Radiomics in nuclear medicine - robustness, reproducibility, standardisation and how to avoid data analysis traps and replication crisis

Supplementary notes

Alex Zwanenburg

Supplementary note 1: Literature review and meta-analysis

The literature review is reported using the guidelines outlined in the PRISMA statement [1] and elaborated in its explanation and elaboration [2].

Introduction

Rationale

Image biomarkers are used in radiomics for diagnostic and prognostic purposes. These biomarkers are computed from medical imaging, such as PET and SPECT imaging. Image biomarkers are sensitive to various factors involved in image acquisition, reconstruction and delineation of volumes of interest, among others. Successful use of image biomarkers in multi-centre settings, including external validation, requires identifying the degree of sensitivity of biomarkers against variations in each factor. For example, if biomarkers are highly sensitive to differences in image resolution (voxel dimensions), we should attempt to harmonize image resolution to enable reproducibility.

Several studies have investigated factors affecting reproducibility of image biomarkers and were included in several recent qualitative reviews [3–6]. A meta-analysis has not been performed to date, but is performed here to obtain quantitative insights.

Objectives

The literature review and the meta-analysis have three objectives:

1. To relate relevant findings on the repeatability and reproducibility of PET and SPECT image biomarkers to the context of multi-center reproducibility.
2. To identify the general sensitivity of radiomic analysis to acquisition, reconstruction, segmentation and radiomics processing factors.
3. To assess the inherent sensitivity to parameter differences for each of the image biomarkers defined by the image biomarker standardisation initiative (IBSI) [7].

Objectives 2 and 3 are addressed in the meta-analysis.

Methods

Eligibility criteria

Reports were considered eligible when they:

- were published in academic journals, and were:
  - peer-reviewed.
  - full-text (conference abstracts etc. were not considered).
  - written in the English language.
  - digitally available at the reviewer's institution.
- assessed repeatability, reproducibility and/or complementarity of image biomarkers in PET or SPECT imaging, as primary or secondary endpoint, using any agreement or reliability metric.
  - only technical and operator factors were considered in this context (see **outcomes**). This excludes reports where repeatability, reproducibility and/or complementarity was solely performed in relation to clinical factors (e.g. histopathology, disease staging).
- assessed biomarkers aside from known, established biomarkers (MATV, SUV_{max}, SUV_{mean}, TLG), and aside from simple statistical SUV intensity derivatives.
- were performed using adult human subjects and/or phantom experiments.

Information sources

The PubMed database was used to search for published literature. In addition, recent reviews on the same topic were perused to identify missing relevant literature [4–6].

Search strategy

The following query was used to query the PubMed database: *(radiomics OR texture OR "quantitative image" OR textural) AND (reproducibility OR repeatability OR robustness OR stability OR sensitivity OR variability) AND (SPECT OR PET)*. All available publications included in the database from its inception to the date of query (23 October 2018) were queried.

Data management

Queried abstracts were ordered by author last name and downloaded as text. An abstract included the following meta-data as well:

- Journal and issue in which the abstract was published
- Title of the work
- Author names and affiliations
- DOI and PubMed identifiers

The meta-data from the query was additionally downloaded as a comma-separated value file. This file had the same order as the downloaded abstracts. The file was converted to an OpenDocument Spreadsheet prior to screening.

After screening the abstracts for eligibility, full-text versions were downloaded and stored offline.

An online Google Sheets was used to list data items from each of the reports selected for inclusion in the review. Another Google Sheets was used to store outcome data from each data to be used in the meta-analysis.

Selection process

The reviewer (AZ) first screened the abstracts for suitability according to the eligibility criteria. Meta-data of the abstracts were available to the reviewer.

Subsequently, full-text versions of potentially eligible reports were sought. If a full-text version was available, the reviewer screened the full-texts for eligibility.

Three recent reviews were then used to identify reports that were not found through the PubMed search, but may meet the eligibility criteria [4–6]. After identification, full-text versions were sought and, if available, screened for eligibility by the reviewer.

Data collection process

Data items were collected by the reviewer as reported by the investigators, and converted to standard units (e.g. megabecquerel instead of milliCurie) when possible. Data items are reported for the repeatability, reproducibility and/or complementarity endpoints only, in case the report describes additional outcomes.

For the meta-analysis (regarding objectives 2 & 3) the following data we attempted to collect the following data for each image biomarker with regard to each reported factor:

- Normalised Bland-Altman statistics (BA (%)), including proportional variability (PV)
- Coefficient of variation (CV) for within-sample (within-subject) differences.
- Intraclass correlation coefficient (ICC)

A report was excluded from meta-analysis if none of the above metrics were reported, or if the data was neither obtained in human subjects or using anthropomorphic phantoms. Simulated data and phantoms may represent highly simplified structures and were not used to avoid potentially biasing the results.

When available, data for the above metrics were extracted directly from tables in the digital version of the report. If the data were reported in tabular form as supplementary data in a pdf format, a pdf to excel converter (pdftoexcel.com) was used to make the data accessible. Otherwise, if the data were only available in raster format figures, a high-resolution version of the figure was downloaded and digitisation software (Engauge Digitizer 10.11) used to manually retrieve the data. If data were not directly available as numeric values, or could not be related to single features, the original investigators were contacted with the request to kindly provide the data.

If image biomarkers could not be uniquely identified by name, the reported mathematical definition (if any) was compared with the standard IBSI definitions to identify a unique match [7]. Image biomarkers that could not be uniquely matched with a standard IBSI definition were excluded from the meta-analysis. Image biomarkers that could be uniquely identified, but underwent a transformation to decrease correlations with volume and/or number of grey levels were also excluded.

Data items

The following data items were identified and collected from the reports included in the qualitative synthesis.

General study data

- *Last name of the first author.*
- *Year of appearance of the report.*
- *Site, type of disease or phantom.*
- *The used imaging tracer.*
- *The number of patients or phantoms*: If available, the total number of evaluated lesions or phantom configurations was additionally reported.
- *Whether the report concerned a multi-centric study*: If imaging was done in one centre only, but using different or multiple scanners, this was noted here. In addition, if the study was multi-centric, but evaluations were only performed using single-centre cohorts, this was noted as well.
- *Number of image biomarkers (features) assessed.*
- *Image biomarker families assessed*: Only the families included in the IBSI reference manual are noted. In case the investigators assessed different image biomarkers as well, this is marked by "others".
- *Study type*: One or more of the following: repeatability, reproducibility and complementarity.
- *Agreement metrics*: Metrics used to assess agreement between biomarker values for two or more different measurements. An example would be Bland-Altman analysis. Note that for reports that use the coefficient of variation (CV) as a metric, we attempted to determine whether this was within-samples CV or between-samples CV.
- *Reliability metrics*: Metrics used to assess reliability of image biomarkers across different settings. An example is the Intraclass Correlation Coefficient (ICC).
- *Reported metric thresholds*: Whether thresholds were reported for the used agreement and reliability metrics.

Factors evaluated for repeatability, reproducibility and complementarity

- *Name of factor*: For a full list, see the **outcomes** section below.

Cohort details

- *Cohort details*: Whether details concerning the patient cohort were described.

Acquisition details:

- *Minimum fasting time*: May be of relevance for FDG-PET, as glucose competes with tracer uptake.
- *Blood glucose concentration*: Same as above. Blood glucose levels are determined from blood samples prior to acquisition.
- *Injected tracer activity*: The administered tracer activity.
- *Tracer uptake time*: Time elapsed between tracer administration and image acquisition.
- *Scan duration*: Time per bed position or total scanning time.

Reconstruction details

- *Reconstruction algorithm*: The algorithm used to reconstruct the image.
- *Number of reconstruction iterations*: The number of iterations used for iterative reconstruction.
- *Number of subsets*: The number of subsets used in iterative reconstruction.
- *Post-reconstruction filter width*: The width of the Gaussian post-reconstruction filter.
- *Pixel size*: Size of in-plane image pixels. If not available, matrix size.
- *Slice thickness*: Thickness of the image slice.
- *Attenuation correction*: Whether the image was attenuation corrected.
- *Scatter correction*: Whether the image was corrected for photon scattering.
- *Dead time correction*: Whether the image was corrected for detector or system dead time.
- *Random coincidence*: Whether the image was corrected for randomly coinciding photons that were not produced by the same decay event.
- *SUV normalisation*: The type of SUV normalisation used.

