Abstract

Introduction. Radiomics is a promising imaging-based tool which could enhance clinical observation and identify representative features. To avoid different interpretations, the Image Biomarker Standardisation Initiative (IBSI) imposed conditions for harmonisation. This study evaluates IBSI-compliant radiomics applications against a known benchmark and clinical datasets for agreements.

Materials and methods. The three radiomics platforms compared were RadiomiX Research Toolbox, LIFEx v7.0.0, and syngo.via Frontier Radiomics v1.2.5 (based on PyRadiomics v2.1). Basic assessment included comparing feature names and their formulas. The IBSI digital phantom was used for evaluation against reference values. For agreement evaluation (including same software but different versions), two clinical datasets were used: 27 contrast-enhanced computed tomography (CECT) of colorectal liver metastases and 39 magnetic resonance imaging (MRI) of breast cancer, including intravoxel incoherent motion (IVIM) and dynamic contrast-enhanced (DCE) MRI. The intraclass correlation coefficient (ICC, lower 95% confidence interval) was used, with 0.9 as the threshold for excellent agreement. Results. The three radiomics applications share 41 (3 shape, 8 intensity, 30 texture) out of 172, 84 and 110 features for RadiomiX, LIFEx and syngo.via, respectively, as well as wavelet filtering. The naming convention is, however, different between them. Syngo.via had excellent agreement with the IBSI benchmark, while LIFEx and RadiomiX showed slightly worse agreement. Excellent reproducibility was achieved for shape features only, while intensity and texture features varied considerably with the imaging type. For intensity, excellent agreement ranged from 46% for the DCE maps to 100% for CECT, while this lowered to 44% and 73% for texture features, respectively. Wavelet features produced the greatest variation between applications, with an excellent agreement for only 3% to 11% features. Conclusion. Even with IBSI-compliance, the reproducibility of features between radiomics applications is not guaranteed. To evaluate variation, quality assurance of radiomics applications should be performed and repeated when updating to a new version or adding a new modality.

1. Introduction

Medical images have primarily been used as diagnostic tools interpreted qualitatively by trained observers, while in the last years - especially in oncology - more quantitative evaluation is performed primarily to tailor personalised healthcare further. In this context, conventional metrics from functional imaging were first explored, such as the maximum standardised uptake value (SUVmax) in 18F-FDG Positron Emission Tomography (PET) and $K^{\text{trans}}$ in perfusion Magnetic Resonance Imaging (MRI), which have been...
The purpose of this study was to investigate the reproducibility of radiomics features between radiomics applications that participated in the IBSI standardisation. Three IBSI-compliant applications were compared in an in-depth evaluation, using a digital phantom and two clinical datasets of various anatomical regions and imaging modalities.

2. Materials and methods

2.1. Study design

The comparison of the three radiomics applications was divided into four phases. In the first phase, we checked the correspondence of the feature names between the IBSI standard and the applications based on their equations. In the second phase, features from the digital phantom were computed on each platform and compared with the reference values provided by IBSI. As applications undergo frequent updates, we tested different versions of the same application for their consistency in the values of the features in the third phase. Lastly, in the fourth phase, the three applications were compared based on their shared features to assess the reproducibility of those features. Clinical data was used in both the third and fourth phases.

2.2. Imaging data sets and segmentation

This study used three different image data sets: a digital phantom and two clinical data sets of two different pathologies, comprising 27 computed tomography (CT) and 39 magnetic resonance imaging (MRI) scans.

IBSI provides a digital phantom with one predefined ROI specifically designed to benchmark feature implementation (Zwanenburg et al. 2016). They benchmarked 169 features values (Zwanenburg et al. 2020). The 3-dimensional image consists of 80 voxels of $2 \times 2 \times 2$ mm$^3$ with intensity ranging from 1 to 9 and is associated with a mask of 74 voxels (figure 1). Both image and mask, provided in NIFTI format, were converted to DICOM format (RT-STRUCT object for the mask) using 3D Slicer v4.11 (Fedorov et al. 2012) (www.slicer.org).

The clinical data sets originated from two retrospective studies in Jules Bordet Institute, Brussels, Belgium. These studies were approved by the Jules Bordet Institute Ethics Committee (CE 2953 and CE 3117).

