Kucukseymen S, Rodriguez J, Pierce P, Goddu B, Nezafat R. Reproducibility of Segmentation-based Myocardial Radiomic Features with Cardiac MRI. Radiol Cardiothorac Imaging 2020;2:e190216.
40. Merisaari H, Taimen P, Shiradkar R, Ettala O, Pesola M, Saunavaara J, Boström PJ, Madabhushi A, Aronen HJ, Jambor I. Repeatability of radiomics and machine learning for DWI: Short-term repeatability study of 112 patients with prostate cancer. Magn Reson Med 2020;83:2293-309.
41. Pandey U, Saini J, Kumar M, Gupta R, Ingalhalikar M. Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images. J Magn Reson Imaging 2021;53:394-407.
42. Gutmann DAP, Rospleszcz S, Rathmann W, Schlett CL, Peters A, Wachinger C, Gatidis S, Bamberg F. MRI-Derived Radiomics Features of Hepatic Fat Predict Metabolic States in Individuals without Cardiovascular Disease. Acad Radiol 2020. [Epub ahead of print]. doi: 10.1016/j.acra.2020.06.030.
43. Scalco E, Belfatto A, Mastropietro A, Rancati T, Avuzzi B, Messina A, Valdagni R, Rizzo G. T2w-MRI signal normalization affects radiomics features reproducibility. Med Phys 2020;47:1680-91.
44. Shiri I, Hajianfar G, Sohrabi A, Abdollahi H, P Shayesteh S, Geramifar P, Zaidi H, Oveisi M, Rahmim A. Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: Test-retest and image registration analyses. Med Phys 2020;47:4265-80.
45. Ta D, Khan M, Ishaque A, Seres P, Eurich D, Yang YH, Kalra S. Reliability of 3D texture analysis: A multicenter MRI study of the brain. J Magn Reson Imaging 2020;51:1200-9.
46. Zhang J, Yao K, Liu P, Liu Z, Han T, Zhao Z, Cao Y, Zhang G, Zhang J, Tian J, Zhou J. A radiomics model for preoperative prediction of brain invasion in meningioma non-invasively based on MRI: A multicentre study. EBioMedicine 2020;58:102933.
47. Han Y, Yang Y, Shi ZS, Zhang AD, Yan LF, Hu YC, Feng LL, Ma J, Wang W, Cui GB. Distinguishing brain inflammation from grade II glioma in population without contrast enhancement: a radiomics analysis based on conventional MRI. Eur J Radiol 2021;134:109467.
48. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, Hoekstra CJ, Boellaard R, Dekker AL, Gillies RJ, Aerts HJ, Lambin P. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52:1391-7.
49. Willaime JM, Turkheimer FE, Kenny LM, Aboagye EO. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. Phys Med Biol 2013;58:187-203.
50. van Velden FH, Nissen IA, Jongsma F, Velasquez LM, Hayes W, Lammertsma AA, Hoekstra OS, Boellaard R. Test-retest variability of various quantitative measures to characterize tracer uptake and/or tracer uptake heterogeneity in metastasized liver for patients with colorectal carcinoma. Mol Imaging Biol 2014;16:13-8.
51. Cheng NM, Fang YH, Tsan DL, Hsu CH, Yen TC. Respiration-Averaged CT for Attenuation Correction of PET Images - Impact on PET Texture Features in Non-Small Cell Lung Cancer Patients. PLoS One 2016;11:e0150509.
52. van Rossum PS, Fried DV, Zhang L, Hofstetter WL, van Vulpen M, Meijer GJ, Court LE, Lin SH. The Incremental Value of Subjective and Quantitative Assessment of 18F-FDG PET for the Prediction of Pathologic Complete Response to Preoperative Chemoradiotherapy in Esophageal Cancer. J Nucl Med 2016;57:691-700.
53. Carvalho S, Leijenaar RTH, Troost EGC, van Timmeren JE, Oberije C, van Elmpt W, de Geus-Oei LF, Bussink J, Lambin P. 18F-fluorodeoxyglucose positron-emission tomography (FDG-PET)-Radiomics of metastatic lymph nodes and primary tumor in non-small cell lung cancer (NSCLC) - A prospective externally validated study. PLoS One 2018;13:e0192859.
54. Jiang Y, Yuan Q, Lv W, Xi S, Huang W, Sun Z, Chen H, Zhao L, Liu W, Hu Y, Lu L, Ma J, Li T, Yu J, Wang Q, Li G. Radiomic signature of (18)F fluorodeoxyglucose PET/CT for prediction of gastric cancer survival and chemotherapeutic benefits. Theranostics 2018;8:5915-28.
55. Lin C, Harmon S, Bradshaw T, Eickhoff J, Perlman S, Liu G, Jeraj R. Response-to-repeatability of quantitative imaging features for longitudinal response assessment. Phys Med Biol 2019;64:025019.
56. Manabe O, Ohira H, Hirata K, Hayashi S, Naya M, Tsujino I, Aikawa T, Koyanagawa K, Oyama-Manabe N, Tomiyama Y, Magota K, Yoshinaga K, Tamaki N. Use of

 

