Physics and Imaging in Radiation Oncology 20 (2021) 69–75

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Review Article

A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features

Elisabeth Pfaehler [a,*,1], Ivan Zhovannik [b,c,1], Lise Wei [d], Ronald Boellaard [a,e], Andre Dekker [c], René Monshouwer [b], Issam El Naqa [d], Jan Bussink [b], Robert Gillies [f], Leonard Wee [c], Alberto Traverso [c]

[a] Department of Nuclear Medicine and Molecular Imaging, Medical Imaging Center, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[b] Department of Radiation Oncology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands
[c] Department of Radiation Oncology (MAASTRO), GROW School for Oncology, Maastricht University Medical Centre+, Maastricht, The Netherlands
[d] Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
[e] Department of Radiology & Nuclear Medicine, VU University Medical Center, Amsterdam, The Netherlands
[f] Department of Radiology, Moffitt Cancer Center, Tampa, FL, USA

ARTICLE INFO

Keywords:
Radiomics
Repeatability
Reproducibility
Review

ABSTRACT

Background and Purpose: Although quantitative image biomarkers (radiomics) show promising value for cancer diagnosis, prognosis, and treatment assessment, these biomarkers still lack reproducibility. In this systematic review, we aimed to assess the progress in radiomics reproducibility and repeatability in the recent years.
Methods and materials: Four hundred fifty-one abstracts were retrieved according to our search pattern criteria, with the publication dates ranging from 2017/05/01 to 2020/12/01. Each abstract including the keywords was independently screened by four observers. Forty-two full-text articles were selected for further analysis. Patient population data, radiomic feature classes, feature extraction software, image preprocessing, and reproducibility results were extracted from each article. To support the community with a standardized reporting strategy, we propose a specific reporting checklist to evaluate the feasibility to reproduce each study.
Results: Many studies continue to under-report essential reproducibility information: all but one clinical and all but two phantom studies missed to report at least one important item reporting image acquisition. The studies included in this review indicate that all radiomic features are sensitive to image acquisition, reconstruction, tumor segmentation, and interpolation. However, the amount of sensitivity is feature dependent, for instance, textural features were, in general, less robust than statistical features.
Conclusions: Radiomics repeatability, reproducibility, and reporting quality can substantially be improved regarding feature extraction software and settings, image preprocessing and acquisition, cutoff values for stable feature selection. Our proposed radiomics reporting checklist can serve to simplify and improve the reporting and, eventually, guarantee the possibility to fully replicate and validate radiomic studies.

1. Introduction

"Radiomics" refers to the automated extraction of imaging biomarkers from patients' scans and has gained an increasing interest in the last decade. Several radiomics studies have reported promising results for cancer diagnosis, prognosis, or evaluation of treatment response [1–3]. Radiomics studies span common volumetric imaging modalities such as Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI). The numbers of studies investigating applications of radiomics have increased dramatically in the last years. In 2019 alone, 728 studies were indexed in PubMed relating to radiomics studies. However, there remains a translational gap between academic study and clinical utilization [4]. One challenge that makes a clinical implementation of radiomics difficult is the problem of replicating published results. These difficulties are due to the unavailability of input images and software used for computations, poor

* Corresponding author at: Department of Nuclear Medicine and Molecular Imaging, University Medical Center Groningen, Groningen, The Netherlands.
E-mail address: elli.pfaehler@gmail.com (E. Pfaehler).
[1] These authors have equally contributed to this work and are in control over the presented data.

https://doi.org/10.1016/j.phro.2021.10.007
Received 1 April 2021; Received in revised form 28 October 2021; Accepted 29 October 2021
Available online 9 November 2021

reporting of study design, and lack of metadata associated with radiomic studies [5]. External validation of radiomic models has been hampered by models trained on small institutional cohorts, prevalence of overfitting and high false positive discovery rates [6]. Additionally, there is a dissonance between the study design methodology and guidelines from TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) [7] that strongly recommends the validation of prediction models on independent (non-randomly partitioned) datasets and full transparency in reporting.

In a previous review aiming to find a consensus in literature on reproducible and repeatable features, the authors indicated a lack of clear consensus in radiomics methodology. Moreover, the review concluded that deficiencies in reporting of study design, methodology and results were hampering the transparency and reproducibility of radiomic studies [5]. However, the previous effort to identify a subset of features likely to be generally reproducible and repeatable, via a quantitative meta-analysis, was encumbered by two issues identified in the literature: (i) incomplete reporting of radiomic analysis procedure and (ii) heterogeneity in metrics used to report feature repeatability and/or reproducibility. Thus, it was not possible in previous work to perform a conclusive quantitative meta-analysis for every imaging modality (CT, PET, MRI).