The first clinical data set included 27 patients (13 women and 14 men; mean age of 61 years) with histologically confirmed colorectal liver metastases who underwent a contrast-enhanced (CE) CT of the abdomen acquired at the portal phase before their curative-intent surgery. In this data set, a single radiologist observer segmented 36 metastases in a semi-automatic approach using syngo.via Frontier Radiomics v1.2.5 (Siemens Healthcare, Erlangen, Germany). The procedure consisted of adjusting a volume box containing the whole tumour, refining the automated threshold...
and manually correcting the contours. The median tumour volume was 3.9 ml (range: 0.2–340.4 ml). An example is shown in figure 2(A).

The second clinical data set was composed of 39 patients (39 women, mean age of 52 years). All patients had pathologically confirmed untreated invasive ductal or lobular breast cancer with an indication for neoadjuvant chemotherapy and surgical resection and had undergone a multiparametric MRI prior to treatment. The MRI acquisition protocol included diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) sequences, where T1 relaxation times...
were determined using a variable flip angle gradient echo acquisition. Two maps (D, f) were computed using an in-house developed software with the intravoxel incoherent motion (IVIM) model from the DWI series. Three DCE-MRI maps (Ktrans, Kep, IAUG) were extracted from the dynamic image series with TumourMetrics (Chao et al. 2017). An MRI specialist semi-automatically segmented the DCE subtraction series in the same approach as described above. Another radiologist observer validated each contour. For one patient, the segmentation of the tumour was composed of two non-associated clusters. The IVIM maps were rigidly registered to the DCE maps. The median tumour volume was 7.7 ml (range: 0.4–127.0 ml). An example of $K_{\text{trans}}$ map and its segmentation is shown in figure 2(B).

Additional details about image acquisition and reconstruction for the two clinical data sets are described in supplementary data (available online at stacks.iop.org/BPEX/8/065008/mmedia) tables 1 and 2. The clinical data structures were stored in syngo.via RTSTRUCT object. They were extraced from the dynamic image series with TumourMetrics. The three commercially available radiomics applications considered in the study were RadiomiX Research Toolbox (Radiomics, Liège, Belgium), LIFEx v7.0.0 (Nioche et al. 2018) (www.lifexsoft.org), and syngo.via Frontier Radiomics v1.2.5 (Siemens Healthcare, Erlangen, Germany), which is based on PyRadiomics v2.1 (van Griethuysen et al. 2017) (pyradiomics.readthedocs.io/en/latest), Numpy v1.15.3, SimpleITK v1.0, PyWavelet v1.0.1 and Python v3.5.6. For the software version comparison, LIFEx v7.0 was compared to v6.65 and PyRadiomics v2.1 with the same python version dependencies as syngo.via was compared to PyRadiomics v3.0.1 (with Numpy v1.19.2, SimpleTK v2.0.2, PyWavelet v1.1.1 and Python v3.7.7).

All three applications, or their original code, participated in the IBSI standardisation where their respective features were compared with reference values of the digital phantom and clinical CT images, and validated by achieving a good reproducibility for multimodal images.

### Table 1.
Summary of the main characteristics of the three radiomics applications considered in the study. Abbreviations: ROI = Region of Interest; LoG = Laplacian of Gaussian; FBS = Fixed Bin Size; FBN = Fixed Bin Number.

| Programming language | RadiomiX | LIFEx | Syngo.via |
|----------------------|----------|-------|----------|
| Image Format         | MATLAB   | Java  | Python   |
| ROI definition       | RT-STRUCT| RT-STRUCT, NIfTI, JFF, PNG, BMP | DICOM SEG, NIfTI and segmentation tools |
| Image and segmentation display | No | Yes | Yes |
| Model building       | Yes | No | Yes |
| Image processing:  
  • Filtering          | LoG, Wavelet | LoG, Wavelet, Mean, Laws | LoG, Wavelet, Square, SquareRoot, Logarithm, Exponential |
  • Normalisation      | Z-score normalisation | No | Z-score normalisation |
  • Resampling interpolator | linear, nearest or cubic | polynomial of degree 5 (Lagrangian polynomial) | linear, B spline, nearest neighbour, gaussian, label gaussian, hammering/welch/ blackman windowed sinc |
  • Discretisation     | FBS and FBN | FBS (absolute or relative) and FBN | FBS |
| Features number in each category | 26 shape, 47 intensity, 96 texture and 3 fractal | 5 shape, 47 intensity and 32 texture | 17 shape, 18 intensity and 75 texture |

### Table 2.
Application-specific calculation settings and their selected values for this study. Parameters that the user cannot change are marked with an asterisk (*). Abbreviations: GLCM = Grey Level Co-occurrence Matrix; MNC = Matrix Not Computed; NGLDM = Neighbouring Grey Level Dependence Matrix; NM = Not Mentioned in the documentation provided by the application.