(18)F-FDG PET/CT texture analysis to diagnose cardiac sarcoidosis. Eur J Nucl Med Mol Imaging 2019;46:1240-7.
57. Vuong D, Tanadini-Lang S, Huellner MW, Veit-Haibach P, Unkelbach J, Andratschke N, Kraft J, Guckenberger M, Bogowicz M. Interchangeability of radiomic features between [18F]-FDG PET/CT and [18F]-FDG PET/MR. Med Phys 2019;46:1677-85.
58. Desseroit MC, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, Hatt M. Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non-Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort. J Nucl Med 2017;58:406-11.
59. Jiang J, Zhou H, Duan H, Liu X, Zuo C, Huang Z, Yu Z, Yan Z; Alzheimer's Disease Neuroimaging Initiative. A novel individual-level morphological brain networks constructing method and its evaluation in PET and MR images. Heliyon 2017;3:e00475.
60. Ahn SJ, Kim JH, Lee SM, Park SJ, Han JK. CT reconstruction algorithms affect histogram and texture analysis: evidence for liver parenchyma, focal solid liver lesions, and renal cysts. Eur Radiol 2019;29:4008-15.
61. Kolossváry M, Szilveszter B, Karády J, Drobni ZD, Merkely B, Maurovich-Horvat P. Effect of image reconstruction algorithms on volumetric and radiomic parameters of coronary plaques. J Cardiovasc Comput Tomogr 2019;13:325-30.
62. Koo HJ, Sung YS, Shim WH, Xu H, Choi CM, Kim HR, Lee JB, Kim MY. Quantitative Computed Tomography Features for Predicting Tumor Recurrence in Patients with Surgically Resected Adenocarcinoma of the Lung. PLoS One 2017;12:e0167955.
63. Lee SH, Cho HH, Lee HY, Park H. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. Cancer Imaging 2019;19:54.
64. Suter Y, Knecht U, Alão M, Valenzuela W, Hewer E, Schucht P, Wiest R, Reyes M. Radiomics for glioblastoma survival analysis in pre-operative MRI: exploring feature robustness, class boundaries, and machine learning techniques. Cancer Imaging 2020;20:55.
65. Altazi BA, Zhang GG, Fernandez DC, Montejo ME, Hunt D, Werner J, Biagioli MC, Moros EG. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. J Appl Clin Med Phys 2017;18:32-48.
66. van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, Hoekstra OS, Smit EF, Boellaard R. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [(18)F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. Mol Imaging Biol 2016;18:788-95.
67. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, Mitra S, Shankar BU, Kikinis R, Haibe-Kains B, Lambin P, Aerts HJ. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9:e102107.
68. Echegaray S, Gevaert O, Shah R, Kamaya A, Louie J, Kothary N, Napel S. Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. J Med Imaging (Bellingham) 2015;2:041011.
69. Echegaray S, Nair V, Kadoch M, Leung A, Rubin D, Gevaert O, Napel S. A Rapid Segmentation-Insensitive "Digital Biopsy" Method for Radiomic Feature Extraction: Method and Pilot Study Using CT Images of Non-Small Cell Lung Cancer. Tomography 2016;2:283-94.
70. Qiu Q, Duan J, Gong G, Lu Y, Li D, Lu J. Reproducibility of radiomic features with GrowCut and GraphCut semiautomatic tumor segmentation in hepatocellular carcinoma. Transl Cancer Res 2017;6:940-8.
71. Owens CA, Peterson CB, Tang C, Koay EJ, Yu W, Mackin DS, Li J, Salehpour MR, Fuentes DT, Court LE, Yang J. Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer. PLoS One 2018;13:e0205003.
72. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, Roesch J, Rudofsky L, Friess M, Veit-Haibach P, Huellner M, Opitz I, Weder W, Frauenfelder T, Guckenberger M, Tanadini-Lang S. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol 2018;57:1070-4.
73. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning-based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. Eur Radiol 2019;29:4765-75.
74. Kocak B, Durmaz ES, Kaya OK, Ates E, Kilickesmez O. Reliability of Single-Slice-Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. AJR Am J Roentgenol 2019;213:377-83.
75. Moltz JH (editor). Stability of radiomic features of liver lesions from manual delineation in CT scans. Medical Imaging 2019: Computer-Aided Diagnosis. Bellingham,

 

Washington: International Society for Optics and Photonics, 2019.
76. Mori M, Benedetti G, Partelli S, Sini C, Andreasi V, Broggi S, Barbera M, Cattaneo GM, Muffatti F, Panzeri M, Falconi M, Fiorino C, De Cobelli F. Ct radiomic features of pancreatic neuroendocrine neoplasms (panNEN) are robust against delineation uncertainty. Phys Med 2019;57:41-6.
77. Qiu Q, Duan J, Duan Z, Meng X, Ma C, Zhu J, Lu J, Liu T, Yin Y. Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability. Quant Imaging Med Surg 2019;9:453-64.
78. Sung P, Lee JM, Joo I, Lee S, Kim TH, Ganeshan B. Evaluation of the Impact of Iterative Reconstruction Algorithms on Computed Tomography Texture Features of the Liver Parenchyma Using the Filtration-Histogram Method. Korean J Radiol 2019;20:558-68.
79. Uthoff J, Nagpal P, Sanchez R, Gross TJ, Lee C, Sieren JC. Differentiation of non-small cell lung cancer and histoplasmosis pulmonary nodules: insights from radiomics model performance compared with clinician observers. Transl Lung Cancer Res 2019;8:979-88.
80. Caballo M, Pangallo DR, Mann RM, Sechopoulos I. Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence. Comput Biol Med 2020;118:103629.
81. Haarburger C, Muller-Franzes G, Weninger L, Kuhl C, Truhn D, Merhof D. Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. Sci Rep 2020;10:12688.
82. Kakino R, Nakamura M, Mitsuyoshi T, Shintani T, Hirashima H, Matsuo Y, Mizowaki T. Comparison of radiomic features in diagnostic CT images with and without contrast enhancement in the delayed phase for NSCLC patients. Phys Med 2020;69:176-82.
83. Kulkarni A, Carrion-Martinez I, Dhindsa K, Alaref AA, Rozenberg R, van der Pol CB. Pancreas adenocarcinoma CT texture analysis: comparison of 3D and 2D tumor segmentation techniques. Abdom Radiol (NY) 2021;46:1027-33.
84. Liu R, Elhalawani H, Radwan Mohamed AS, Elgohari B, Court L, Zhu H, Fuller CD. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. Clin Transl Radiat Oncol 2019;21:11-8.
85. Nguyen K, Schieda N, James N, McInnes MDF, Wu M, Thornhill RE. Effect of phase of enhancement on texture analysis in renal masses evaluated with non-contrast-enhanced, corticomedullary, and nephrographic phase-enhanced CT images. Eur Radiol 2021;31:1676-86.
86. Ren J, Yuan Y, Qi M, Tao X. Machine learning-based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation. Eur Radiol 2020;30:6858-66.
87. Adduru VR, Michael AM, Helguera M, Baum SA, Moore GJ. Leveraging Clinical Imaging Archives for Radiomics: Reliability of Automated Methods for Brain Volume Measurement. Radiology 2017;284:862-9.
88. Lee M, Woo B, Kuo MD, Jamshidi N, Kim JH. Quality of Radiomic Features in Glioblastoma Multiforme: Impact of Semi-Automated Tumor Segmentation Software. Korean J Radiol 2017;18:498-509.
89. Bologna M, Corino VDA, Montin E, Messina A, Calareso G, Greco FG, Sdao S, Mainardi LT. Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images. J Digit Imaging 2018;31:879-94.
90. Saha A, Harowicz MR, Mazurowski MA. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. Med Phys 2018;45:3076-85.
91. Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik JC, Thomassin-Naggara I, Fournier L, Lecler A. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One 2019;14:e0213459.
92. Koçak B. Reliability of 2D Magnetic Resonance Imaging Texture Analysis in Cerebral Gliomas: Influence of Slice Selection Bias on Reproducibility of Radiomic Features. Istanbul Med J 2019;20. doi: 10.4274/imj.galenos.2019.09582
93. Lecler A, Duron L, Balvay D, Savatovsky J, Bergès O, Zmuda M, Farah E, Galatoire O, Bouchouicha A, Fournier LS. Combining Multiple Magnetic Resonance Imaging Sequences Provides Independent Reproducible Radiomics Features. Sci Rep 2019;9:2068.
94. Tixier F, Um H, Young RJ, Veeraraghavan H. Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features. Med Phys 2019;46:3582-91.
95. Traverso A, Kazmierski M, Welch ML, Weiss J, Fiset S, Foltz WD, Gladwish A, Dekker A, Jaffray D, Wee L, Han K. Sensitivity of radiomic features to inter-observer