Considering the increase in radiomic studies, this work is an updated review about the repeatability and reproducibility of radiomic features, with multiple aims: 1. to identify a set of repeatable/reproducible features for each imaging modality; 2. to verify if consensus has recently emerged regarding a list of major factors impacting on reproducibility/repeatability; 3. to isolate a set of repeatable/reproducible features across different modalities in both human and phantom data; 4. to verify if one of the largest issues identified in previous work, being the poor quality of reporting, has been addressed in new studies. To address aim (4) and support high-quality reporting of radiomic studies, we propose a checklist, which includes all necessary steps to fully reproduce a radiomic study.

2. Methods and Materials

2.1. Eligibility criteria

Peer-reviewed full-text articles in the English language eligible for this review must have been published between 2017/05/01 to 2020/12/01. One electronic database (PubMed) was used to search for records (see supplementary material for query filter). Only studies investigating radiomic features extracted A) from one of the imaging modalities CT, PET, or MRI and B) from radiologic phantoms or from human persons suffering from at least one primary tumor were eligible for review. Included articles had to report on the repeatability/reproducibility of radiomic features at least oneof the following aspects: image acquisition/reconstruction parameters, effect of image pre-processing such as smoothing, or segmentation method. Studies must report a statistical metric assessing the degree of robustness (such as the Interclass-Correlation Coefficient (ICC)).

2.2. Study records

Selection process: After the literature search, titles and abstracts were checked for matching the described criteria by four independent observers. Each reviewer voted if an article was eligible for review. In case of disagreement amongst reviewers, a consensus was obtained by joint discussion. EP and IZ reviewed the phantom studies, while AT and LIW reviewed the patient studies.

Data extraction: We extracted information about the datasets used for radiomics (e.g. primary tumor type, or phantom details), details about the segmentation method used, details about radiomic feature extraction such as the interpolation method, details about the statistical analysis and the summary of results. Details of the reviewer form are given in Table 1a. DICOM attributes about the important imaging details are summarized in Table 1b for the imaging modalities.

Creation of the radiomic reporting checklist: In order to generate a radiomic reporting checklist, we attempted to literally reproduce each step of a study by following the description in each manuscript. The checklist complements the points mentioned by Vallieres et al. [8] and the IBSI reference manual [9,10] and necessary information to reproduce a patient study (i.e. patient inclusion, statistical analysis). This checklist was also inspired by the QUADAS-2, a tool for the quality assessment of diagnostic accuracy studies [9]. The development of "signaling questions" has been carried out by the authors of this paper, who have at least 5 years of expertise in the radiomic domain and are members of tasks forces for radiomic standardization such as the IBSI [10]. Each of the authors defined a list of signaling questions, which were fundamental to reproduce a study. After a round of discussion, the final list of signaling questions was obtained. If a step was not reported and could therefore not be replicated, this step was noted, and we attempted to continue to the next step with a "best guess". This procedure was followed until all steps of the radiomics workflow was completed or guessed.

The risk probability was estimated as the number of "no's divided by the number of "signaling" questions, expressed in percentages.

$$risk\ of\ bias = \frac{|questions\ answered\ with\ no|}{|questions|} \bullet 100\%$$

E.g. if two out of four questions were answered with a "no", the risk of bias is calculated as:

$$risk\ of\ bias = \frac{2}{4} \bullet 100\% = 50\%$$

Supplementary tables 2a-2b show the risk assessment sheets for human and phantom studies, respectively. The only differences between human and phantom studies can be found in section A. This radiomic checklist was proposed in order to tackle the limitations of using only "quality scores" to evaluate such a complex question as methodological quality.

2.3. Outcomes and prioritization

The primary outcome of this review was the degree of repeatability/reproducibility of a radiomic feature. The secondary outcomes were the impact of image acquisition and reconstruction settings, preprocessing steps, and tumor segmentation on the reliability/reproducibility of radiomic features. Additional outcomes were the metrics used for reporting on reliability/reproducibility. Finally, the radiomic reporting checklist was used to evaluate the quality of reporting of analyzed studies.

2.4. Risk of bias in individual studies

Two reviewers independently reviewed the studies. In a discussion round, both observers merged their results. This step was performed in order to avoid that the results were biased based on single reviewer's judgement. Forced consensus was used.

3. Results

3.1. Literature search

A total of 451 abstracts were found while searching PubMed with the aforementioned search filter. After reviewing the abstracts, 42 studies fulfilled the inclusion criteria and seven additional studies had been included as prior knowledge in the field. The PRISMA flowchart illustrating the selection process is shown as Fig. 1.