|                           | RadiomiX | LIFEx | Syngo.via |
|---------------------------|----------|-------|----------|
| GLCM distance             | 1        | 1     | 1        |
| GLCM symmetry             | NM       | Yes*  | Yes      |
| Texture matrix weighting  | NM       | None* | None     |
| Voxel array shift (intensity) | 0°     | 0°     | 0        |
| NGLDM cutoff value        | NM       | MNC   | 0        |
| Wavelet filter:  
  • Family                  | NM       | Coiflets | Coiflets |
  • Start level             | NM       | NM    | 0        |
  • Order                   | NM       | 1     | 1        |
  • Level                   | NM       | 1     | 1        |
  • Padding Method          | NM       | Reflect | NM      |
(Zwanenburg et al. 2020). They were therefore considered IBSI compliant in this study. The main characteristics of the three applications are summarised in Table 1. PyRadiomics is the only open-source project, allowing a thorough investigation of the implementation. The other software packages are closed-source projects, where only documentation with written formulas is available.

2.4. Feature extraction

Features were extracted directly from the same DICOM images over the different applications, except for PyRadiomics where conversion to SimpleITK images with SimpleITK v2.0.2 package in Python is performed before sending the images to PyRadiomics. Features were extracted from the same segmented structures but with the file format required by the different tools (i.e. NIfTI for LIFEx, syngo.via and PyRadiomics; RTSTRUCT for RadiomiX).

Except for image discretisation, no image processing in the radiomics applications was performed (neither spatial resampling nor normalisation). Two approaches to discretisation can be used: a fixed number of bins (FBN) which introduces a normalising effect, recommended for arbitrary intensity units (e.g. raw MRI), and a fixed bin size (FBS) where the relationship between image intensity and physiological meaning is kept. As syngo.via has only the FBS option, this approach was used. A bin size of 1 was applied for the digital phantom. For the clinical datasets, the bin size was selected to produce a bin count between 30 and 128 for the majority of the tumours, as this range is frequently used in radiomics studies (Tixier et al. 2012, Schwier et al. 2019, Dreher et al. 2020, Traverso et al. 2020a). Based on tumour intensity ranges, this resulted in a bin size of 3 for CECT, 18 for IVIM-D, 1 for IVIM-f, 12 for IAUG, 28 for Ktrans, and 32 for Kep.

Calculation settings for each application were left to their default value. In that way, they were matched as closely as possible according to the information provided in the documentation. Table 2 shows the settings available in each application and the values chosen for this study.

Investigated features were categorised by shape, intensity, texture and filter-based. The shape category includes all the morphological features. The intensity category combines local intensity features, intensity-based statistical features, and intensity histogram features referred to as discretised intensity features in this study. The texture category includes grey level co-occurrence matrix (GLCM), grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM), grey level distance zone matrix (GLDZM), neighbourhood grey tone difference matrix (NGTDM), and neighbouring grey level dependence matrix (NGLDM). The filter used in this work is the wavelet filter. Since IBSI has not yet standardised the filter-based category, it was not analysed in phase 2. The feature names used in this study correspond to the IBSI nomenclature.

2.5. Statistical analysis

After matching feature names based on the provided formulas, we calculated, in phase 2 (comparison to IBSI reference values), the absolute relative difference (RD) between each application and the IBSI benchmark values for shared features. RD is the absolute value of the percentage difference between the application value and the reference value divided by the reference value. The features were classified to indicate an excellent (RD ≤ 1%), good (1% < RD ≤ 5%), moderate (5% < RD ≤ 10%) or poor (RD > 10%) match to IBSI reference.

Agreement on common radiomics features across software versions (phase 3) and across the different applications (phase 4) was analysed using the two-way mixed effect absolute agreement single rater intraclass correlation coefficient (ICC) and its 95% confidence interval. The ICC quantifies the absolute agreement between data sets by comparing the variability in feature values across software packages to the variability in values across patients. Using the lower boundary of the 95% confidence interval (ICC LCI), feature reliability was stratified to poor (values less than 0.5), moderate (values between 0.5 and 0.75), good (values between 0.75 and 0.9) and excellent (values greater than 0.9) (Koo and Li 2016). ICC was computed for each feature using pingouin library v0.3.11 (Vallat 2018) (pingouin-stats.org), an open-source statistical package written in Python.