 

variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients. Radiother Oncol 2020;143:88-94.
96. Alis D, Yergin M, Asmakutlu O, Topel C, Karaarslan E. The influence of cardiac motion on radiomics features: radiomics features of non-enhanced CMR cine images greatly vary through the cardiac cycle. Eur Radiol 2021;31:2706-15.
97. Chen H, He Y, Zhao C, Zheng L, Pan N, Qiu J, Zhang Z, Niu X, Yuan Z. Reproducibility of radiomics features derived from intravoxel incoherent motion diffusion-weighted MRI of cervical cancer. Acta Radiol 2021;62:679-86.
98. Granzier RWY, Verbakel NMH, Ibrahim A, van Timmeren JE, van Nijnatten TJA, Leijenaar RTH, Lobbes MBI, Smidt ML, Woodruff HC. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. Sci Rep 2020;10:14163.
99. Lin YC, Lin CH, Lu HY, Chiang HJ, Wang HK, Huang YT, Ng SH, Hong JH, Yen TC, Lai CH, Lin G. Deep learning for fully automated tumor segmentation and extraction of magnetic resonance radiomics features in cervical cancer. Eur Radiol 2020;30:1297-305.
100. Pati S, Verma R, Akbari H, Bilello M, Hill VB, Sako C, Correa R, Beig N, Venet L, Thakur S, Serai P, Ha SM, Blake GD, Shinohara RT, Tiwari P, Bakas S. Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset. Med Phys 2020;47:6039-52.
101. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, Chen W. Robustness of Radiomic Features in [(11)C]Choline and [(18)F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization. Mol Imaging Biol 2016;18:935-45.
102. Bashir U, Azad G, Siddique MM, Dhillon S, Patel N, Bassett P, Landau D, Goh V, Cook G. The effects of segmentation algorithms on the measurement of (18) F-FDG PET texture parameters in non-small cell lung cancer. EJNMMI Res 2017;7:60.
103. Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, Fallanca F, Passoni P, Vanoli EG, Calandrino R, Di Muzio N, Picchio M, Fiorino C. Quantifying the robustness of [(18)F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. Phys Med 2018;49:105-11.
104. Yang P, Xu L, Cao Z, Wan Y, Xue Y, Jiang Y, Yen E, Luo C, Wang J, Rong Y, Niu T. Extracting and Selecting Robust Radiomic Features from PET/MR Images in Nasopharyngeal Carcinoma. Mol Imaging Biol 2020;22:1581-91.
105. Bogowicz M, Riesterer O, Bundschuh RA, Veit-Haibach P, Hüllner M, Studer G, Stieb S, Glatz S, Pruschy M, Guckenberger M, Tanadini-Lang S.. Stability of radiomic features in CT perfusion maps. Phys Med Biol 2016;61:8736-49.
106. Ger RB, Zhou S, Chi PM, Lee HJ, Layman RR, Jones AK, Goff DL, Fuller CD, Howell RM, Li H, Stafford RJ, Court LE, Mackin DS. Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. Sci Rep 2018;8:13047.
107. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep 2018;8:10545.
108. Defeudis A, De Mattia C, Rizzetto F, Calderoni F, Mazzetti S, Torresin A, Vanzulli A, Regge D, Giannini V. Standardization of CT radiomics features for multi-center analysis: impact of software settings and parameters. Phys Med Biol 2020;65:195012.
109. Park BW, Kim JK, Heo C, Park KJ. Reliability of CT radiomic features reflecting tumour heterogeneity according to image quality and image processing parameters. Sci Rep 2020;10:3852.
110. Kim D, Wang N, Ravikumar V, Raghuram DR, Li J, Patel A, Wendt RE 3rd, Rao G, Rao A. Prediction of 1p/19q Codeletion in Diffuse Glioma Patients Using Pre-operative Multiparametric Magnetic Resonance Imaging. Front Comput Neurosci 2019;13:52.
111. Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, Aerts HJWL, Kikinis R, Fennessy FM, Fedorov A. Repeatability of Multiparametric Prostate MRI Radiomics Features. Sci Rep 2019;9:9441.
112. Fan M, Liu Z, Xu M, Wang S, Zeng T, Gao X, Li L. Generative adversarial network-based super-resolution of diffusion-weighted imaging: Application to tumour radiomics in breast cancer. NMR Biomed 2020;33:e4345.
113. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys 2020;21:179-90.
114. Branchini M, Zorz A, Zucchetta P, Bettinelli A, De Monte F, Cecchin D, Paiusco M. Impact of acquisition count statistics reduction and SUV discretization on PET radiomic features in pediatric 18F-FDG-PET/MRI examinations. Phys Med 2019;59:117-26.