Of these 42 studies, 29 studies were clinical (human subjects) studies. A summary on the study characteristics are displayed in

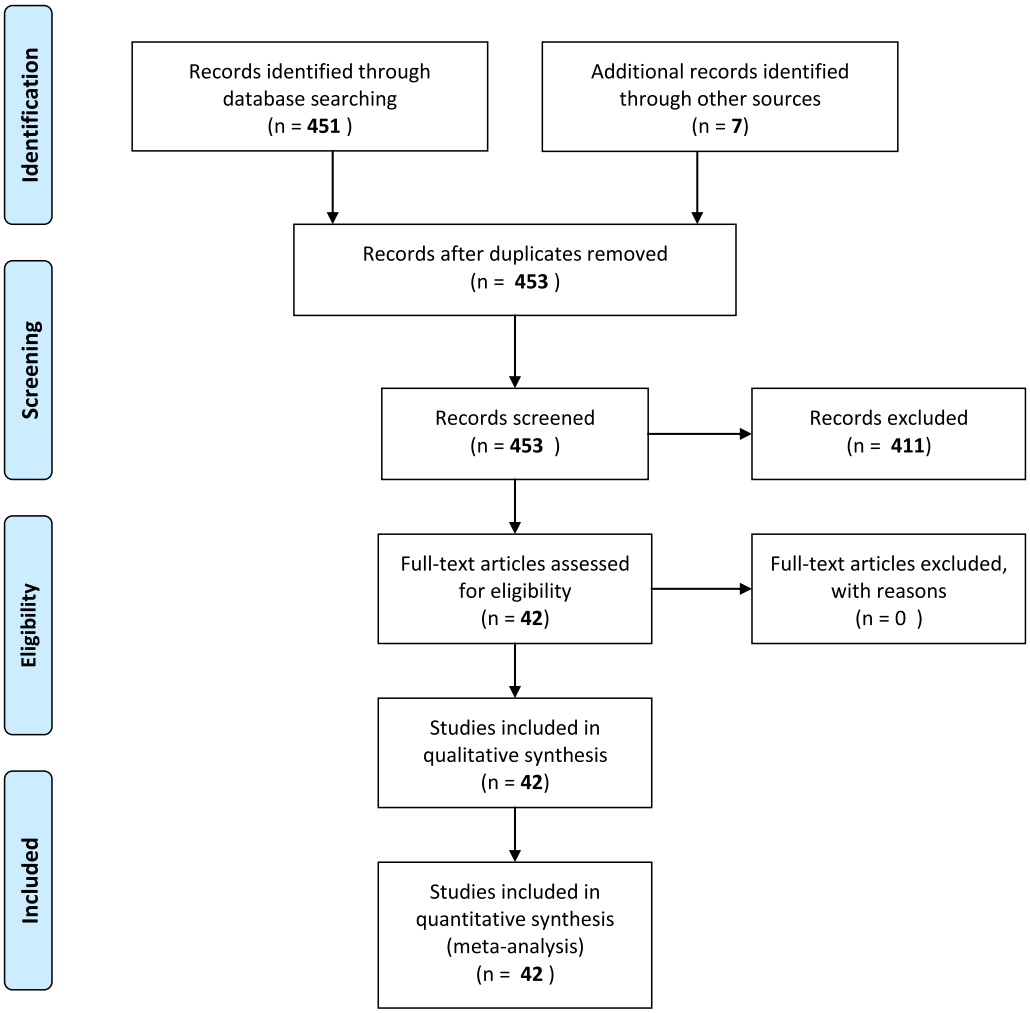

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow chart. The primary PubMed search returned 168 studies. 7 studies were added from primary knowledge. After abstract screening and full-text analysis, a total of 26 studies were included in the qualitative synthesis.

Table 3a. Eight of these studies reported on PET, nine on MR, and twelve on CT images, with only one prospective study [11]. The number of patients included in the studies ranged from a minimum of 14 to a maximum of 465. Five consisted of multi-institutional studies. Most of the studies were focused on reproducibility, with one investigating both reproducibility and repeatability [12].

The other thirteen studies used imaging phantoms to assess repeatability/reproducibility. The general characteristics are summarized in Supplementary table 3b. Two studies reported on MR, six on PET and five on CT images. All studies were retrospective, with one study that provided a publicly available dataset [13]. Six phantom articles reported on feature reproducibility and seven on repeatability.

Five clinical studies [14–17] and two phantom studies [13] made their image dataset publicly available or used a publicly available dataset. A second phantom study used a digital phantom that was already publicly available [18].

Seventeen patient and five phantom studies used publicly available software to calculate radiomic feature values. Pyradiomics was used in eight human studies and one phantom study, and was the most frequently used software [19]. Three patient studies used the open-source software CGITA [20], two patient studies used MaZda [21], one phantom study used LifeX [22], and one phantom and one patient study used IBEX [23]. Furthermore, three studies used open-source code written in Matlab. Six human and three phantom studies failed to mention the programming language used [24].