3. Results

3.1. Phase 1: matching terminology and formulas

The naming conventions of matrices and features were not entirely consistent with IBSI nomenclature and between the three applications. For example, in the case of matrix names, NGLDM is tagged as GLDM in syngo.via, and LIFEx NGLDM corresponds, in fact, to NGTDM. Another example of different naming conventions for features calculated with the same equation, the GLSZM Gr ayLevelNonUniformity in syngo.via is referred to as GLNU in LIFEx and IN in RadiomiX.

Supplementary Material contains a complete list of feature names from each application and their correspondence to each other and to the IBSI nomenclature.

The percentage of shared features of RadiomiX, LIFEx and syngo.via compared to the total number of reference values in IBSI are 92%, 29% and 62%, respectively. Forty-one common features were identified between the three applications, consisting of 3 shape features, 8 intensity features, and 30 texture features (see Table 3 and Supplementary Material for more details). In those common features, Kurtosis is not a feature standardised by IBSI, which standardised Excess kurtosis instead (by subtracting the kurtosis by 3 to centre on zero for normal distributions). For the filter-based features, the wavelet filter is available in all three applications except in LIFEx v6.65 where it was
not yet implemented. The wavelet filter adds 304 features to the phase 4 analysis (intensity and texture features for the 8 different decompositions of the original image).

Differences in the implementation of some features between the three applications were recorded based on their documentation. Firstly, shape features with length units are calculated in centimetre in LIFEx and Radiomix, contrary to millimetre in IBSI and syngo.via. Consequently, these shape features were converted to millimetre as a post-processing step before comparison. Secondly, the applications do not handle segmentation containing more than one cluster the same way: LIFEx takes into account only the largest cluster within the ROI, syngo.via and RadiomiX treat the multiple clusters as one volume to compute every shape feature for syngo.via and only 6 shape features from 26 implemented for RadiomiX. Therefore, the only patient with two clusters after segmentation was discarded from the analysis. The IBSI standard defines the volume of a ROI in two ways, a voxel-based approach and a mesh-based approach. The latter is used to compute shape features with volume in their formula. Syngo.via computes both options for volume calculation and uses mesh-based for subsequent shape features, as per IBSI definition. RadiomiX and LIFEx only implement the voxel-based volume, which can impact the subsequent features (e.g. SurfaceArea) compared to the IBSI reference or syngo.via. Finally, the definition for the intensity_MedianAbsoluteDeviation feature did not match between IBSI and RadiomiX. This feature was therefore discarded in the remainder of the analysis.

3.2. Phase 2: comparison to IBSI benchmark values with phantom data

Syngo.via provided feature values with excellent agreement (RD < 1%) with the IBSI reference values across all 104 features.

3.3. Phase 3: application version comparison with clinical data

In LIFEx v6.65, the wavelet filter is not implemented; the filter-based features were thus not compared. LIFEx had excellent reliability for all features between v7.0 and v6.65, except for discretised_Mean, which had moderate reliability (0.75 > ICC LCI > = 0.5) for IVIMD and Kep images and good reliability (0.9 > ICC LCI > = 0.75) for CECT and IAUG images. The discretised_Mean in LIFEx v6.65 was systematically one unit lower than in v7.0.

For the comparison between the PyRadiomics versions and all its other dependencies, there was excellent agreement for most features (98/107) for CECT, IVIMf, IVIMD and IAUG images. For Ktrans and Kep images, features with poor agreement were registered for intensity (2/18 for Ktrans and 16/18 for Kep), texture (36/75 for Kep) and wavelet (61/744 for Ktrans and 382/744 for Kep) categories. Maximum and Minimum features are included in those poorly reliable intensity features, with a substantial difference between versions: for Kep images, the mean Minimum and Maximum feature calculations are 18.9 and 11680 for v2.1 compared to 8.2 and 3927 for v3.0.1, respectively. These Minimum and Maximum value changes affect dependent features, such as Kurtosis, Skewness, and some texture and wavelet features.

3.4. Phase 4: inter-software comparison with clinical data

The three shape features in common across RadiomiX, LIFEx and syngo.via (i.e. Sphericity, SurfaceArea and VolumeVoxel) were in excellent agreement (ICC LCI > 0.9) for all image types (CEPT, IVIMD, IVIMf, IAUG, Ktrans, and Kep).