 

115. Whybra P, Parkinson C, Foley K, Staffurth J, Spezi E. Assessing radiomic feature robustness to interpolation in (18)F-FDG PET imaging. Sci Rep 2019;9:9649.
116. Foy JJ, Robinson KR, Li H, Giger ML, Al-Hallaq H, Armato SG 3rd. Variation in algorithm implementation across radiomics software. J Med Imaging (Bellingham) 2018;5:044505.
117. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 2012;53:693-700.
118. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, Aerts HJ, Gillies RJ, Lambin P. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep 2015;5:11075.
119. Bogowicz M, Leijenaar RTH, Tanadini-Lang S, Riesterer O, Pruschy M, Studer G, Unkelbach J, Guckenberger M, Konukoglu E, Lambin P. Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models. Radiother Oncol 2017;125:385-91.
120. Lv W, Yuan Q, Wang Q, Ma J, Jiang J, Yang W, Feng Q, Chen W, Rahmim A, Lu L. Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT. Eur Radiol 2018;28:3245-54.
121. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, Wahl RL, Kurland BF, Schwarz AJ, Gönen M, Zahlmann G, Kondratovich MV, O'Donnell K, Petrick N, Cole PE, Garra B, Sullivan DC; QIBA Technical Performance Working Group. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res 2015;24:27-67.
122. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, Sabater S. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology 2018;288:407-15.
123. Mannil M, von Spiczak J, Hermanns T, Alkadhi H, Fankhauser CD. Prediction of successful shock wave lithotripsy with CT: a phantom study using texture analysis. Abdom Radiol (NY) 2018;43:1432-8.
124. Li Y, Tan G, Vangel M, Hall J, Cai W. Influence of feature calculating parameters on the reproducibility of CT radiomic features: a thoracic phantom study. Quant Imaging Med Surg 2020;10:1775-85.
125. Nardone V, Reginelli A, Guida C, Belfiore MP, Biondi M, Mormile M, Banci Buonamici F, Di Giorgio E, Spadafora M, Tini P, Grassi R, Pirtoli L, Correale P, Cappabianca S, Grassi R. Delta-radiomics increases multicentre reproducibility: a phantom study. Med Oncol 2020;37:38.
126. Song YS, Park CM, Lee SM, Park SJ, Cho HR, Choi SH, Lee JM, Kiefer B, Goo JM. Reproducibility of histogram and texture parameters derived from intravoxel incoherent motion diffusion-weighted MRI of FN13762 rat breast Carcinomas. Anticancer Res 2014;34:2135-44.
127. Baeßler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Invest Radiol 2019;54:221-8.
128. Bologna M, Corino VDA, Mainardi LT. Assessment of the effect of intensity standardization on the reliability of T1-weighted MRI radiomic features: experiment on a virtual phantom. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:413-6.
129. Bologna M, Corino V, Mainardi L. Technical Note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain. Med Phys 2019;46:5116-23.
130. Cattell R, Chen S, Huang C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis Comput Ind Biomed Art 2019;2:19.
131. Bianchini L, Botta F, Origgi D, Rizzo S, Mariani M, Summers P, García-Polo P, Cremonesi M, Lascialfari A. PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. Phys Med 2020;71:71-81.
132. Dreher C, Kuder TA, König F, Mlynarska-Bujny A, Tenconi C, Paech D, Schlemmer HP, Ladd ME, Bickelhaupt S. Radiomics in diffusion data: a test-retest, inter- and intra-reader DWI phantom study. Clin Radiol 2020;75:798 e13-22.
133. Eresen A, Yang J, Shangguan J, Benson AB, Yaghmai V, Zhang Z. Detection of Immunotherapeutic Response in a Transgenic Mouse Model of Pancreatic Ductal Adenocarcinoma Using Multiparametric MRI Radiomics: A Preliminary Investigation. Acad Radiol 2021;28:e147-54.
134. Rai R, Holloway LC, Brink C, Field M, Christiansen RL, Sun Y, Barton MB, Liney GP. Multicenter evaluation of MRI-based radiomic features: A phantom study. Med Phys 2020;47:3054-63.
135. Bianchini L, Santinha J, Loução N, Figueiredo M, Botta F, Origgi D, Cremonesi M, Cassano E, Papanikolaou N, Lascialfari A. A multicenter study on radiomic features from T2 -weighted images of a customized MR pelvic