The most frequently used metric to assess reproducibility was the ICC which was used in 19 human studies and four phantom studies. Five human studies and one phantom study used the Concordance Correlation Coefficient (CCC), however a total of four human studies and two phantom studies used other metrics not described above. Some studies used more than one metric to assess feature stability. In general, the cut-off value to dichotomize stable versus unstable features were highly heterogeneous across studies. One study failed to mention the cut-off value for the ICC, while the threshold for "excellent feature" stability differed from 0.5 to 0.85 in the other studies. One study used the half-width of the ICC confidence interval as threshold for repeatable features. CCC values above 0.8, above 0.7 or equal or higher than 0.85 were considered as robust.

3.2. *Factors impacting radiomic feature values*

The variety of analyzed settings makes it difficult to draw a general conclusion. In summary, all patient studies confirm that differences in image acquisition, reconstruction, preprocessing, and discretization have an impact on radiomic feature values [11,15,17,24,25–30]. While the sensitivity to these factors is feature dependent, this sensitivity is found for all cancer types and imaging modalities. To mitigate this effect, Park et al. demonstrated that CT feature reproducibility can be improved by using a CNN-based super resolution algorithm [31]. For CT studies, Erdal et al. pointed out that a slice thickness of 2 mm leads to the most accurate shape features [32]. It remains unclear if radiomic features extracted from images acquired under different conditions also

lead to different conclusions. In one head and neck PET study, strong dependencies of radiomic features with respect to digital image pre-processing parameters was shown, but these differences were not important enough to affect the prognostic power of radiomic features. Same conclusion holds in the study about nasopharyngeal carcinoma [25].

Many studies also reported on the sensitivity of radiomic features to differences in segmentation [11,14,16,33–43,55]. Also here, the robustness is feature dependent. Some studies demonstrated that the majority of features **was** stable to differences in tumor segmentation for lung [39] and oesophageal cancer. However, in one lung PET study [38], the shape metric sphericity was found to change its prognostic value when using different delineations. **Also** in CT studies, contours delineated by different clinicians impacted the prognostic value of radiomic features. For head and neck cancer, in one PET and one MR study features were found to be very sensitive to differences in tumor delineation [32,33,37,44]. Overall, semi-automated and automated algorithms produced more stable results.

Individual studies showed that the interpolation method, image discretization, voxel size and motion blurring had an effect on radiomic feature values [32,45–48] as listed in detail in the Supplementary material.

Also results from PET phantom studies agreed that image reconstruction protocols (in particular matrix size) had strong impact on feature reproducibility, but with different level of sensitivity per feature categories. In [49], the authors pointed out that smaller volumes seem to result in lower repeatability of feature values. In contrast to these results, Ger et al. reported that most features resulted in a good or excellent reliability when the phantom was scanned on the same scanner with different acquisition protocols [50].

All CT phantom studies focused on differences in acquisition settings such as tube current and agreed that difference in acquisition protocols strongly impact feature reproducibility. However, it is difficult to draw a consensus on which features are stable: The authors in [51] demonstrated that the use of delta radiomic features, i.e. differences of feature values between two different scans of the same patient, increases the repeatability of features.

In one MR phantom study, Baessler et al. [52] investigated both reproducibility and repeatability of radiomic features using a physical phantom scanned using different sequences. The investigators showed that radiomic features extracted from FLAIR (Fluid Attenuated Inversion Recovery) images were more repeatable than features from T1- and T2-weighted images.

3.3. Stable feature categories

As stated above, the variety of tumor types and investigated parameters makes it difficult to identify features which are in general robust. The feature groups found to be stable by the majority of studies were statistical and morphological features, as well as GLCM and GLRLM features, while GLSZM and NGLDM features were found to be less robust. However, a few studies reported the contrary: Yang et al. found in simulated PET lung cancer data that NGLDM features were the most robust, while GLCM features were the least robust feature group [53]. One CT phantom study showed that statistical features are less robust than textural features [54]. Baessler et al. also showed in their MR phantom study that GLSZM features were more robust than GLCM features [52]. One study showed that features extracted from Fourier transformed images are the most robust, a feature group none of the other papers investigated. The authors in [51] demonstrated that the use of delta radiomic features, i.e. differences of feature values between two different scans of the same patient, increases the repeatability of features.

3.4. Radiomic reporting checklist

Supplementary tables 4a-5b summarizes the radiomic reporting checklist risk assessment for clinical and phantom studies. None of the studies scored a zero-risk bias probability. An example of the check-list can be found in the Supplementary material.

For clinical studies, the lowest median risk was achieved in the "study design" domain. Information about the creation of the binary mask (question B4-imaging domain) was the least reported item with only one study providing detailed information. Twelve patient studies missed to report on feature specific parameters (questions C5-C6-radiomic pipeline domain) such as feature aggregation and. In contrast, only three patient studies missed to provide any detail of the software which is a clear improvement when compared with the previous review.