For the intensity features, agreement errors appear in the DWI and DCE maps, as shown in figure 4. The Minimum feature was in the excellent and good category for CECT and IVIMf, respectively, and in the poor category for IAUG, Ktrans and Kep. LIFEx was

| Feature group | RadiomiX | LIFEx | Syngo.via | Common features |
|---------------|----------|-------|-----------|-----------------|
| Shape         | 26 (21)  | 5 (3) | 17 (14)   | 3 (3)           |
| Intensity     | 47 (41)  | 47 (15) | 18 (16)   | 8 (7)           |
| GLCM          | 26 (25)  | 7 (6) | 24 (23)   | 5 (5)           |
| GLRLM         | 16 (16)  | 11 (11) | 16 (16)   | 11 (11)         |
| GLSZM         | 16 (16)  | 11 (11) | 16 (16)   | 11 (11)         |
| NGTDM         | 5 (5)    | 3 (3) | 5 (5)     | 3 (3)           |
| GLDM          | 16 (16)  | —     | —         | —               |
| NGLDM         | 17 (16)  | —     | 14 (14)   | —               |
| Fractal       | 3 (0)    | —     | —         | —               |
| Total         | 172 (156)| 84 (49) | 110 (104) | 41 (40)         |
computing negative values for *Minimum*, even if this feature cannot be negative for maps. The *Maximum* feature was also unreliable across applications for Ktrans and Kep images, with the same discrepancy that appeared for the PyRadiomics version difference (LIFEx closer to the values computed by PyRadiomics v3.0.1 compared to the other applications). A boxplot of these two features for Kep images is shown in figure 5. Those unreliable features impact the results of other *intensity*, *texture* and *wavelet* features. No intensity features are in excellent agreement between the different applications for Kep maps; however, *Entropy* and *Uniformity* are still categorised as good features.

Figure 6 shows bar graphs of the number of *texture* features in excellent, good, moderate or poor agreement for each image type. The largest discrepancy across all image types was observed for the *wavelet* features, where most features showed poor agreement (percentage of *wavelet* features with poor agreement across all software platforms: 69%...
for Kep, 77% for Ktrans, 78% for CECT, 84% for IAUG, 93% for IVIMf and 94% for IVIMD).

Overall, the level of reliability for each feature was inconsistent across the clinical data sets. The list of features with their ICC and LCI values for the three applications is available in supplementary data.

4. Discussion

This study aimed to investigate the reproducibility of radiomics features between three fully automated IBSI-compliant radiomics applications by comparing their feature nomenclature, formulas and output value. In prior studies (Foy et al 2018, Defeudis et al 2020, McNitt-Gray et al 2020, Doshi et al 2021), the lack of reproducibility of the same features was reported for applications that were not all compliant with the IBSI standard, with one study (Fornacon-Wood et al 2020) reporting better reliability if only IBSI-compliant applications were used in the analysis. To our knowledge, the study reported here is the first to investigate feature reproducibility across exclusively IBSI-compliant applications on multiple imaging modalities (CECT and MRI, including maps from IVIM and DCE).
We showed that, for the same image and segmentation, the feature values with corresponding mathematical definitions and settings were not always identical between applications. These discrepancies can affect the performance of radiomics models and limit their generalisability (Bogowicz et al 2017, Liang et al 2019, Korte et al 2021). *Shape* features were the most reliable features on clinical data, contrarily to the other study’s results (Fornacon-Wood et al 2020) that compared IBSI-compliant software, reporting an uncertainty for the *Sphericity* feature that was not present in our study. *Wavelet* features were the least reproducible, in line with the study results that also analysed this filter (Bogowicz et al 2017). These results were expected as IBSI has not yet tackled the standardisation of filter-based features. The reproducibility of *intensity* and *texture* features was heavily dependent on the imaging modality used. The results on CECT images were better than in the studies that compared applications not IBSI-compliant (Bogowicz et al 2017, Foy et al 2018), but worse results were observed with perfusion MRI-based maps. It was linked to the substantial difference in the *Minimum* and *Maximum* values between RadiomiX, LIFEx and syngo.via for those images (figure 5), which would have a knock-on effect on dependent features in the *intensity* category, as well as on the discretisation process, which results in most *texture* features not matching each other. Those large discrepancies in extreme intensity values can be linked to two different sources. The first source is the orientation resampling executed automatically in LIFEx, where images acquired with an orientation matrix different from the identity matrix (which is often the case for MRI) are reoriented in the DICOM reference coordinate system. The interpolation linked to the resampling modifies the extreme values, especially visible for the *Minimum* feature in Kep images where negative intensities appeared. This resampling can be disabled in the application’s settings, resolving one issue. We have not found a definitive explanation for the significant increase of the *Maximum* feature from LIFEx and PyRadiomics v3.0.1 to RadiomiX and syngo.via. A possible explanation is that the conversion between the input images and the ones used for feature computation is different inside each application, using unknown and different DICOM tags. As this behaviour appeared for the MRI maps, it may be due to some mishandling of DICOM tags during the pre-processing steps of the creation of those maps.