Segmentation

- *Segmentation method*: The type of segmentation method.

Radiomics processing details

- *Interpolation*: The kind of interpolation performed.
- *Discretisation*: Discretisation methods and parameters used to discretise the image before computation of texture features.
- *Texture matrix parameters*: Whether computation of texture matrices could be reproduced using the parameters reported by the authors.
- *Radiomics software*: The radiomics software used in the report.

Data availability

- *Availability of image data*: Allows re-producing results, or computation of additional image biomarkers.
- *Availability of feature value per factor setting*: Allows computation of other agreement and reliability metrics.
- *Availability of robustness data*: Allows meta-analysis.

Outcomes

The following factors affecting repeatability, reproducibility and complementarity were assessed as outcomes. A short description of each factor is provided.

- Repeatability

    - *Test-retest analysis*: a human subject or phantom is imaged twice within a time frame of minutes to days using the same scanner and the same protocols.
- Acquisition factors:
    - *Injected activity*: amount of PET-tracer activity that is injected before imaging.
    - *Competing substance levels*: Glucose in the blood and FDG-tracers compete for uptake in the metabolically active volume(s). High blood glucose levels prevent adequate uptake of FDG-PET. In FLT-PET, there is some evidence that thymidine concentration affects uptake.
    - *Uptake time before acquisition*: Tracer uptake is a dynamic process. Thus, the acquired image depends on the moment of acquisition.
    - *Scan duration*: As PET-imaging is based measuring pairs of gamma rays from radioactive tracer decay, a longer scan duration allows more decay events to be measured. Images with longer scan durations therefore contain less statistical noise.
    - *Scanner differences*: Images from scans from different scanners may vary due to differences in the detector, reconstruction algorithms, etc.
- Respiratory motion factors:
    - *4D breathing acquisition frames*: 4D PET bins acquisitions into a number of frames, based on the breathing cycle. The resulting images may differ.
    - *Static (3D) versus gated imaging (4D)*: Unlike 4D PET, 3D PET uses all measurements for reconstruction. As a result, 3D PET images may contain motion blur.
    - *4D breathing patterns*: 4D PET imaging quality may depend on the exhibited breathing pattern, e.g. normal, shallow, and irregular breathing.
    - *Tumour motion*: Breathing causes tumour motion, but the degree of motion depends on the location of the tumour and local tissue characteristics.
- Reconstruction factors:
    - *Choice of reconstruction algorithm*: Raw PET scans are reconstructed for human interpretation. Various algorithms exist. Most modern algorithms contain some sort of iterative reconstruction. Dependent on the scanner, TOF (time-of-flight) information may be available for reconstruction, as well as PSF (point-spread function) modelling.
    - *Number of iterations in iterative reconstruction*: Iterative reconstruction optimises its objective function over a number of iterations. The reconstructed image thus depends on the number of iterations.
    - *Number of subsets in iterative reconstruction*: Iterative reconstruction is typically performed within smaller subsets to increase computational efficiency.
    - *Width of Gaussian post-reconstruction filter*: Gaussian smoothing filters are typically used to suppress noise incurred in the iterative reconstruction.
    - *Voxel dimension difference*: Images may differ in in-plane resolution and slice thickness, which together make up the dimension of a voxel. The in-plane resolution is determined by the matrix size.
    - *Voxel size harmonisation* (*rec.*): Some image biomarkers implicitly or explicitly take voxel spacing into account. Harmonising voxel sizes ensures that voxel spacing is consistent across a cohort. This can be done through reconstructing to the same voxel spacing.
    - *Partial volume corrections*: Partial volume corrections try to decrease the effect of spillover between neighbouring voxels due to measurement uncertainties. Partial volume corrections are performed by modelling a point-spread function and performing a deconvolution on the image.
- Segmentation factors
    - *Interobserver delineation variability*: Volumes of interest (VOI) are segmented automatically, semi-automatically or manually. Multiple observers using manual or semi-automatic methods may not produce the exact same delineation, which leads to variability.

    - *Choice of (semi-) automatic segmentation algorithm*: This is similar to interobserver delineation variability, but with focus on different (semi-) automatic segmentation algorithms.
- Radiomics processing factors
    - *Voxel size harmonisation* (*intp.*): Some image biomarkers implicitly or explicitly take voxel spacing into account. Harmonising voxel sizes ensures that voxel spacing is consistent across a cohort. This can be done through image interpolation.
    - *Image interpolation algorithm*: Image interpolation is performed using algorithms that compute the image intensity at new voxel coordinates. Different algorithms have different assumptions regarding local smoothness, and may produce different images.
    - *Discretisation method for intensity binning*: Image intensities are binned for the calculation of histogram-based biomarkers and texture-matrix based biomarkers. Two common methods are fixed bin size (FBS), which divides image intensities in equally sized SUV bins, and fixed bin number (FBN), which divides intensities based on the intensity range of the VOI.
    - *Discretisation levels*: Either the bin size (FBS) or the bin number (FBN) have to provided, which affects the coarseness of the discretised images.
    - *Influence of texture parameters*:
        - *Texture matrix aggregation*: All texture-matrix based biomarkers are computed from texture matrices that are either computed per image slice (2D), per volume (3D) or after merging 2D matrices over the volume (2.5D). In addition, co-occurrence matrices and run length matrices involve directional matrices which may also be aggregated in different ways.
        - *Co-occurrence matrix symmetry*: Co-occurrence matrix may either be symmetric or asymmetric. This may result in differences if biomarkers are calculated by averaging over feature values from each co-occurrence matrix.
        - *Co-occurrence matrix distance*: The co-occurrence matrix assesses co-occurrence of intensities at a set (Chebyshev) distance.
        - *Size zone matrix linkage distance*: Size zone matrices quantify zones of voxels with the same intensity. Zones are defined by linked voxels, i.e. voxels that are within a certain distance from one another (typically: adjacent).
        - *Distance zone matrix linkage distance*: The distance zone matrix builds upon the same “zone” concept as the size zone matrix, and therefore also has a linkage distance.
        - *Distance zone matrix zone distance norm*: Distance zone matrices use the distance of zones to the border of the VOI. This distance depends on the distance norm used. The IBSI suggests using Manhattan distance norms.
        - *Neighbourhood grey tone difference matrix distance*: The NGTDM assesses intensities in a neighbourhood around a centre voxel. The distance parameter determines the size of this neighbourhood.
        - *Neighbouring grey level dependence matrix distance*: The NGLDM compares intensities from a neighbourhood around a centre voxel with that of the centre voxel. The distance parameter determines the size of this neighbourhood.
        - *Neighbouring grey level dependence matrix coarseness*: Intensities are compared with a certain tolerance, the coarseness parameter.
- Complementarity
    - *Correlation with image biomarkers*: Image biomarkers may offer similar information. Complementarity is determined by assessing correlation with, e.g. MATV, SUVmax, SUVmean, TLG. Image biomarkers should be compared with other, already established biomarkers [8].
- Other

- *Noise sensitivity*: This is similar to scan duration, but the noise is introduced on purpose.
- *Radiomic software implementation*: Radiomics software is used to perform radiomic analyses on medical imaging.
- *Static versus dynamic, parameterised scans*: Dynamic PET imaging allows quantification of tracer uptake dynamics. These can be compared with SUV in static scans.
- *SUV normalisation function*: SUV can be normalised in various ways. The main aim of normalisation is to increase comparability of SUV values between patients, who may differ in weight, for example.
- *Disease type and site*: Image biomarker robustness may dependent on the disease type and site.

Risk of bias in individual studies

Repeatability, reproducibility and complementarity studies were not seen as being exposed to selection bias. However, some bias may for example exist if voxel dimensions were not isotropic and not harmonised across the study cohort [9,10]. Other potential biases are introduced when acquisition parameters and reconstruction parameters differ across the cohort. Potential biases are assigned a quality score QS that is used to assess the quality of a report, as noted in the list below. Note that this list can be expanded in the future. For example, when the IBSI benchmarks are finalised, non-compliant software could be assessed as a potential bias.