For phantom studies, there were no risks of biases in the "study design" and "statistical analysis" domains. Compared to clinical studies a) all but one phantom study included a table reporting all statistical results (signaling question E3- data and metadata availability domain), and b) all but two studies failed to report on how the binary mask was created from the segmentation (question B4-imaging domain).

4. Discussion

The studies included in this review confirm almost all findings of the previous published review. Major issues remain the little number of studies that made their data publicly available, the heterogeneity of used metrics and cut-off values for the assessment of feature robustness, and the lack of detailed reporting. To ease the way to reproduce a study, we invite again the radiomics community to make their data and metadata publicly available. The variability in metrics and cut-offs used to categorize the features into good/poor reproducibility/repeatability, makes it difficult to compare the results of the studies. While there is no evidence that a specific metric should be used for analysis, a description about the metrics, as well as statistical hypothesis underlying the data analysis and specific cut-offs applied should be reported to guarantee the transparency and the reproducibility of the study. We also strongly recommend the users to append as supplementary material the raw results of the analysis to facilitate meta-analyses. We strongly advise to follow guidelines provided by the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement for a transparent reporting of model design, development and evaluation.

Even though it was hard to get a consensus on which features are repeatable/reproducible, all studies agreed that reconstruction settings, image noise, and segmentation method have a high impact on radiomic feature values for all imaging modalities. This implies that multi-center radiomic studies require harmonized images in terms of image reconstruction setting and signal to noise ratio. Using images from different centers without harmonizing the images itself can lead to wrong conclusions. Further harmonization can be achieved with correction methods applied before feature extraction such as e.g. resampling the images to cubic voxels or by standardizing images via post-processing such as by histogram equalization of MR images or post-reconstruction smoothing [56,57]. Moreover, radiomics harmonization can be achieved by image domain adaptation to reduce the influence of image acquisition settings as Chen et al showed in a simulation study [58]. However, it still has to be validated if these methods can be used for radiomic analysis as it might be that by standardizing the images important textural information gets lost.

Additionally to the standardization of images, there are methods aligning radiomic features. One of the most popular methods is the so-called ComBat, which has been applied to CT [59] and PET features

[57], but still requires a large-scale validation. However, even though, these algorithms to correct for multi-center effects are being developed, it is still important to keep this correction as small as possible. Therefore, when using multi-centric data, it is essential that the images are as comparable as possible in terms of image acquisition. Moreover, patient cohorts across institutions/scanners should be comparable, i.e. the number of patients with a positive/negative outcome should be comparable across the datasets of each institution/scanner type. Otherwise the findings using a radiomic model can be due to inter-scanner differences of radiomic features and not to differences caused by variability in tumor characteristics.

To ensure a valid and reproducible analysis of PET studies, it should be carefully checked if reported tracer dose and uptake time are correct and the conversion from image data in Bq/ml to SUV units is accurate. If the liver is displayed in the image, this can be checked by drawing a 3 cm^2 in the liver and verifying that the mean SUV inside the 'liver' sphere is in the range between 1.5 and 2.5. Higher/lower values are an indication for calibration or other errors and these images should be verified, corrected or excluded from the analysis. If the liver is not displayed in the image such as for brain images, a digital reference object can be used to verify the correct conversion to SUV as proposed by Pierce et al. [60]

To minimize the effect of different segmentations, a (semi-) automatic segmentation method might be preferred, as automatic approaches reduce inter-observer variability and yield a higher reproducibility than manual segmentations [61]. The most suitable segmentation method for radiomic analysis has to be identified what has to be done for each imaging modality and cancer type separately. Likely, several segmentation methods will be a good candidate as they yield similar accuracy and repeatability.

Regarding the quality of reporting, we invite the users to provide not only their software but also the metadata associated with it such as information about the programming language. In general, we recommend that each software used for feature calculation should be tested if it complies with the benchmarks provided by IBSI. In this way, feature values extracted by different software packages become comparable which is one important step in the standardization of radiomic feature values.

However, also many studies included in this review missed to report details of preprocessing steps. This missing information has the consequence that the study itself becomes non reproducible and the results are not comparable with other studies. In summary, to make radiomic studies comparable across centers, pre-processing steps should be standardized for each imaging modality as suggested by Park et al. [31]. This includes the discretization method as well as the bin number/bin width of choice as it has an impact on radiomic feature values [62,63]. Since radiomic features can be sensitive to differences in voxel size, it is recommended to interpolate the images before feature extraction to an isotropic voxel size. This step and the used interpolation method should be reported if applied and the radiomic community should agree on which kind of resampling is the preferred one. Almost all studies did not report any information related to the generation of the binary mask from the original data. However, different software tools are available to go from a contour to the final binary mask, therefore it is important to state it.