The matching of features and settings between applications was not straightforward. The naming convention should be further harmonised to better correspond to the unique identifier defined in the IBSI guidelines, as shown in phase 1. Therefore, this study has provided a detailed overview of matching nomenclature between the radiomics applications and the IBSI standard. In addition, it is challenging to match the feature extraction settings between RadiomiX, LIFEx and syngo.via. Different settings are available between applications, and sufficient documentation is not always provided to match the settings across platforms. For example, several options are available for the wavelet filter in LIFEx and syngo.via but not in RadiomiX, where the implementation is not well documented and unclear.

The utilisation of the IBSI digital phantom enabled the detection of differences and errors in feature implementation. The volume definition differences resulted in inferior reproducibility of *shape* features for RadiomiX and LIFEx in phase 2. A discretisation error in LIFEx v7.0 was discovered due to the poor reproducibility of the *discretised intensity* group and some *texture* features, with deviations from the IBSI reference values that were not present in the earlier version (v6.65). Those variations in *shape*, *discretised intensity* and *texture* features were not visible in clinical data where the intensity range and volume are larger than in the phantom, which could explain the inferior results obtained with the digital phantom. This part of the analysis shows that it is important to test each application and each version, even if the application is considered to be IBSI compliant and has therefore already been tested with the digital phantom. The result may vary from one version to another and some errors may appear in most recent versions if this phantom check is not systematically performed by the application during updates. The use of this phantom with known ground-truth values also helps to confirm the feature matching and the default parameters used in the applications.

Performing resampling and normalisation is recommended as part of a workflow in a radiomics clinical environment. Indeed, these pre-processing steps can improve feature reproducibility (Shafiq-ul-Hassan et al 2017, Scalco et al 2020, Traverso et al 2020a) and radiomics performance. Harmonisation, for raw MRI especially, reduces variability in intensity distributions between patients, imaging protocols and scanners and therefore produces more robust radiomics studies (Isaksson et al 2020, Hu et al 2021, Saint Martin et al 2021). Different harmonisation solutions are emerging, such as ComBat which shows promising results for all types of imaging biomarkers (Orlhac et al 2022), as well as normalisation using deep learning (Mali et al 2021). Despite the importance of normalisation in radiomics studies, we decided to minimise the pre-processing steps as we focused on how the features were mathematically implemented in each application, not the pre-processing around it. Indeed, IBSI provides recommendations for pre-processing but not a clear workflow. Therefore, the definition of users’ dependant workflow (e.g. how to perform image resampling: up-sampling, down-sampling, interpolation…), and their different implementations, or their unavailability, among all the applications (table 1) would introduce even more sources of variation. Image pre-processing would reduce the observed differences in feature values between radiomics applications when...
there should be no difference at all for the same image and segmentation.

Another important part of the radiomics workflow is the strong correlation between the volume and the voxel size. It was demonstrated that features, and hence radiomics signature, can be a surrogate for volume (Welch et al 2019). Analysis to identify features that have unanticipated cross-correlation with known prognostic features have to be executed to avoid false conclusions, as is proposed in published selection procedures (Traverso et al 2020b, Pfehler et al 2021). An alternative to eliminate the dependence on voxel size and VOI size is to normalise radiomics features by including the number of voxels in their mathematical definitions, as proposed by (Shaﬁq-ul-Hassan et al 2017, 2018). This could be a next step for IBSI to tackle, to improve the formulas to avoid volume-con founding effect.

There are some limitations to this study. The impact of resampling and normalisation was not tested as discussed before. The lack of pre-processing might have led to greater inter-patient variability, which might bias the clinical data results due to the metric used (ICC). As the images used are not raw images as it can be the case for some MRI sequences, and they come from the same scanner with the same imaging protocols, the variations between patients should remain limited. This study also used in-house developed programs to create the maps and perform the registration between them. As much as this should be avoided to allow easier reproducibility and validation of radiomics studies, the use of internal scripts to compensate for the lack of comprehensive applications remains a reality. Despite those limitations, the results in this study and the subsequent message remain: for equivalent inputs (image, segmentation and settings), IBSI-compliant radiomics applications currently do not provide the same output (feature values).