1. *Low study cohort size*:
   - **QS NA**: phantom studies.
   - **QS 2**: ≥ 50 patients.
   - **QS 1**: between 20 and 50 patients.
   - **QS 0**: below 20 patients.
2. *Potential competition for tracer uptake*: Some radioactive tracers (notably 18F-FDG) compete with substances present in the body (e.g. blood glucose).
   - **QS NA**: no known strong competing effect or not applicable (phantom studies).
   - **QS 2**: direct measurement of presence/concentration of competing substance.
   - **QS 1**: patient instructions (e.g. fasting) only.
   - **QS 0**: otherwise.
3. *Unknown injected activity*:
   - **QS 1**: injective activity is reported.
   - **QS 0**: otherwise.
4. *Unknown tracer uptake time*: tracer uptake is a dynamic process, and static scans therefore depend on the time of measurement.
   - **QS NA**: dynamic scans are performed or not applicable (phantom studies).
   - **QS 1**: uptake time is reported.
   - **QS 0**: otherwise.
5. *Mismatch in reconstruction parameters:* Reconstruction parameters influence the appearance of the reconstructed image. Parameters under consideration were: the reconstruction algorithm, number of iterations (if applicable), number of subsets (if applicable), post-processing Gaussian filter width.
   - **QS NA**: the only factor assessed was a single reconstruction parameter.
   - **QS 2**: reconstruction parameters were consistent across the cohort.
   - **QS 1**: reconstruction parameters varied, but were comparable.
   - **QS 0**: otherwise.
6. *SUV normalization*: Standardised uptake values should be comparable across the cohort:
   - **QS NA**: phantom studies.
   - **QS 1**: SUV values were normalised (e.g. against body weight, lean body weight).
   - **QS 0**: otherwise.
7. *Manual segmentation:* Manual segmentation may introduce variability.

   - **QS NA**: segmentation was the only factor assessed, or ground truth was known (simulation studies).
   - **QS 1**: a (semi-)automatic method for segmentation was used consistently.
   - **QS 0**: otherwise.

8. *Unknown texture parameters*: whether the texture matrix parameters used for calculating texture matrix based image biomarkers are fully reported.
   - **QS NA**: no texture image biomarkers were assessed.
   - **QS 1**: relevant texture parameters could be fully identified [7].
   - **QS 0**: otherwise.
9. *Mismatch between texture and voxel dimensions*: If textures are computed in 3D, voxels should be isotropic.
   - **QS NA**: no texture image biomarkers were assessed, or this potential bias was assessed as an outcome.
   - **QS 1**: if voxels are isotropic, or textures are computed in 2D and have isotropic in-plane resolution.
   - **QS 0**: otherwise.
10. *Mismatch between voxel dimensions within the cohort*: Image biomarkers (particularly texture biomarkers) are sensitive to mismatching voxel dimensions.
    - **QS NA**: no texture image biomarkers were assessed, or this potential bias was assessed as an outcome.
    - **QS 3**: if voxel dimensions are the same across the study cohort.
    - **QS 1**: if voxel dimensions are not the same across the study cohort, but only agreement is assessed.
    - **QS 0**: otherwise.

To assess overall study quality, quality scores for all potential biases were summed. A study was seen as high quality if it achieved at least 70% of the total attainable points. Potential biases that were not scored (NA) were not taken into account when determining the total number of attainable points.

The procedure outlined above was defined before any studies were scored.

Data synthesis

Data was extracted from reports included in quantitative meta analysis, as described in the section **data collection process**. We assessed proportional variability (PV) from the Bland-Altman analysis, within-samples CV and ICC separately as response variables of a simple linear model. Sadly, we were not able to combine PV and CV metrics as in [11]. Tests using synthetic data show that PV and CV are only related by $CV = PV/\sqrt{2}$ if there are no consistent bias terms between methods. This can only be safely assumed for test-retest repeatability.

The linear model contains the factor and the image biomarker as predictor variables and was defined as follows:

$$y_i = \beta_j \delta(factor = j) + \beta_k \delta(IB = k) + \beta_m \delta(IB = k)\delta(discr = FBS) + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma)$$

In essence, the linear model has a coefficient β_j for every factor j, a coefficient β_k for every image biomarker k, and a separate coefficient β_m that captures the interaction between image biomarkers and FBS discretisation. Most studies used FBN discretisation, and the coefficient was added to distinguish between image biomarker values computed with FBN and FBS. Errors were modelled according to a normal distribution with standard deviation σ.

Prior to final model assessment, the ICC value was transformed, so that all three metrics have an optimal value of 0.0:

$$\hat{y}_{ICC} =- (y_{ICC} - 1)$$

We then tested the model fit using linear regression with an Ordinary Least Squares solver in R [12] by calculating the model R^2. Model fits were 0.479, 0.422 and 0.418 for PV, CV and ICC respectively. We then attempted to increase model fit by applying a logarithmic transformation:

$$\hat{y}_{BA} = \log(y_{BA} + 1)$$
$$\hat{y}_{CV} = \log(y_{CV} + 1)$$
$$\hat{y}_{ICC} = \log(-(y_{ICC} - 1) + 1)$$

As before, all three metrics have an optimal value of 0.0. The logarithmic transformation produced considerable improvement in model fits for PV and CV (0.751 and 0.736), and a marginal improvement for ICC (0.448). In addition, the logarithmic transformation has the added advantage that the model coefficients can be interpreted as multipliers on the original scale:

$$y^*_{PV} = \exp(\beta_j)\exp(\beta_k) - 1$$
$$y^*_{CV} = \exp(\beta_j)\exp(\beta_k) - 1$$
$$y^*_{ICC} = 2 - \exp(\beta_j)\exp(\beta_k)$$

Where y^* is the modelled metric value for factor j and biomarker (with no or FBN discretisation) k. Note that the optimal values for the modelled values are 0.0, 0.0 and 1.0 for PV, CV, and ICC, respectively, as parameters β_j and β_k are non-negative.

Model fitting was finally performed through Bayesian modelling using the Stan interface in R [13,14], as this allowed the inclusion of prior information, i.e. that parameters β_j and β_k should be non-negative. All parameters had weakly informed priors. The expected value of each coefficient was reported with a 95% credibility interval.

The linear model enables us to establish the general sensitivity introduced by each factor and overall sensitivity for each individual image biomarker. Inclusion of other variables would make the model more accurate, but were not added because of lack of supporting data. Variables that could potentially be added include the interaction between factor and image biomarker and interactions between image biomarker with various reconstruction parameters.

For consistent plotting, all metric values were scaled through division by the weighted mean metric value of a set of 9 image biomarkers. This set consisted of volume$_{\text{morph}}$ (MATV), mean$_{\text{IS}}$ (SUV$_{\text{mean}}$), max$_{\text{IS}}$ (SUV$_{\text{max}}$), entropy$_{\text{IH}}$, correlation$_{\text{CM}}$, dissimilarity$_{\text{CM}}$, zone percentage$_{\text{SZM}}$ and LZHGE$_{\text{SZM}}$, after Lasnon et al. [15]. Weighting was conducted using the number of studies for each biomarker as weight. This yielded scaling factors of 1.31, 1.52 and 0.137 for BA, CV and ICC, respectively. The scaled metric values were then used to select the appropriate color from a color gradient.

Assessing meta-biases

Meta-biases were not assessed. In our estimation, repeatability and reproducibility studies do not carry an inherent publication bias, as there are no clear positive or negative results. Some studies only reported on image biomarkers that were previously found to be reproducible, which may limit cumulative evidence, but does not introduce biases directly.