Most studies reported on the robustness of first order and local textural features such as GLCM and GLRLM features, while global textural features (such as GLSZM features) were found to be less robust.

One limitation of the current review is that it was not possible to draw a general conclusion on which features are reproducible and can be used in the clinic what was one of the original aims of this review. The variety of analyzed settings and used metrics makes it impossible to perform a quantitative synthesis of the analyzed articles. Unfortunately, since the last review, the radiomics community did not come to a consensus on which metric is the most adequate to use in a radiomics setting. The checklist proposed in this review will hopefully help to increase the reproducibility of radiomic studies. Moreover, we hope that future radiomic studies will consider only repeatable and reproducible features and will focus on the transferability of the results.

Funding

This work is part of the research program STRaTeGy with project number 14929, which is (partly) financed by the Netherlands Organization for Scientific Research (NWO).

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andre Dekker reports grants from Varian Medical Systems, personal fees from Medical Data Works BV, personal fees from UHN Toronto, personal fees from Hanarth Fund, personal fees from Johnson & Johnson, outside the submitted work; In addition, Andre Dekker has a patent Systems, methods and devices for analyzing quantitative information obtained from radiological images US Patent 9721340 B2 issued.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2021.10.007.

References

[1] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1). https://doi.org/10.1038/ncomms5006.

[2] Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017;6. doi: 10.7554/eLife.23421.

[3] Tixier F, Hatt M, Valla C, Fleury V, Lamour C, Ezzouhri S, et al. Visual versus quantitative assessment of intratumor 18F-FDG PET uptake heterogeneity: prognostic value in non-small cell lung cancer. J Nucl Med 2014;55(8):1235–41. https://doi.org/10.2967/jnumed.113.133389.

[4] Buvat I, Orlhac F. The dark side of radiomics: on the paramount importance of publishing negative results. J Nucl Med 2019;60(11):1543–4. https://doi.org/10.2967/jnumed.119.235325.

[5] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol 2018;102(4):1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.

[6] Chalkidou A, O'Doherty MJ, Marsden PK, Rubin DL. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS ONE 2015;10(5):e0124165. https://doi.org/10.1371/journal.pone.0124165.

[7] Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13(1):1. https://doi.org/10.1186/s12916-014-0241-z.

[8] Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. J Nucl Med 2018;59(2):189–93. https://doi.org/10.2967/jnumed.117.200501.

[9] Whiting PF. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

[10] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295(2):328–38. https://doi.org/10.1148/radiol.2020191145.

[11] Zhuang M, García DV, Kramer GM, Frings V, Smit EF, Dierckx R, et al. Variability and repeatability of quantitative uptake metrics in 18F-FDG PET/CT of non–small cell lung cancer: Impact of segmentation method, uptake interval, and reconstruction protocol. J Nucl Med 2019;60(5):600–7. https://doi.org/10.2967/jnumed.118.216028.