 

phantom setting the basis for robust radiomic models in clinics. Magn Reson Med 2021;85:1713-26.
136. Gallivanone F, Interlenghi M, D'Ambrosio D, Trifiro G, Castiglioni I. Parameters Influencing PET Imaging Features: A Phantom Study with Irregular and Heterogeneous Synthetic Lesions. Contrast Media Mol Imaging 2018;2018:5324517.
137. Ger RB, Meier JG, Pahlka RB, Gay S, Mumme R, Fuller CD, Li H, Howell RM, Layman RR, Stafford RJ, Zhou S, Mawlawi O, Court LE. Effects of alterations in positron emission tomography imaging parameters on radiomics features. PLoS One 2019;14:e0221877.
138. Pfaehler E, Beukinga RJ, de Jong JR, Slart RHJA, Slump CH, Dierckx RAJO, Boellaard R. Repeatability of (18) F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. Med Phys 2019;46:665-78.
139. Pfaehler E, van Sluis J, Merema BBJ, van Ooijen P, Berendsen RCM, van Velden FHP, Boellaard R. Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. J Nucl Med 2020;61:469-76.
140. Yang F, Simpson G, Young L, Ford J, Dogan N, Wang L. Impact of contouring variability on oncological PET radiomics features in the lung. Sci Rep 2020;10:369.
141. Qiao M, Li C, Suo S, Cheng F, Hua J, Xue D, Guo Y, Xu J, Wang Y. Breast DCE-MRI radiomics: a robust computer-aided system based on reproducible BI-RADS features across the influence of datasets bias and segmentation methods. Int J Comput Assist Radiol Surg 2020;15:921-30.
142. Qiao M, Li C, Suo S, Cheng F, Hua J, Xue D, Guo Y, Xu J, Wang Y. Diffusion and perfusion MRI radiomics obtained from deep learning segmentation provides reproducible and comparable diagnostic model to human in post-treatment glioblastoma. Eur Radiol 2020;15:921-30.
143. Kunimatsu A, Kunimatsu N, Kamiya K, Watadani T, Mori H, Abe O. Comparison between Glioblastoma and Primary Central Nervous System Lymphoma Using MR Image-based Texture Analysis. Magn Reson Med Sci 2018;17:50-7.
144. Sutton EJ, Huang EP, Drukker K, Burnside ES, Li H, Net JM, Rao A, Whitman GJ, Zuley M, Ganott M, Bonaccio E, Giger ML, Morris EA; TCGA group. Breast MRI radiomics: comparison of computer-and human-extracted imaging phenotypes. Eur Radiol Exp 2017;1:22.
145. Hinzpeter R, Wagner MW, Wurnig MC, Seifert B, Manka R, Alkadhi H. Texture analysis of acute myocardial infarction with CT: First experience study. PLoS One 2017;12:e0186876.
146. Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295:328-38.
147. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, Huang SH, Purdie TG, O'Sullivan B, Aerts HJWL, Jaffray DA. Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol 2019;130:2-9.
148. Summers RM. Are we at a crossroads or a plateau? Radiomics and machine learning in abdominal oncology imaging. Abdom Radiol (NY) 2019;44:1985-9.
149. Stanzione A, Gambardella M, Cuocolo R, Ponsiglione A, Romeo V, Imbriaco M. Prostate MRI radiomics: A systematic review and radiomic quality score assessment. Eur J Radiol 2020;129:109095.
150. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, Shin JH, Kim JH. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 2020;30:523-36.
151. Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, Lambin P. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother Oncol 2018;127:349-60.
152. Wakabayashi T, Ouhmich F, Gonzalez-Cabrera C, Felli E, Saviano A, Agnus V, Savadjiev P, Baumert TF, Pessaux P, Marescaux J, Gallix B. Radiomics in hepatocellular carcinoma: a quantitative review. Hepatol Int 2019;13:546-59.
153. Wang H, Zhou Y, Li L, Hou W, Ma X, Tian R. Current status and quality of radiomics studies in lymphoma: a systematic review. Eur Radiol 2020;30:6228-40.
154. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud 2011;48:661-71.
155. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med 2012;31:3972-81.
156. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30.

 

157. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
158. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, Lambin P. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography 2016;2:361-5.
159. Sullivan DC, Obuchowski NA, Kessler LG, Raunig DL, Gatsonis C, Huang EP, Kondratovich M, McShane LM, Reeves AP, Barboriak DP, Guimaraes AR, Wahl RL; RSNA-QIBA Metrology Working Group. Metrology Standards for Quantitative Imaging Biomarkers. Radiology 2015;277:813-25.
160. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. Available online: https://arxivorg/abs/161207003
161. Kopp-Schneider A, Hielscher T. How to evaluate agreement between quantitative measurements. Radiother Oncol 2019;141:321-6.
162. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm 2013;9:330-8.
163. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PLoS One 2012;7:e37908.
164. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008;31:466-75.

**Cite this article as:** Xue C, Yuan J, Lo GG, Chang ATY, Poon DMC, Wong OL, Zhou Y, Chu WCW. Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review. Quant Imaging Med Surg 2021;11(10):4431-4460. doi: 10.21037/qims-21-86

 

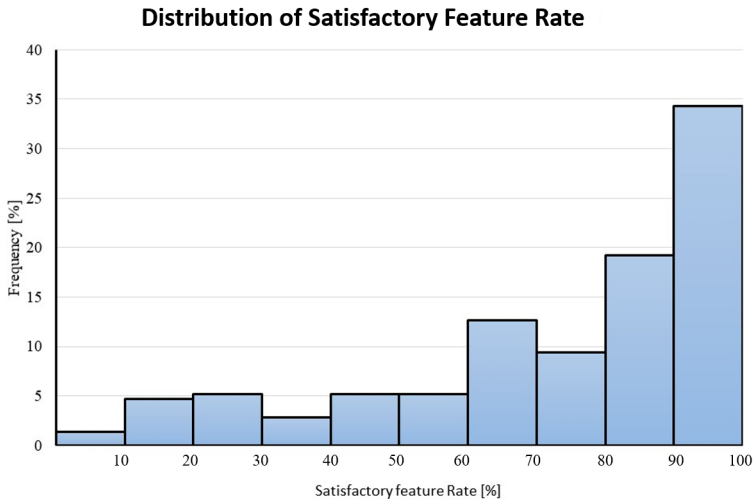


**Figure S1** Satisfactory feature rate distribution in the studies reported ICC results regarding the feature reliability due to image segmentation for intra-/inter-observer agreement.