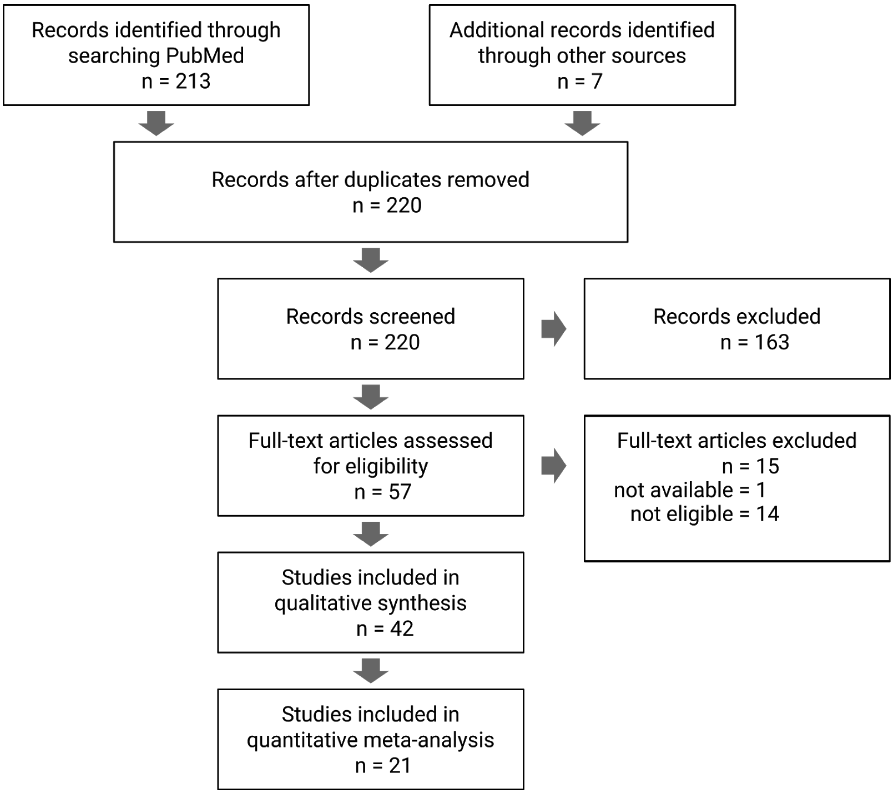
Confidence in cumulative evidence

Quality of evidence for the influence of factors and biomarkers on reproducibility was assigned to one of four categories based on a cumulative quality score (CQS). To compute this score, take the number of involved high-quality studies and add half the number of low-quality studies. The following categories were devised:

- Absent: factor or image biomarker was not assessed by any study.
- Strong: factor or image biomarker had a CQS ≥ 5.
- Moderate: factor or image biomarker had 3 ≤ CQS < 5.
- Weak: factor or image biomarker had CQS < 3.

Supplementary note 2: PRISMA flow diagram

We screened 220 abstracts, and assessed 57 full text articles. Of these, 14 were not eligible and 1 was not available to us [16] as a full-text article. Thus, 42 studies were included in the qualitative synthesis. We were able to collect data from 21 studies for inclusion in the quantitative meta-analysis. A PRISMA flow diagram is shown in **Suppl. Figure 1**.


Records identified through searching PubMed n = 213
Additional records identified through other sources n = 7
Records after duplicates removed n = 220
Records screened n = 220
Records excluded n = 163
Full-text articles assessed for eligibility n = 57
Full-text articles excluded n = 15 not available = 1 not eligible = 14
Studies included in qualitative synthesis n = 42
Studies included in quantitative meta-analysis n = 21

Suppl. Figure 1 | PRISMA flow diagram.

Supplementary note 3: literature overview

study	year	disease	tracer	cohort size	multicentric study
Altazi et al. [17]	2017	cervical cancer	18F-FDG	88	no
Bailly et al. [18]	2016	gastro-entero-pancreatic neuroendocrine tumours	68Ga-DOTANOC	26 (44 lesions)	no
Bashir et al. [19]	2017	NSCLC	18F-FDG	53	multiple scanners (2)
Belli et al. [20]	2018	HNC, pancreatic cancer	18F-FDG	50 (25 HNC; 25 PC; total 70 lesions)	multiple scanners (3)
Bogowicz et al. [21]	2017	HNSCC	18F-FDG	128	multiple scanners (5)
Carles et al. [22]	2017	phantom	18F-FDG	3 (28 lesions)	no
Carles et al. [23]	2018	lung cancer	18F-FDG	31 (36 lesions)	no
Desseroit et al. [24]	2017	NSCLC	18F-FDG	74	yes
Doumou et al. [25]	2015	oesophageal cancer	18F-FDG	64	multiple scanners (2)
Forgacs et al. [26]	2016	NSCLC	18F-FDG	65	no
Forgacs et al. [26]	2016	phantom	18F; 11C	1	multiple scanners (3)
Galavis et al. [27]	2010	solid tumours	18F-FDG	20	no
Gallivanone et al. [28]	2018	anthropomorphic phantom	18F-FDG	1 (38 lesions)	no
Grootjans et al. [29]	2016	lung cancer	18F-FDG	60	no
Hatt et al.[30]	2013	oesophageal cancer	18F-FDG	50	no
Hatt et al. [31]	2015	various	18F-FDG	555	yes
Lasnon et al. [15]	2016	lung cancer	18F-FDG	60 (71 lesions)	no
Leijenaar et al. [32]	2013	NSCLC	18F-FDG	11 (18 lesions) | 23 (27 lesions)	single-centre cohorts (2)
Leijenaar et al. [33]	2015	NSCLC	18F-FDG	35	no
Lovat et al. [34]	2017	neurofibromatosis-1	18F-FDG	54	multiple scanners (2)
Lu et al. [35]	2016	NPC	18F-FDG & 11C-Choline	40	no
Lv et al. [36]	2018	NPC/CN	18F-FDG	106	no
Manabe et al. [37]	2018	cardiac sarcoidosis / various malignancies	18F-FDG	37 / 52	no
Mu et al. [38]	2015	cervical cancer	18F-FDG	42	no
Nyflot et al. [39]	2015	simulated phantom		150	
Oliver et al. [40]	2015	NSCLC	18F-FDG	23	multiple scanners (2)
Oliver et al. [41]	2017	phantom	68Ge	1	no
Oliver et al. [41]	2017	NSCLC		31 (32 lesions)	no
Orlhac et al. [42]	2014	NSCLC, breast cancer, MCC	18F-FDG	106 (188 lesions)	single-centre cohorts (3)
Orlhac et al. [43]	2015	NSCLC	18F-FDG	48	no
Orlhac et al. [44]	2017	simulated phantom		10 (200)	
Presotto et al. [45]	2018	phantom	18F	2 (7 lesions)	no
Reuzé et al. [46]	2017	cervical cancer	18F-FDG	115	multiple scanners (2)
Shiri et al. [47]	2017	phantom	18F-FDG	1 (4)	no
Shiri et al. [47]	2017	various	18F-FDG	25	multiple scanners (2)
Takeda et al. [48]	2017	NSCLC	18F-FDG	26	no
Tixier et al. [49]	2012	oesophageal cancer	18F-FDG	16	no
Tixier et al. [50]	2016	NSCLC	18F-FDG	20	no
Van Velden et al. [51]	2014	MCC	18F-FDG	29	yes
Van Velden et al. [52]	2016	NSCLC	18F-FDG	11 (19 lesions)	no
Willaime et al. [53]	2013	breast cancer	18F-FLT	9	no
Wu et al. [54]	2016	NSCLC	18F-FDG	77	multiple scanners (2)
Yan et al. [55]	2015	lung lesions	18F-FDG	17 (24 lesions)	no
Yip et al. [56]	2014	NSCLC	18F-FDG	26 (34 lesions)	no
Yip et al. [57]	2017	NSCLC	18F-FDG	348	multiple scanners (7)

Suppl. Table 1 | General study information of reports included in the qualitative synthesis. CN: chronic nasopharyngitis; HNC: head and neck cancer; MCC: metastatic colorectal carcinoma; NPC: nasopharyngeal carcinoma; NSCLC: non-small-cell lung cancer; PC: pancreatic cancer.