5. Conclusion

IBSI-compliant radiomics applications were explored in detail using different datasets (digital and clinical). We showed that the reproducibility of features between those radiomics platforms is not guaranteed, and the translation between applications of radiomics models directly using these features should be performed with care. Further harmonisation is required to better adhere to the IBSI guidelines regarding terminology, settings and calculation. Based on the results, we recommend performing a commissioning of the radiomics application before starting a radiomics study. This commissioning should be accomplished in two steps, first with the IBSI phantom and then with clinical data, as both allow the discovery of different sources of errors. This commissioning of radiomics software should be repeated when updating to a new version or when using a new imaging modality.

Acknowledgments

Zelda Paquier has a research grant from ‘L’Association Jules Bordet’. The radiology and medical physics departments in Institut Jules Bordet have a research agreement with Siemens.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Zelda Paquier  https://orcid.org/0000-0002-1041-0009

References

An C, Park Y W, Ahn S S, Han K, Kim H and Lee S-K 2021 Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results PLoS One 16 e0236152

Bogowicz M, Leijenaar R T H, Tanadini-Lang S, Risterer O, Pruschy M, Studer G, Unkelbach J, Guckenberger M, Konukoglu E and Lambin P 2017 Post-radiochemotherapy PET radiomics in head and neck cancer—the influence of radiomics implementation on the reproducibility of local control tumor models Radiother. Oncol. 125 356–91

Chalkidou A, O’Doherty M J and Marsden P K 2015 False discovery rates in PET and CT studies with texture features: a systematic review PLoS One 10 1–18

Chao S L, Metens T and Lemort M 2017 TumourMetrics: a comprehensive clinical solution for the standardization of DCE-MRI analysis in research and routine use Quant. Imaging Med. Surg. 7 496–510

Defeudis A, De Mattia C, Rizzetto F, Calderoni F, Mazzetti S, Torresin A, Vanzulli A, Regge D and Giannini V 2020 Standardization of CT radiomics features for multi-center analysis: impact of software settings and parameters Phys. Med. Biol. 65 195012

Dercle L, Henry T, Carré A, Paragios N, Deutsch E and Robert C 2021 Reinventing radiation therapy with machine learning and imaging bio-markers (radiomics): State-of-the-art, challenges and perspectives Methods 188 44–60

Doshi A M et al 2021 Assessment of renal cell carcinoma by texture analysis in clinical practice: a six-site, six-platform analysis of reliability Am. J. Roentgenol. 217 1132–40

Dreher C, Kuder T A, König F, Mlynarska-Bujny A, Tenconi C, Paech D, Schlemmer H-P, Ladd M E and Bickelhaupt S 2020 Radiomics in diffusion data: a test–retest, inter- and intra-reader DWI phantom study Clin. Radiol. 75 798.e13–98.e22

Fedorov A et al 2012 3D Slicer as an image computing platform for the quantitative imaging network Magn. Reson. Imaging 30 1323–41

Fornacon-Wood I, Mistry H, Ackermann C J, Blackhall F, McPartlin A, Fairey-Finn C, Price G J and O’Connor J P B 2020 Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform Eur. Radiol. 30 6241–50

Foy J, Robinson K R, Li H, Giger M L, Al-Hallaq H and Arrmato S G 2018 Variation in algorithm implementation across radiomics software J. Med. Imaging 5 1
Gillies R J, Kinahan P E and Hricak H 2016 Radiomics: images are more than pictures, they are data *Radiology* **278** 563–77

Hu Z, Zhuang Q, Xiao Y, Wu G, Shi Z, Chen L, Wang Y and Yu J 2021 MIL normalization —— prerequisites for accurate MRI radiomics analysis *Comput. Biol. Med.* **133** 104403

Isaksson L J et al 2020 Effects of MRI image normalization techniques in prostate cancer radiomics *Phys. Med. Biol.* **71** 1–13

Jha A K, Mithun S, Jaiswar V, Sherkhane U B, Purandare N C, Prabhash K, Rangarajan V, Dekker A, Wee L and Traverso A 2021 Repeatability and reproducibility study of radiomic features on a phantom and human cohort *Sci. Rep.* **11** 2035

Koo T K and Li M Y 2016 A guideline of selecting and reporting intraclass correlation coefficients for reliability research *J. Chiropr. Med.* **15** 135–63

Korte J C et al 2021 Radiomics feature stability of open-source software evaluated on apparent diffusion coefficient maps in head and neck cancer *Sci. Rep.* **11** 17633

Lambin P et al 2017 Radiomics: the bridge between medical imaging and personalized medicine *Nat. Rev. Clin. Oncol.* **14** 749–62