[12] Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. Radiother Oncol 2019;135:107–14. https://doi.org/10.1016/j.radonc.2019.03.001.
[13] Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, et al. Learning from scanners: bias reduction and feature correction in radiomics. Clin Transl Radiat Oncol 2019;19:33–8. https://doi.org/10.1016/j.ctro.2019.07.003.
[14] Xia W, Chen Y, Zhang R, Yan Z, Zhou X, Zhang Bo, et al. Radiogenomics of hepatocellular carcinoma: multiregion analysis-based identification of prognostic imaging biomarkers by integrating gene data—a preliminary study. Phys Med Biol 2018;63(3):035044. https://doi.org/10.1088/1361-6560/aaa609.
[15] Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 2019;9 (1). https://doi.org/10.1038/s41598-019-45766-z.
[16] Haarburger C, Müller-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. https://doi.org/10.1038/s41598-020-69534-6.
[17] 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(1):179–90. https://doi.org/10.1002/acm2.v21.110.1002/acm2.12795.
[18] Midya A, Chakraborty J, Gönen M, Do RKG, Simpson AL. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. J Med Imaging 2018;5(01):1. https://doi.org/10.1117/1.JMI.5.1.011020.
[19] Griethuysen JJM Van, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype 2017;77: 104–8. doi: 10.1158/0008-5472.CAN-17-0339.
[20] Fang Y-H, Lin C-Y, Shih M-J, Wang H-M, Ho T-Y, Liao C-T, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. Biomed Res Int 2014;2014:1–9. https://doi.org/10.1155/2014/248505.
[21] Strzelecki M, Szczypinski P, Materka A, Klepaczko A. A software tool for automatic classification and segmentation of 2D/3D medical images. Nucl Instrum Methods Phys Res Sect A Accel Spectrometers, Detect Assoc Equip 2013;702:137–40. https://doi.org/10.1016/j.nima.2012.09.006.
[22] Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, et al. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res 2018;78(16): 4786–9. https://doi.org/10.1158/0008-5472.CAN-18-0125.
[23] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42:1341–53. https://doi.org/10.1118/1.4908210.
[24] Altazi BA, Zhang GG, Fernandez DC, Montejo ME, Hunt D, Werner J, et al. 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(6):32–48. https://doi.org/10.1002/acm2.2017.18.issue-610.1002/acm2.12170.
[25] Lv W, Yuan Q, Wang Q, Ma J, Jiang J, Yang W, et al. Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT. Eur Radiol 2018;28(8):3245–54. https://doi.org/10.1007/s00330-018-5343-0.
[26] Traverso A, Kazmierski M, Shi Z, Kalendralis P, Welch M, Nissen HD, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. Phys Medica 2019;61: 44–51. https://doi.org/10.1016/j.ejmp.2019.04.009.
[27] Carles M, Bach T, Torres-Espallardo I, Baltas D, Nestle U, Martí-Bonmatí L. Significance of the impact of motion compensation on the variability of PET image features. Phys Med Biol 2018;63(6):065013. https://doi.org/10.1088/1361-6560/aab180.
[28] Loi S, Mori M, Benedetti G, Partelli S, Broggi S, Cattaneo GM, et al. Robustness of CT radiomic features against image discretization and interpolation in characterizing pancreatic neuroendocrine neoplasms. Phys Medica 2020;76: 125–33. https://doi.org/10.1016/j.ejmp.2020.06.025.
[29] Meyer M, Ronald J, Vernuccio F, Nelson RC, Ramirez-Giraldo JC, Solomon J, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology 2019;293(3):583–91. https://doi.org/10.1148/radiol.2019190928.
[30] Yamashita R, Perrin T, Chakraborty J, Chou JF, Horvat N, Koszalka MA, et al. Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation. Eur Radiol 2020;30(1):195–205. https://doi.org/10.1007/s00330-019-06381-8.
[31] Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. Korean J Radiol 2019;20(7):1124. https://doi.org/10.3348/kjr.2018.0070.
[32] Erdal BS, Demirer M, Little KJ, Amadi CC, Ibrahim GFM, O'Donnell TP, et al. Are quantitative features of lung nodules reproducible at different CT acquisition and reconstruction parameters? PLoS ONE 2020;15(10):e0240184. https://doi.org/10.1371/journal.pone.0240184.
[33] Kocak B, Yardimci AH, Bektas CT, Turkcanoglu MH, Erdim C, Yucetas U, et al. Textural differences between renal cell carcinoma subtypes: Machine learning-based quantitative computed tomography texture analysis with independent external validation. Eur J Radiol 2018;107:149–57. https://doi.org/10.1016/j.ejrad.2018.08.014.
[34] Qu J, Shen C, Qin J, Wang Z, Liu Z, Guo J, et al. The MR radiomic signature can predict preoperative lymph node metastasis in patients with esophageal cancer. Eur Radiol 2019;29(2):906–14. https://doi.org/10.1007/s00330-018-5583-z.
[35] 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(8):3582–91. https://doi.org/10.1002/mp.v46.810.1002/mp.13624.
[36] Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, et al. A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. Eur Radiol 2018;28(6): 2255–63. https://doi.org/10.1007/s00330-017-5154-8.
[37] Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, et al. 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 Medica 2018;49:105–11. https://doi.org/10.1016/j.ejmp.2018.05.013.
[38] Hatt M, Laurent B, Fayad H, Jaouen V, Visvikis D, Le Rest CC. Tumour functional sphericity from PET images: prognostic value in NSCLC and impact of delineation method. Eur J Nucl Med Mol Imaging 2018;45(4):630–41. https://doi.org/10.1007/s00259-017-3865-3.
[39] Guan Y, Li W, Jiang Z, Chen Y, Liu S, He J, et al. Whole-lesion apparent diffusion coefficient-based entropy-related parameters for characterizing cervical cancers. Acad Radiol 2016;23(12):1559–67. https://doi.org/10.1016/j.acra.2016.08.010.
[40] Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O, et al. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. Radiol Phys Technol 2018;11(1):27–35. https://doi.org/10.1007/s12194-017-0433-2.
[41] Takeda K, Takanami K, Shirata Y, Yamamoto T, Takahashi N, Ito K, et al. Clinical utility of texture analysis of 18F-FDG PET/CT in patients with Stage I lung cancer treated with stereotactic body radiotherapy. J Radiat Res 2017;58:862–9. doi: 10.1093/jrr/rrx050.
[42] Bektas CT, Kocak B, Yardimci AH, Turkcanoglu MH, Yucetas U, Koca SB, et al. Clear cell renal cell carcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of fuhrman nuclear grade. Eur Radiol 2019;29(3):1153–63. https://doi.org/10.1007/s00330-018-5698-2.
[43] Feng Z, Rong P, Cao P, Zhou Q, Zhu W, Yan Z, et al. Machine learning-based quantitative texture analysis of CT images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma. Eur Radiol 2018;28 (4):1625–33. https://doi.org/10.1007/s00330-017-5118-z.
[44] Zhang X, Zhong L, Zhang B, Zhang Lu, Du H, Lu L, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: evaluation with two disease groups. Cancer Imaging 2019;19(1). https://doi.org/10.1186/s40644-019-0276-7.
[45] Park S, Lee SM, Do K-H, Lee J-G, Bae W, Park H, et al. Deep learning algorithm for reducing CT slice thickness: effect on reproducibility of radiomic features in lung cancer. Korean J Radiol 2019;20(10):1431. https://doi.org/10.3348/kjr.2019.0212.
[46] Whybra P, Parkinson C, Foley K, Staffurth J, Spezi E. Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. Sci Rep 2019;9:9649. https://doi.org/10.1038/s41598-019-46030-0.
[47] Lee S-H, Cho H, 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. https://doi.org/10.1186/s40644-019-0239-z.
[48] Lafata K, Cai J, Wang C, Hong J, Kelsey CR, Yin F-F. Spatial-temporal variability of radiomic features and its effect on the classification of lung cancer histology. Phys Med Biol 2018;63(22):225003. https://doi.org/10.1088/1361-6560/aae56a.
[49] Pfaehler E, Beukinga RJ, de Jong JR, Slart RHJA, Slump CH, Dierckx RAJO, et al. Repeatability of 18 F-FDG PET radiomic features: a phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. Med Phys 2018. https://doi.org/10.1002/mp.13322.
[50] Ger RB, Meier JG, Pahlka RB, Gay S, Mumme R, Fuller CD, et al. Effects of alterations in positron emission tomography imaging parameters on radiomics features. PLoS ONE 2019;14(9):e0221877. https://doi.org/10.1371/journal.pone.0221877.
[51] Nardone V, Reginelli A, Guida C, Belfiore MP, Biondi M, Mormile M, et al. Delta-radiomics increases multicentre reproducibility: a phantom study. Med Oncol 2020;37(5). https://doi.org/10.1007/s12032-020-01359-9.
[52] Baeßler B, Weiss K, Pinto dos Santos D. Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. Invest Radiol 2019;54(4):221–8. https://doi.org/10.1097/RLI.0000000000000530.
[53] 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. https://doi.org/10.1038/s41598-019-57171-7.
[54] Varghese BA, Hwang D, Cen SY, Levy J, Liu D, Lau C, et al. Reliability of CT-based texture features: Phantom study. J Appl Clin Med Phys 2019;20(8):155–63. https://doi.org/10.1002/acm2.v20.810.1002/acm2.12666.
[55] Johnson PB, Young LA, Lamichhane N, Patel V, Chinea FM, Yang F. Quantitative imaging: correlating image features with the segmentation accuracy of PET based tumor contours in the lung. Radiother Oncol 2017;123(2):257–62. https://doi.org/10.1016/j.radonc.2017.03.008.
[56] Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS ONE 2017;12 (9):e0178524. https://doi.org/10.1371/journal.pone.0178524.
[57] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in