study	year	type	n. features	feature families
Altazi et al. [17]	2017	human	79	morph, IS/IH, IVH, CM, RLM, SZM, NGTDM
Bailly et al. [18]	2016	human	17	IS, CM, RLM, SZM
Bashir et al. [19]	2017	human	83	morph, IS/IH, CM, SZM, NGTDM, others
Belli et al. [20]	2018	human	73	morph, IS/IH, LI, CM, RLM, SZM, NGTDM, NGLDM, others
Bogowicz et al. [21]	2017	human	649	morph, IS/IH, CM, RLM, SZM, others
Carles et al. [22]	2017	phantom	11	IS, IVH, CM
Carles et al. [23]	2018	human	31	morph, IS/IH, CM, RLM, NGTDM
Desseroit et al. [24]	2017	human	40	morph, IS/IH, IVH, CM, SZM, NGTDM
Doumou et al. [25]	2015	human	57	CM, RLM, SZM, NGTDM, others
Forgacs et al. [26]	2016	human	27	IS/IH, CM, RLM, SZM
Forgacs et al. [26]	2016	phantom	27	IS/IH, CM, RLM, SZM
Galavis et al. [27]	2010	human	50	IS/IH, CM, RLM, NGTDM, NGLDM
Gallivanone et al. [28]	2018	phantom	58	morph, IS/IH, CM, RLM, SZM
Grootjans et al. [29]	2016	human	7	morph, IS/IH, CM, SZM
Hatt et al.[30]	2013	human	10	morph, IS/IH, IVH, CM, SZM
Hatt et al. [31]	2015	human	9	morph, IS/IH, CM, SZM
Lasnon et al. [15]	2016	human	9	morph, IS/IH, IVH, CM, SZM
Leijenaar et al. [32]	2013	human	106	morph, IS/IH, LI, IVH, CM, RLM, SZM
Leijenaar et al. [33]	2015	human	44	CM, RLM, SZM
Lovat et al. [34]	2017	human	99	morph, IS/IH, CM, RLM, SZM, NGTDM, others
Lu et al. [35]	2016	human	88	morph, IS/IH, IVH, LI, CM, RLM, SZM, NGTDM
Lv et al. [36]	2018	human	57	morph, IS/IH, LI, CM, RLM, SZM, NGTDM
Manabe et al. [37]	2018	human	36	IS/IH, CM, RLM, SZM, NGTDM
Mu et al. [38]	2015	human	58	morph, IS/IH, CM, RLM, SZM, NGTDM, others
Nyflot et al. [39]	2015	simulation	35	IS/IH, CM, SZM, NGTDM
Oliver et al. [40]	2015	human	56	morph, IS/IH, IVH, CM, RLM
Oliver et al. [41]	2017	phantom	81	morph, IS/IH, IVH, CM, RLM, SZM
Oliver et al. [41]	2017	human	81	morph, IS/IH, IVH, CM, RLM, SZM
Orlhac et al. [42]	2014	human	41	morph, IS/IH, LI, CM, RLM, SZM, NGTDM
Orlhac et al. [43]	2015	human	9	morph, IS/IH, CM, RLM, SZM
Orlhac et al. [44]	2017	simulation	8	IS/IH, CM, LRM, SZM
Presotto et al. [45]	2018	phantom	39	CM, RLM, SZM, NGTDM, NGLDM
Reuzé et al. [46]	2017	human	11	IS/IH, LI, CM, RLM, SZM
Shiri et al. [47]	2017	phantom	100	morph, IS/IH, LI, CM, RLM, SZM, NGTDM, NGLDM, others
Shiri et al. [47]	2017	human	100	morph, IS/IH, LI, CM, RLM, SZM, NGTDM, NGLDM, others
Takeda et al. [48]	2017	human	7	morph, IS/IH, CM, SZM
Tixier et al. [49]	2012	human	23	IS/IH, CM, SZM
Tixier et al. [50]	2016	human	9	morph, IS/IH, CM, SZM
Van Velden et al. [51]	2014	human	18	morph, IS/IH, IVH
Van Velden et al. [52]	2016	human	105	morph, IS/IH, IVH, LI, CM, RLM, others
Willaime et al. [53]	2013	human	31	IS/IH, IVH, CM, SZM, NGTDM
Wu et al. [54]	2016	human	70	morph, IS/IH, CM, others
Yan et al. [55]	2015	human	61	IS/IH, LI, CM, RLM, SZM, NGTDM, NGLDM
Yip et al. [56]	2014	human	5	CM, RLM, NGTDM
Yip et al. [57]	2017	human	66	morph, IS/IH, CM, RLM, SZM

Suppl. Table 2 | Image biomarkers assessed in the reports in the qualitative analysis. IBSI nomenclature is used: CM: grey level co-occurrence matrix; IS/IH: intensity-based statistics and intensity histogram; IVH: intensity-volume histogram; LI: local intensity; morph: morphology; NGLDM: neighbouring grey level dependence matrix; NGTDM: neighbourhood grey tone difference matrix; RLM: grey level run length matrix; SZM: grey level size zone matrix.

study	year	type	assessment type	agreement metrics	reliability metrics
Altazi et al. [17]	2017	human	reproducibility	BA (%), DICE	ICC
Bailly et al. [18]	2016	human	reproducibility	CV (ws)	
Bashir et al. [19]	2017	human	reproducibility	JI	ICC
Belli et al. [20]	2018	human	repeatability, reproducibility	DICE, z-test	ICC
Bogowicz et al. [21]	2017	human	repeatability		ICC
Carles et al. [22]	2017	phantom	repeatability, reproducibility	BA (%), CV (?), pearson, RD, WSR	
Carles et al. [23]	2018	human	complementarity, reproducibility	BA (%), CV (?), pearson, spearman, WSR	
Desseroit et al. [24]	2017	human	complementarity, repeatability	BA (%), spearman	ICC
Doumou et al. [25]	2015	human	reproducibility	CCC	
Forgacs et al. [26]	2016	human	complementarity		
Forgacs et al. [26]	2016	phantom	complementarity, repeatability, reproducibility	CV (ws)	
Galavis et al. [27]	2010	human	reproducibility	percent difference	
Gallivanone et al. [28]	2018	phantom	repeatability, reproducibility	CV (ws), Friedman test, MWU	ICC
Grootjans et al. [29]	2016	human	reproducibility	WSR, KW, BA (%)	
Hatt et al.[30]	2013	human	complementarity, reproducibility	BA (%)	
Hatt et al. [31]	2015	human	complementarity, reproducibility	spearman	
Lasnon et al. [15]	2016	human	reproducibility	BA (%), Friedman test	
Leijenaar et al. [32]	2013	human	repeatability, reproducibility	BA (%)	ICC (1,1), ICC (3,1), PRC
Leijenaar et al. [33]	2015	human	reproducibility		ICC, PRC
Lovat et al. [34]	2017	human	reproducibility	WSR, spearman	ICC
Lu et al. [35]	2016	human	complementarity, reproducibility		ICC
Lv et al. [36]	2018	human	reproducibility		ICC
Manabe et al. [37]	2018	human	complementarity, reproducibility		ICC
Mu et al. [38]	2015	human	complementarity, reproducibility	DICE, Hausdorff distance, pearson	
Nyflot et al. [39]	2015	simulation	reproducibility	CV (ws), percent difference	
Oliver et al. [40]	2015	human	reproducibility	CCC, percent difference	
Oliver et al. [41]	2017	phantom	reproducibility	CCC, percent difference	
Oliver et al. [41]	2017	human	reproducibility	CCC, percent difference	
Orlhac et al. [42]	2014	human	complementarity, reproducibility	BA (%), pearson	
Orlhac et al. [43]	2015	human	complementarity		
Orlhac et al. [44]	2017	simulation	reproducibility	WSR, CV (ws)	
Presotto et al. [45]	2018	phantom	reproducibility	Cohen's d	η^2
Reuzé et al. [46]	2017	human	reproducibility	WSR	
Shiri et al. [47]	2017	phantom	reproducibility	CV (bs)	
Shiri et al. [47]	2017	human	reproducibility	CV (bs)	
Takeda et al. [48]	2017	human	reproducibility		ICC
Tixier et al. [49]	2012	human	repeatability, reproducibility	BA (%), paired t-test	ICC
Tixier et al. [50]	2016	human	reproducibility	BA (%), spearman	
Van Velden et al. [51]	2014	human	repeatability	BA (%), paired t-test	ICC
Van Velden et al. [52]	2016	human	repeatability	BA (%)	ICC
Willaime et al. [53]	2013	human	repeatability	BA (%)	ICC
Wu et al. [54]	2016	human	reproducibility		ICC
Yan et al. [55]	2015	human	reproducibility	CV (ws)	
Yip et al. [56]	2014	human	reproducibility	CV (ws), percent difference, KW, spearman	
Yip et al. [57]	2017	human	reproducibility	CV (ws), spearman	

Suppl. Table 3 | Type of assessments performed, and the involved agreement and reliability metrics. BA (%): relative Bland-Altman analysis; CCC: concordance correlation coefficient; CV: coefficient of variation; CV (?): unspecified CV; CV (bs): between-samples CV; CV (ws): within-samples CV; DICE: Sørensen–Dice coefficient; ICC: intraclass correlation coefficient; JI: Jaccard index; KW: Kruskal-Wallis test; MWU: Mann-Whitney U test; PRC: patient ranking correlation; RD: relative deviation; WSR: Wilcoxon signed rank test.