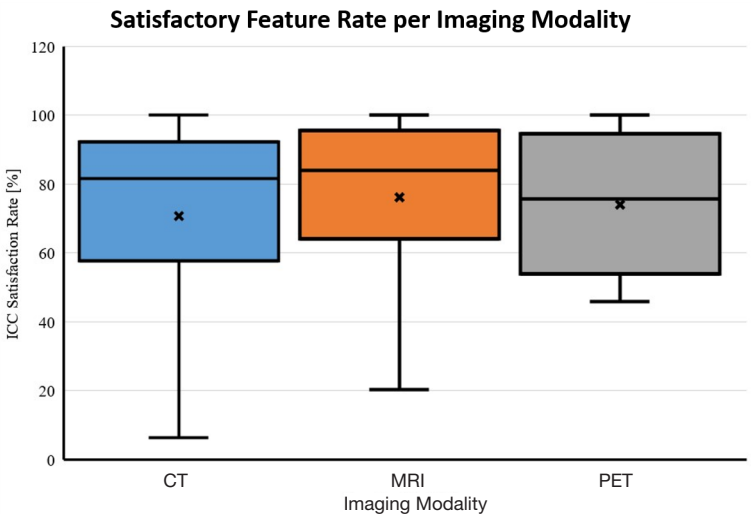


**Figure S2** The difference of satisfactory feature rate in different imaging modality in the studies reported ICC results regarding the feature reliability due to image segmentation for intra-/inter-observer agreement.


https://dx.doi.org/10.21037/qims-21-86

**Table S1** The available ICC forms described in the studies

| PMID/doi | First author | Publication year | Disease or organ | Patient number | ICC purpose | ICC form described in the article | ICC model | ICC type | ICC definition |
|---|---|---|---|---|---|---|---|---|---|
| 23807457 | van Velden FHP et al. | 2014 | Colorectal cancer | 29 | Acquisition | 2-way random single measure model | Yes | Yes | No |
| 24047337 | Leijenaar RTH et al. | 2013 | Lung cancer | 34 | Acquisition | ICC(1,1) for test -retest reliability and ICC(3,k) for inter-rater simulation | Yes | Yes | Yes |
| 25025374 | Parmar C et al. | 2014 | Lung cancer | 20 | Segmentation | Inter-observer segmentation (two-way mixed effect model, case 3A(McGraw), absolute agreement), intra-observer segmentation (one-way model, case 1(McGraw), form-specific formula given | Yes | Yes | Yes |
| 26163091 | Panth KM et al. | 2015 | Phantom | 23 | Segmentation | ICC(1,1), ICC(2,1) | Yes | Yes | Yes |
| 26242464 | Leijenaar RTH et al. | 2015 | Lung cancer | 35 | Quantification | Formula given, but cannot be specified to a certain form | No | No | No |
| 26587549 | Echegaray S et al. | 2015 | Liver cancer | 29 | Segmentation | A1 method (McGraw) | Yes | Yes | Yes |
| 26795288 | Rossum PSNR et al. | 2016 | HNC | 217 | Acquisition | An absolute agreement definition in a 2-way mixed-effects model | Yes | No | Yes |
| 26920355 | van Velden FHP et al. | 2016 | Lung cancer | 11 | Reconstruction | One-way random single-measure | Yes | Yes | No |
| 27669756 | Hu P et al. | 2016 | Colorectal cancer | 40 | Acquisition | Two-way mixed effect model, with formula given | Yes | Yes | Yes |
| 27893446 | Bogowicz M et al. | 2016 | HNC and lung cancer | 22 | Processing | Two-way mixed model, consistency | Yes | No | Yes |
| 28612050 | Echegaray S et al. | 2016 | Lung cancer | 100 | Segmentation | A1 method (McGraw) | Yes | Yes | Yes |
| 29122358 | Bogowicz M et al. | 2017 | HNC | 178 | Quantification | Two-way mixed single measures | Yes | Yes | Yes |
| 29494598 | Carvalho S et al. | 2018 | Lung cancer | 215 | Acquisition | ICC(1,k) | Yes | Yes | Yes |
| 29513054 | Pavic M et al. | 2018 | HNC and Lung Cancer | 11 | Segmentation | ICC(3,1), consistency | Yes | Yes | Yes |
| 29520429 | Lv W et al. | 2018 | NPC | 106 | Quantification | Formula given, but cannot be specified to a certain form | No | No | No |
| 29633002 | Shi Z et al. | 2018 | Vascular disease | 96 | Segmentation | Two-way random model with absolute measurements | Yes | Yes | Yes |
| 29663411 | Saha A et al. | 2018 | Breast cancer | 50 | Segmentation | ICC(3,1), consistency | Yes | Yes | Yes |
| 29725965 | Bologna M et al. | 2018 | HNC | 36 | Segmentation | Two-way mixed effect model | Yes | No | No |
| 30002441 | Hassan M et al. | 2018 | Lung cancer | 18 | Processing | Form-specific formula given | Yes | Yes | Yes |
| 30135549 | Bibault J et al. | 2018 | Colorectal cancer | 95 | Segmentation | Two-way mixed effect model, formula given | Yes | Yes | Yes |
| 30158540 | Ger RB et al. | 2018 | HNC and lung cancer | 50 | Processing | ICC (2, 1) (two-way random effects, absolute agreement, single ratter/measurement) and ICC (3, 1) (two-way random effects, consistency, single rater/measurement) | Yes | Yes | Yes |
| 30230556 | Soufi M et al. | 2018 | Lung cancer | 305 | Acquisition | Two-ray random effect model, absolute agreement condition | Yes | No | Yes |
| 30286184 | Owens CA et al. | 2018 | Lung cancer | 10 | Segmentation | Two-way mixed-effects model | Yes | No | No |
| 30463529 | Zheng B et al. | 2018 | Liver cancer | 319 | Segmentation | A two-way random, single measure (absolute agreement) | Yes | Yes | Yes |