Liang Z-G et al 2019 Comparison of radiomics tools for image analyses and clinical prediction in nasopharyngeal carcinoma *Br. J. Radiol.* **92** 20190271

Mali S A, Ibrahim A, Woodruff H C, Andrearczyk V, Müller H, Korte J C, Isaksson L J, Gillies R J, Kinahan P E and Hricak H 2016 Radiomics: images are more than pictures, they are data *Radiology* **278** 563–77

Moros E 2018 Voxel size and gray level normalization of CT radiomics features on voxel size and number of gray levels *Magn. Reson. Mater. Physics, Biol. Med.* **34** 355–66

McNitt-Gray M et al 2020 Standardization in quantitative imaging: a multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets *Tomography* **6** 118–28

Nieco C, Orlha C, Bouhdad S, Reuzé S, Goya-Outi J, Robert C, Pellot-Baraté C, Sousann M, Frouin F and Buvalt I 2021 LIFEx: A Freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity *Cancer Res.* **78** 4786–9

O’Connor J P et al 2017 Imaging biomarker roadmap for cancer studies *Nat. Rev. Clin. Oncol.* **14** 169–86

Orlha F, Eertink J J, Ciotterou A-S, Zijlstra J M, Thieblemont C, Meignan M, Boellaard R and Buvalt I 2022 A guide to combat harmonization of imaging biomarkers in multicenter studies *J. Nucl. Med.* **63** 172–9

Park J E, Kim H S, Kim D, Park S Y, Kim J Y, Cho S J and Kim J H 2020 A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical validity and quality improvement using high-dimensional imaging features *BMC Cancer* **20** 29

Pfaehler E, Mesotten L, Zhovannik I, Piepenbosch S, Thomeer M, Vanhoore V, Adriaensens P and Boellaard R 2021 Plausibility and redundancy analysis to select FDG-PET textural features in non-small cell lung cancer *Med. Phys.* **48** 1226–38

Rizzo S, Botta F, Raimondi S, Origi D, Fanciullo C, Morganti A G and Bellomi M 2018 Radiomics: the facts and the challenges of image analysis *Eur. Radiol. Exp.* **2** 36

Rogers W et al 2020 Radiomics: from qualitative to quantitative imaging *Br. J. Radiol.* **93** 20190948

Sanduleanu S, Woodruff H C, de Jong F E C, van Timmeren J E, Jochens A, Dubois I and Lambin P 2018 Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score *Radiother. Oncol.* **127** 349–60

Scalo E, Belfatto A, Mastropietro A, Rancati T, Avuzzi B, Messina A, Valdagni R and Rizzo G 2020 T2w-MRI signal normalization affects radiomics features reproducibility *Med. Phys.* **47** 1680–91

Schwier M, van Griethuysen J, Vangel M G, Shah M, Pieper S, Peled S, Temppany C, Aerts H J W L, Kikinis R, Fennissy F M and Fedorov A 2019 Repeatability of multiparametric prostate MRI radiomics features *Sci. Rep.* **9** 14941

Shafig-ul-Hassan M et al 2017 Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels *Med. Phys.* **44** 1050–62

Shafig-ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R and Moros E 2018 Voxel size and gray level normalization of CT radiomics features in lung cancer *Sci. Rep.* **8** 10545

Song I, Yin Y, Wang H, Chang Z, Liu Z and Cui L 2020 A review of original articles published in the emerging field of radiomics *Eur. J. Radiol.* **127** 108991

Tixier F, Hatt M, Le Rest c, Le Pogam A, Corcos L and Visvikis D 2012 Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18 F-FDG PET *J. Nucl. Med.* **53** 693–700

Traverso A et al 2020a Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients *Radiother. Oncol.* **143** 88–94

Traverso A, Kazmierski M, Zhovannik I, Welch M, Wee L, Jaffray D, Dekker A and Hope A 2020b Machine learning helps identifying volume-confounding effects in radiomics *Phys. Medica* **71** 24–30

Vallat R 2018 Pingouin: statistics in python *J. Open Source Softw.* **3** 1026

van Griethuysen J J M, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G H, Fillon-Robin J-C, Pieper S and Aerts H J W L 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7

Welch M L et al 2019 Vulnerabilities of radiomic signature development: the need for safeguards *Radiother. Oncol.* **130** 2–9

Yip S S F and Aerts H J W L 2016 Applications and limitations of radiomics *Nat. Rev. Clin. Oncol.* **13** 610–20

Zwanenburg A, Leger S, Vallières M and Löck S 2016 Image biomarker roadmap for cancer studies *arXiv:1612.07003*