Supplementary note 4: Quality scores

Study quality was assessed by scoring ten items according to defined criteria. If studies received at least 70% of the obtainable points, it was ranked as a high-quality study. 21 of 42 studies were high-quality studies, and the remaining 21 low-quality.

study	year	type	1	2	3	4	5	6	7	8	9	10	Total score	Total obtainable	Study quality
Altazi et al. [17]	2017	human	2	0	1	1	2	0	0	0	0	3	9	15	LQ
Bailly et al. [18]	2016	human	1	1	1	1	2	0	1	0	0	3	10	15	LQ
Bashir et al. [19]	2017	human	2	NA	1	1	0	0	NA	0	0	3	7	12	LQ
Belli et al. [20]	2018	human	2	0	1	1	0	0	NA	0	0	0	4	14	LQ
Bogowicz et al. [21]	2017	human	2	2	1	1	0	0	1	0	1	3	11	15	HQ
Carles et al. [22]	2017	phantom	NA	NA	1	NA	0	NA	1	1	1	3	7	9	HQ
Carles et al. [23]	2018	human	1	2	1	1	2	0	1	1	1	3	13	15	HQ
Desseroit et al. [24]	2017	human	2	2	1	1	0	1	1	0	0	0	8	15	LQ
Doumou et al. [25]	2015	human	2	1	1	1	2	0	1	0	1	1	10	15	LQ
Forgacs et al. [26]	2016	human	2	2	1	0	0	0	1	0	0	1	7	15	LQ
Forgacs et al. [26]	2016	phantom	NA	NA	1	NA	0	NA	1	0	0	1	3	9	LQ
Galavis et al. [27]	2010	human	1	0	1	1	2	1	1	0	0	3	10	15	LQ
Gallivanone et al. [28]	2018	phantom	NA	NA	1	NA	2	NA	1	0	1	3	8	9	HQ
Grootjans et al. [29]	2016	human	2	0	1	1	2	0	1	0	0	3	10	15	LQ
Hatt et al.[30]	2013	human	2	1	1	1	2	0	1	1	1	3	13	15	HQ
Hatt et al. [31]	2015	human	2	1	0	1	2	1	0	1	0	1	9	15	LQ
Lasnon et al. [15]	2016	human	2	2	1	1	2	0	1	1	1	3	14	15	HQ
Leijenaar et al. [32]	2013	human	1	1	1	1	2	0	1	1	0	3	11	15	HQ
Leijenaar et al. [33]	2015	human	1	2	1	1	2	1	0	1	0	3	12	15	HQ
Lovat et al. [34]	2017	human	2	2	1	1	2	0	0	0	0	3	11	15	HQ
Lu et al. [35]	2016	human	1	1	1	1	2	1	0	1	1	3	12	15	HQ
Lv et al. [36]	2018	human	2	1	1	1	2	1	0	1	0	3	12	15	HQ
Manabe et al. [37]	2018	human	2	1	1	1	2	1	0	1	1	3	13	15	HQ
Mu et al. [38]	2015	human	1	1	1	1	2	1	1	0	0	3	11	15	HQ
Nyflot et al. [39]	2015	simulation	NA	NA	NA	NA	2	NA	NA	1	0	3	6	7	HQ
Oliver et al. [40]	2015	human	1	0	1	1	0	0	1	1	0	0	5	15	LQ
Oliver et al. [41]	2017	phantom	NA	NA	1	NA	1	NA	1	0	0	0	3	9	LQ
Oliver et al. [41]	2017	human	1	0	0	0	0	0	1	0	0	0	2	15	LQ
Orlhac et al. [42]	2014	human	2	2	1	1	2	1	1	1	0	3	14	15	HQ
Orlhac et al. [43]	2015	human	1	2	1	1	2	1	1	1	1	3	14	15	HQ
Orlhac et al. [44]	2017	simulation	NA	NA	NA	NA	2	NA	NA	1	1	3	7	7	HQ
Presotto et al. [45]	2018	phantom	NA	NA	1	NA	2	NA	0	0	1	3	7	9	HQ
Reuzé et al. [46]	2017	human	2	0	1	1	2	1	1	0	NA	NA	8	11	HQ
Shiri et al. [47]	2017	phantom	NA	NA	1	NA	2	NA	1	0	0	0	4	9	LQ
Shiri et al. [47]	2017	human	1	2	1	1	2	0	1	0	0	0	8	15	LQ
Takeda et al. [48]	2017	human	1	1	1	1	2	0	NA	0	0	3	9	14	LQ
Tixier et al. [49]	2012	human	0	0	1	1	2	0	1	0	0	0	5	15	LQ
Tixier et al. [50]	2016	human	1	2	1	NA	2	1	1	0	0	3	11	14	HQ
Van Velden et al. [51]	2014	human	1	2	1	1	0	1	0	NA	NA	0	6	13	LQ
Van Velden et al. [52]	2016	human	0	0	1	1	2	1	1	0	1	3	10	15	LQ
Willaime et al. [53]	2013	human	0	0	1	1	2	1	0	1	0	3	9	15	LQ
Wu et al. [54]	2016	human	2	1	1	1	0	0	NA	0	0	0	5	14	LQ
Yan et al. [55]	2015	human	0	2	1	1	2	0	1	1	0	3	11	15	HQ
Yip et al. [56]	2014	human	1	0	1	1	2	0	0	0	0	3	8	15	LQ
Yip et al. [57]	2017	human	2	0	1	1	0	0	1	0	1	3	9	15	LQ

Suppl. Table 4 | Study quality as assessed using the scoring criteria described in the *risk of bias in individual studies* section. In short, the following criteria were scored: 1. Study size; 2. Competing tracer uptake; 3. Injected activity; 4. Tracer uptake time; 5. Mismatch in reconstruction parameters in the cohort; 6. SUV normalization method; 7. Manual segmentation; 8. Texture parameters; 9. Mismatch between texture and voxel isotropy; 10. Mismatch between voxel dimensions in the cohort. HQ: high quality; LQ: low quality.

Supplementary note 5: Factor-dependent sensitivity

factor	evidence	E	LCI	UCI	eE	eLCI	eUCI
Scan duration	0|1 (o)	0.64	0.05	1.42	1.89	1.05	4.12
Static 3D vs. gated 4D	1|1 (o)	0.28	0.03	0.65	1.33	1.03	1.92
Static vs. Dynamic	1|0 (o)	1.11	0.44	1.77	3.04	1.55	5.89
Test-retest repeatability	1|5 (+)	1.33	1.11	1.58	3.80	3.03	4.87
Reconstruction method	0|1 (o)	1.05	0.75	1.37	2.86	2.12	3.93
Partial volume corrections	1|0 (o)	1.23	0.69	1.77	3.41	2.00	5.89
Voxel harmonisation (rec.)	1|0 (o)	0.86	0.48	1.24	2.35	1.61	3.46
Delineation variability	1|0 (o)	1.13	0.85	1.41	3.09	2.33	4.10
Segmentation method	2|1 (o)	0.86	0.61	1.13	2.36	1.85	3.09
Voxel harmonisation (intp.)	1|0 (o)	0.67	0.29	1.06	1.95	1.33	2.90
Discretisation levels	0|1 (o)	0.15	0.01	0.45	1.16	1.01	1.58

Suppl. Table 5 | Estimated coefficients for factors based on proportional variability. Evidence is denoted as # high-quality studies | # low-quality studies and cumulative evidence (o: weak; + moderate; ++: strong). E: expected value; LCI: lower bound of 95% credibility interval; UCI: upper bound of 95% credibility interval; eE: value of exp(E); eLCI: value of exp(LCI); eUCI: value of exp(UCI).

factor	evidence	E	LCI	UCI	eE	eLCI	eUCI
Scan duration	0|1 (o)	1.07	0.73	1.43	2.93	2.06	4.18
4D breathing frames	0|1 (o)	0.25	0.01	0.82	1.29	1.01	2.26
Reconstruction method	2|1 (o)	0.83	0.64	1.04	2.30	1.90	2.84
Number of iterations	2|1 (o)	0.59	0.40	0.80	1.81	1.49	2.22
Number of subsets	1|0 (o)	0.08	0.00	0.28	1.08	1.00	1.32
Gaussian filter width	2|1 (o)	0.80	0.61	1.01	2.23	1.83	2.75
Voxel dimension difference	2|1 (o)	1.29	1.10	1.50	3.63	2.99	4.47
Segmentation method	1|0 (o)	1.07	0.84	1.31	2.92	2.31	3.71