**Table S1** (*continued*)


https://dx.doi.org/10.21037/qims-21-86

**Table S1** (*continued*)

| PMID/doi | First author | Publication year | Disease or organ | Patient number | ICC purpose | ICC form described in the article | ICC model | ICC type | ICC definition |
|---|---|---|---|---|---|---|---|---|---|
| 30506687 | Pfaehler E et al. | 2019 | Phantom | 1 | Reconstruction, segmentation | Two-way single measure model for consistency | Yes | Yes | Yes |
| 30679599 | Alex Z et al. | 2019 | HNC and Lung Cancer | 50 | Acquisition | ICC(1,1) | Yes | Yes | Yes |
| 30714158 | Vuong D et al. | 2019 | Lung cancer | 19 | Acquisition | ICC(1,1) for 4DPETMR, ICC(3,1) for PETCT-PETMR | Yes | Yes | Yes |
| 30738530 | Mori M et al. | 2019 | Pancreatic cancer | 31 | Segmentation | One-way random single-measure model | Yes | Yes | No |
| 30765732 | Lecler A et al. | 2019 | Breast cancer | 207 | Segmentation | ICC(2,1), absolute agreement | Yes | Yes | Yes |
| 30840747 | Foy JJ et al. | 2018 | Healthy | 39 | Quantification | ICC(a,1) | Yes | Yes | Yes |
| 30845221 | Duron L et al. | 2019 | Breast cancer | 104 | Segmentation | 2-way random intraclass correlation coefficient (ICC) (absolute agreement, average type) | Yes | Yes | Yes |
| 30928060 | Branchini M et al. | 2019 | Pediatric | 21 | Processing | One way random single measure model | Yes | Yes | No |
| 30961895 | Peeken JC et al. | 2019 | Soft tissue sarcoma | 221 | Segmentation | ICC(3,1) | Yes | Yes | Yes |
| 31015155 | Fiset S et al. | 2019 | Cervical cancer | 62 | Acquisition | ICC(1,1) for test -retest reliability and diagnostic-simulation, ICC(2,1) for inter-observer | Yes | Yes | Yes |
| 31032192 | Qiu Q et al. | 2019 | Liver cancer | 106 | Segmentation | Form-specific formula given | Yes | Yes | Yes |
| 31063427 | Kocak B et al. | 2019 | Kidney cancer | 30 | Segmentation | Two-way model, single-rating, and absolute agreement | Yes | Yes | Yes |
| 31131906 | Tixier F et al. | 2019 | GBM | 90 | Segmentation | Two-way random effects model | Yes | No | No |
| 31273242 | Whybra P et al. | 2019 | HNC | 441 | Processing | 2-way mixed-effects model, single rater, absolute agreement | Yes | Yes | Yes |
| 31375452 | Huang L et al. | 2019 | Lung cancer | 31 | Acquisition | 2-way fixed effect, absolute agreement, and single measurement model | Yes | Yes | Yes |
| 31375883 | Ugga L et al. | 2019 | HNC | 89 | Segmentation | Absolute agreement ICC value | No | No | Yes |
| 31392481 | Yamashita R et al. | 2020 | Pancreatic cancer | 55 | Acquisition | Mixed effects linear model | Yes | No | No |
| 31420497 | Pfaehler E et al. | 2020 | Phantom | 3 | Acquisition, quantification | 2-way single-measure model for consistency of features | Yes | Yes | Yes |
| 31423714 | Ta D et al. | 2020 | Healthy | 6 | Acquisition | Two-way mixed-effects model | Yes | No | No |
| 31477335 | Traverso A et al. | 2020 | Cervical cancer | 81 | Processing | ICC(2,1) | Yes | Yes | Yes |
| 31539450 | Bologna M et al. | 2019 | Phantom | 1 | Acquisition, processing | Measure agreement in mixed-effect models, equivalent to the (A,1) model described in McGraw et al | Yes | Yes | Yes |
| 31703155 | Merisaari H et al. | 2020 | Prostate cancer | 112 | Acquisition | ICC(3,1) | Yes | Yes | Yes |
| 31743889 | Cong M et al. | 2020 | Lung cancer | 50 | Segmentation | Absolute agreement method | No | No | Yes |
| 31761666 | Wang Y et al. | 2019 | Gastric cancer | 30 | Segmentation | Two-way mixed effects, absolute agreement, single rater ICC test | Yes | Yes | Yes |