PET. J Nucl Med 2018;59(8):1321–8. https://doi.org/10.2967/jnumed.117.199935.
[58] Chen J, Zhang C, Traverso A, Zhovannik I, Dekker A, Wee L, et al. Generative models improve radiomics reproducibility in low dose CTs: a simulation study. Phys Med Biol 2021;66(16):165002. https://doi.org/10.1088/1361-6560/ac16c0.
[59] Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 2019;291(1):53–9. https://doi.org/10.1148/radiol.2019182023.
[60] Pierce LA, Elston BF, Clunie DA, Nelson D, Kinahan PE. A digital reference object to analyze calculation accuracy of PET standardized uptake value. Radiology 2015;277(2):538–45. https://doi.org/10.1148/radiol.2015141262.
[61] Kolinger GD, Vállez García D, Kramer GM, Frings V, Smit EF, de Langen AJ, et al. Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. EJNMMI Res 2019;9(1). https://doi.org/10.1186/s13550-019-0481-1.
[62] Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik J-C, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS ONE 2019;14(3):e0213459. https://doi.org/10.1371/journal.pone.0213459.
[63] Leijenaar RTH, Nalbantov G, Carvalho S, van Elmpt WJC, Troost EGC, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep 2015;5(1). https://doi.org/10.1038/srep11075.