Suppl. Table 6 | Estimated coefficients for factors based on the within-subject coefficient of variation. Evidence is denoted as # high-quality studies | # low-quality studies and cumulative evidence (o: weak; + moderate; ++: strong). E: expected value; LCI: lower bound of 95% credibility interval; UCI: upper bound of 95% credibility interval; eE: value of exp(E); eLCI: value of exp(LCI); eUCI: value of exp(UCI).

factor	evidence	E	LCI	UCI	eE	eLCI	eUCI
Test-retest repeatability	2|4 (+)	0.018	0.001	0.043	1.018	1.001	1.044
Reconstruction method	0|1 (o)	0.068	0.004	0.193	1.071	1.004	1.213
Delineation variability	1|2 (o)	0.003	0.000	0.015	1.003	1.000	1.015
Segmentation method	1|1 (o)	0.032	0.004	0.073	1.033	1.004	1.076
Discretisation levels	2|1 (o)	0.167	0.134	0.202	1.182	1.143	1.224
Texture matrix aggregation	1|0 (o)	0.084	0.025	0.145	1.087	1.026	1.156
CM symmetry	1|0 (o)	0.022	0.001	0.076	1.022	1.001	1.079
CM distance	1|0 (o)	0.034	0.002	0.100	1.034	1.002	1.105
SZM distance	1|0 (o)	0.108	0.016	0.205	1.114	1.016	1.227
NGTDM distance	1|0 (o)	0.084	0.005	0.241	1.087	1.005	1.273

Suppl. Table 7 | Estimated coefficients for factors based on the intraclass correlation coefficient. Evidence is denoted as # high-quality studies | # low-quality studies and cumulative evidence (o: weak; + moderate; ++: strong). E: expected value; LCI: lower bound of 95% credibility interval; UCI: upper bound of 95% credibility interval; eE: value of exp(E); eLCI: value of exp(LCI); eUCI: value of exp(UCI).

Supplementary note 6: Inherent sensitivity of image biomarkers

factor	evidence	E	LCI	UCI	eE	eLCI	eUCI
morph: Volume	3|3 (+)	1.50	0.94	2.07	4.46	2.56	7.93
morph: Surface area	1|2 (o)	1.30	0.53	2.07	3.68	1.70	7.89
morph: Surface to volume ratio	1|2 (o)	1.03	0.30	1.80	2.80	1.35	6.04
morph: Compactness 1	1|2 (o)	1.16	0.37	1.91	3.18	1.44	6.76
morph: Compactness 2	1|1 (o)	1.58	0.71	2.46	4.85	2.03	11.74
morph: Spherical disproportion	1|2 (o)	0.71	0.10	1.48	2.03	1.10	4.38
morph: Sphericity	1|3 (o)	0.72	0.13	1.41	2.05	1.13	4.10
morph: Asphericity	1|0 (o)	1.99	1.09	2.88	7.31	2.98	17.75
morph: Maximum 3D diameter	1|1 (o)	1.37	0.52	2.25	3.92	1.68	9.46
morph: Integrated intensity	1|2 (o)	0.84	0.11	1.69	2.31	1.11	5.40
morph: Moran's I index	0|1 (o)	1.30	0.15	2.70	3.66	1.16	14.81
morph: Geary's C measure	0|1 (o)	0.87	0.06	2.23	2.38	1.06	9.27
LI: Local intensity peak	2|2 (+)	0.43	0.03	1.10	1.54	1.03	3.00
IS: Mean	3|6 (++)	1.28	0.81	1.75	3.61	2.25	5.74
IS: Standard deviation	2|5 (+)	1.35	0.78	1.91	3.84	2.18	6.72
IS: Variance	1|1 (o)	1.70	0.81	2.59	5.47	2.25	13.32
IS: Skewness	3|6 (++)	2.14	1.67	2.62	8.47	5.33	13.76
IS: Kurtosis	3|6 (++)	1.63	1.14	2.11	5.11	3.12	8.23
IS: Median	1|1 (o)	1.29	0.42	2.16	3.63	1.53	8.64
IS: Minimum	1|2 (o)	1.69	0.93	2.48	5.44	2.54	11.95
IS: Maximum	2|6 (++)	0.82	0.27	1.36	2.27	1.31	3.89
IS: Range	1|1 (o)	1.08	0.25	1.97	2.96	1.29	7.17
IS: Mean absolute deviation	1|1 (o)	1.11	0.29	1.99	3.04	1.33	7.32
IS: Median absolute deviation	0|1 (o)	1.40	0.16	2.87	4.08	1.18	17.57
IS: Coefficient of variation	1|5 (+)	0.82	0.28	1.39	2.27	1.32	4.01
IS: Energy	1|2 (o)	1.47	0.73	2.23	4.35	2.07	9.33
IS: Root mean square	1|2 (o)	1.00	0.25	1.75	2.73	1.28	5.76
IH: Entropy	3|4 (++)	0.41	0.04	0.90	1.50	1.04	2.46
IH: Uniformity	3|3 (+)	1.50	0.97	2.03	4.50	2.65	7.62
IVH: Vol. fraction at 10% intensity	1|1 (o)	0.54	0.03	1.47	1.71	1.04	4.37
IVH: Vol. fraction at 90% intensity	1|1 (o)	2.70	1.84	3.57	14.9	6.32	35.56
IVH: Intensity at 10% volume	1|1 (o)	1.30	0.45	2.16	3.67	1.58	8.70
IVH: Intensity at 90% volume	1|1 (o)	0.93	0.15	1.83	2.54	1.16	6.21
IVH: Vol. fraction diff. (10-90%)	1|2 (o)	0.49	0.03	1.24	1.64	1.04	3.45
IVH: Intensity diff. (10-90%)	1|2 (o)	0.86	0.18	1.62	2.36	1.19	5.04
IVH: Area under the IVH curve	2|4 (+)	0.93	0.33	1.53	2.53	1.39	4.61
CM: Joint maximum	1|2 (o)	1.42	0.67	2.16	4.14	1.95	8.65
CM: Joint average	0|1 (o)	1.27	0.40	2.13	3.56	1.50	8.41
CM: Joint variance	1|3 (o)	2.19	1.51	2.86	8.94	4.55	17.54
CM: Joint entropy	5|6 (++)	0.69	0.23	1.15	2.00	1.26	3.17
CM: Difference average	0|1 (o)	1.23	0.33	2.11	3.44	1.39	8.25
CM: Difference variance	0|2 (o)	1.36	0.62	2.11	3.90	1.85	8.26
CM: Difference entropy	1|3 (o)	0.78	0.17	1.45	2.18	1.19	4.28
CM: Sum average	1|3 (o)	1.32	0.64	1.98	3.75	1.89	7.27
CM: Sum variance	1|3 (o)	1.90	1.22	2.58	6.66	3.40	13.25
CM: Sum entropy	1|3 (o)	0.86	0.22	1.53	2.37	1.25	4.62
CM: Angular second moment	3|5 (++)	1.59	1.03	2.13	4.89	2.80	8.39
CM: Contrast	3|5 (++)	1.34	0.79	1.89	3.82	2.21	6.63
CM: Dissimilarity	5|6 (++)	1.38	0.90	1.84	3.96	2.47	6.31
CM: Inverse difference	2|3 (+)	0.74	0.17	1.37	2.10	1.18	3.95
CM: Inverse diff. norm.	1|1 (o)	0.31	0.01	1.21	1.36	1.01	3.35
CM: Inverse diff. moment	4|5 (++)	0.99	0.47	1.50	2.69	1.60	4.50
CM: Inverse diff. moment norm.	1|1 (o)	0.30	0.01	1.19	1.35	1.01	3.30
CM: Inverse variance	1|2 (o)	1.16	0.37	1.90	3.19	1.45	6.71
CM: Correlation	3|5 (++)	1.95	1.39	2.51	7.04	4.02	12.31
CM: Autocorrelation	1|3 (o)	1.31	0.63	2.00	3.72	1.88	7.42
CM: Cluster tendency	1|2 (o)	1.73	0.97	2.48	5.67	2.64	11.95
CM: Cluster shade	1|2 (o)	2.66	1.91	3.40	14.23	6.74	29.93
CM: Cluster prominence	1|3 (o)	2.00	1.32	2.70	7.41	3.75	14.84
CM: Information correlation 1	1|3 (o)	0.84	0.21	1.50	2.31	1.24	4.49