**Table S1** (*continued*)


https://dx.doi.org/10.21037/qims-21-86

**Table S1** (*continued*)

| PMID/doi | First author | Publication year | Disease or organ | Patient number | ICC purpose | ICC form described in the article | ICC model | ICC type | ICC definition |
|---|---|---|---|---|---|---|---|---|---|
| 31824843 | Wu J et al. | 2019 | Colorectal cancer | 102 | Segmentation | Single rater; definition: absolute agreement; model: inter-user ICC: two-way random effects; intra-user ICC: two-way mixed effects | Yes | Yes | Yes |
| 31824847 | Chen W et al. | 2019 | Gastric cancer | 30 | Segmentation | Intraclass (single-rating, consistency, 2-way mixed effects model), interclass (multiple-rating, consistency, 2-way random-effects model) | Yes | Yes | Yes |
| 31903240 | Xu X et al. | 2019 | Lung cancer | 132 | Segmentation | Two-way fixed effect, absolute agreement, and single measurement | Yes | Yes | Yes |
| 31918370 | Kakino R et al. | 2020 | Lung cancer | 256 | Segmentation | Two-way random effects model under an absolute agreement condition | Yes | No | Yes |
| 31941949 | Yang F et al. | 2020 | Phantom | 1 | Segmentation | 2-way random effect model to estimate the absolute agreement of multiple raters per measurement | Yes | Yes | Yes |
| 31971614 | Scalco E et al. | 2020 | Prostate cancer | 14 | Acquisition | Two-way mixed effect model, single rater type, both consistency and absolute agreement were used | Yes | Yes | Yes |
| 32123281 | Park BW et al. | 2020 | Bladder cancer | 83 | Processing | ICC(c,1) | Yes | Yes | Yes |
| 32174316 | Caballo M et al. | 2020 | Breast cancer | 93 | Segmentation | ICC(3,1) | Yes | Yes | Yes |
| 32229081 | Palle S et al. | 2020 | Endometrial cancer | 30 | Segmentation | Two-way random effects, single rater, agreement | Yes | Yes | Yes |
| 32236847 | Nardone V et al. | 2020 | Phantom | 1 | Acquisition | Mean rating, absolute agreement, and two-way mixed effects model | Yes | Yes | Yes |
| 32240418 | Cattell R et al. | 2019 | Phantom | 5 | Quantification | 2-way mixed-effects model, single rater, absolute agreement | Yes | Yes | Yes |
| 32277703 | Rai R et al. | 2020 | Phantom | 11 | Acquisition | Two-way random effect with absolute agreement model | Yes | No | Yes |
| 32395833 | Vuong D et al. | 2020 | Lung cancer | 124 | Acquisition | ICC(3,1) | Yes | Yes | Yes |
| 32399621 | Haider SP et al. | 2020 | HNC | 50 | Segmentation | Two-way mixed effects, absolute agreement, single rater | Yes | Yes | Yes |
| 32557189 | Yang P et al. | 2020 | NPC | 21 | Processing | ICC(3,10) | Yes | Yes | Yes |
| 32593138 | Ramon R et al. | 2020 | GBM | 100 | Segmentation | Two-way random effects, absolute agreement, single rater/measurement | Yes | Yes | Yes |
| 32705290 | Cuocolo R et al. | 2020 | HNC | 89 | Segmentation | Two-way, absolute agreement and single rater ICC | Yes | Yes | Yes |
| 32723501 | Dreher C et al. | 2020 | Phantom | 17 | Segmentation | Two-way mixed model for absolute agreement and single measures | Yes | Yes | Yes |
| 32728098 | Haarburger C et al. | 2020 | Kidney cancer, liver cancer, lung cancer | 1216 | Segmentation | ICC(1,1) | Yes | Yes | Yes |
| 32737518 | Cysouw MCF et al. | 2021 | Prostate cancer | 76 | Segmentation | Two-way mixed effect model, absolute agreement | Yes | No | Yes |

**Table S1** (*continued*)


https://dx.doi.org/10.21037/qims-21-86

**Table S1** (*continued*)

| PMID/doi | First author | Publication year | Disease or organ | Patient number | ICC purpose | ICC form described in the article | ICC model | ICC type | ICC definition |
|---|---|---|---|---|---|---|---|---|---|
| 32758279 | Suter Y et al. | 2020 | GBM | 63 | Reconstruction | ICC(2,1), absolute agreement | Yes | Yes | Yes |
| 32767049 | Song X et al. | 2021 | Ovarian cancer | 104 | Segmentation | Single-rater, absolute- agreement, 2-way mixed-effects model | Yes | Yes | Yes |
| 32800693 | Gutmann DAP et al. | 2020 | Liver cancer | 310 | Processing | ICC(1,1) for test reliability and ICC(3,k) for inter-rater agreement | Yes | Yes | Yes |
| 32822054 | Kim D et al. | 2020 | Lung cancer | 35 | Segmentation | Inter-rater agreement (absolute agreement, 2-way random effect model, k = 2), intra-rater agreement (absolute agreement, 2-way mixed effect model, k = 2) | Yes | Yes | Yes |
| 32827069 | Park YW et al. | 2021 | Lung cancer | 51 | Segmentation | One-way random effects model | Yes | No | No |
| 32843663 | Granzier RWY et al. | 2020 | Breast cancer | 138 | Segmentation | Two-way random single measure ICC(2,1) | Yes | Yes | Yes |
| 32939634 | Kulkarni A et al. | 2020 | Pancreatic cancer | 128 | Segmentation | Two-way model with single rating and absolute agreement | Yes | Yes | Yes |
| 32968131 | Crobe A et al. | 2020 | Soft tissue sarcoma | 70 | Segmentation | 2-way random model, agreement between raters and 6 raters | Yes | Yes | Yes |
| 32970859 | Bianchini L et al. | 2021 | Phantom | 1 | Acquisition | Two-way random effects for absolute agreement and single rater/measurement | Yes | Yes | Yes |
| 33118182 | Pati S et al. | 2020 | GBM | 31 | Segmentation | ICC(3,1) | Yes | Yes | Yes |
| 33128598 | Park JE et al. | 2020 | GBM | 422 | Segmentation | Two-way mixed-effects model | Yes | No | No |
| 33137621 | Tsarochi M et al. | 2020 | Breast cancer | 73 | Segmentation | Two-way mixed effect model, single measurement for absolute agreement | Yes | Yes | Yes |
| 33228815 | Wang X et al. | 2020 | Gastric cancer | 539 | Segmentation | Multiple-rating, consistency, 2-way random-effects model | Yes | Yes | No |
| 10.1109/ACCESS.2019.2923755 | Li Z et al. | 2019 | Brain | 15 | Acquisition, processing | ICC(1,1) for intra-observer repeatability, ICC(3,1) for inter-observer repeatability | Yes | Yes | Yes |
| 10.1117/12.2512406 | Hendrik MJ et al. | 2019 | Liver cancer | 13 | Segmentation | ICC(2,1) | Yes | Yes | Yes |
| 10.21037/trc.2017.09.47 | Qingtao Q et al. | 2017 | Liver cancer | 15 | Segmentation | ICC(A,1) (inter-observer segmentation), ICC(C,1) (intra-observer segmentation), formulas given | Yes | Yes | Yes |
| 10.4274/imj.galenos.2019.09582 | Burak K et al. | 2019 | GBM | 70 | Segmentation | Two-way model, single-rating, and absolute agreement | Yes | Yes | Yes |


https://dx.doi.org/10.21037/qims-21-86