CM: Information correlation 2	1|3 (o)	0.48	0.04	1.11	1.61	1.04	3.04
RLM: Short runs emph.	3|2 (+)	0.23	0.01	0.78	1.26	1.01	2.19
RLM: Long runs emph.	3|2 (+)	0.74	0.12	1.42	2.10	1.13	4.15
RLM: Low grey level run emph.	3|2 (+)	1.66	0.99	2.33	5.24	2.69	10.24
RLM: High grey level run emph.	3|2 (+)	1.29	0.62	1.97	3.63	1.85	7.20
RLM: Short run low grey level emph.	3|2 (+)	1.51	0.82	2.18	4.52	2.26	8.85
RLM: Short run high grey level emph.	3|2 (+)	1.13	0.46	1.83	3.09	1.58	6.21
RLM: Long run low grey level emph.	3|2 (+)	1.55	0.85	2.23	4.72	2.35	9.32
RLM: Long run high grey level emph.	3|2 (+)	1.34	0.66	2.01	3.83	1.93	7.46
RLM: Grey level non-uniformity	3|2 (+)	1.25	0.56	1.92	3.48	1.75	6.83
RLM: Grey level non-uniformity norm.	0|1 (o)	1.61	0.30	3.04	4.99	1.35	20.99
RLM: Run length non-uniformity	3|2 (+)	1.32	0.64	2.00	3.74	1.89	7.40
RLM: Run length non-uniformity norm.	0|1 (o)	1.32	0.17	2.76	3.76	1.18	15.72
RLM: Run percentage	3|2 (+)	0.84	0.18	1.52	2.31	1.20	4.57
SZM: Small zone emph.	2|4 (+)	1.40	0.81	1.98	4.04	2.25	7.22
SZM: Large zone emph.	2|4 (+)	1.84	1.26	2.42	6.29	3.52	11.28
SZM: Low grey level emph.	2|4 (+)	2.38	1.80	2.95	10.84	6.02	19.18
SZM: High grey level emph.	4|5 (++)	1.85	1.36	2.33	6.37	3.91	10.25
SZM: Small zone low grey level emph.	2|4 (+)	2.38	1.79	2.99	10.85	5.98	19.81
SZM: Small zone high grey level emph.	2|4 (+)	1.96	1.37	2.53	7.10	3.92	12.58
SZM: Large zone low grey level emph.	2|4 (+)	2.21	1.64	2.77	9.09	5.14	15.98
SZM: Large zone high grey level emph.	2|4 (+)	2.06	1.45	2.65	7.82	4.28	14.10
SZM: Grey level non-uniformity	1|2 (o)	1.18	0.33	2.03	3.26	1.39	7.60
SZM: Zone size non-uniformity	1|2 (o)	1.13	0.31	1.99	3.09	1.36	7.35
SZM: Zone percentage	4|5 (++)	1.46	0.98	1.93	4.30	2.67	6.91
SZM: Grey level variance	2|3 (+)	1.85	1.25	2.44	6.38	3.50	11.53
SZM: Zone size variance	2|3 (+)	2.14	1.55	2.72	8.49	4.70	15.22
NGTDM: Coarseness	2|3 (+)	1.55	0.92	2.17	4.73	2.50	8.74
NGTDM: Contrast	2|3 (+)	1.81	1.19	2.43	6.13	3.29	11.31
NGTDM: Busyness	1|2 (o)	3.40	2.77	4.04	30.04	16.01	56.58
NGTDM: Complexity	1|3 (o)	1.40	0.71	2.10	4.06	2.04	8.13
NGTDM: Strength	1|3 (o)	1.27	0.61	1.93	3.56	1.85	6.86

Suppl. Table 8 | Estimated coefficients for image biomarkers based on the proportional variability. Evidence is denoted as # high-quality studies | # low-quality studies and cumulative evidence (o: weak; + moderate; ++: strong). E: expected value; LCI: lower bound of 95% credibility interval; UCI: upper bound of 95% credibility interval; eE: value of exp(E); eLCI: value of exp(LCI); eUCI: value of exp(UCI).

factor	evidence	E	LCI	UCI	eE	eLCI	eUCI
morph: Volume	1|0 (o)	1.68	1.15	2.22	5.38	3.17	9.24
morph: Surface area	1|0 (o)	1.29	0.74	1.82	3.65	2.09	6.16
morph: Surface to volume ratio	1|0 (o)	1.18	0.66	1.72	3.25	1.93	5.57
morph: Spherical disproportion	1|0 (o)	0.73	0.22	1.26	2.08	1.25	3.54
morph: Sphericity	1|0 (o)	0.74	0.22	1.28	2.09	1.24	3.59
LI: Local intensity peak	1|0 (o)	0.81	0.19	1.45	2.24	1.21	4.27
IS: Mean	2|1 (o)	1.55	1.19	1.9	4.69	3.28	6.70
IS: Standard deviation	1|0 (o)	2.79	2.25	3.33	16.31	9.49	28.05
IS: Variance	2|0 (o)	2.57	2.13	3.00	13.1	8.40	20.06
IS: Skewness	2|0 (o)	2.39	1.96	2.82	10.88	7.07	16.71
IS: Kurtosis	2|0 (o)	1.51	1.07	1.94	4.51	2.93	6.97
IS: Median	1|0 (o)	2.74	2.20	3.28	15.55	9.06	26.52
IS: Minimum	1|0 (o)	3.12	2.57	3.66	22.62	13.06	38.74
IS: Maximum	2|1 (o)	1.89	1.53	2.24	6.63	4.62	9.42
IS: Mean absolute deviation	1|0 (o)	2.83	2.30	3.38	16.95	10.01	29.32
IS: Coefficient of variation	1|0 (o)	1.19	0.55	1.84	3.28	1.73	6.32
IS: Energy	2|0 (o)	2.44	2.02	2.88	11.52	7.50	17.88
IS: Root mean square	1|0 (o)	2.69	2.17	3.24	14.78	8.72	25.45
IH: Entropy	2|0 (o)	0.60	0.18	1.02	1.83	1.20	2.78
IH: Uniformity	1|0 (o)	1.76	1.22	2.30	5.80	3.40	9.94
CM: Joint maximum	1|0 (o)	1.47	0.83	2.12	4.36	2.29	8.36
CM: Joint variance	2|0 (o)	1.69	1.25	2.12	5.41	3.48	8.31
CM: Joint entropy	2|1 (o)	0.34	0.05	0.68	1.40	1.05	1.98
CM: Difference variance	1|0 (o)	1.56	0.90	2.22	4.77	2.46	9.19

CM: Difference entropy	1|0 (o)	0.71	0.15	1.35	2.04	1.16	3.85
CM: Sum average	2|0 (o)	1.05	0.62	1.46	2.85	1.85	4.32
CM: Sum variance	1|0 (o)	1.62	0.98	2.27	5.07	2.65	9.73
CM: Sum entropy	1|0 (o)	0.97	0.32	1.64	2.63	1.38	5.14
CM: Angular second moment	2|1 (o)	1.46	1.09	1.81	4.29	2.98	6.08
CM: Contrast	2|1 (o)	1.79	1.43	2.15	6.01	4.19	8.55
CM: Dissimilarity	2|1 (o)	1.24	0.89	1.60	3.47	2.43	4.95
CM: Inverse difference	2|1 (o)	1.08	0.73	1.44	2.96	2.07	4.24
CM: Inverse diff. norm.	1|0 (o)	0.79	0.18	1.42	2.20	1.20	4.14
CM: Inverse diff. moment	1|0 (o)	1.32	0.67	1.98	3.73	1.95	7.22
CM: Inverse diff. moment norm.	1|0 (o)	0.80	0.20	1.43	2.22	1.23	4.19
CM: Correlation	2|1 (o)	2.12	1.76	2.47	8.33	5.80	11.85
CM: Autocorrelation	2|0 (o)	1.64	1.20	2.07	5.14	3.32	7.95
CM: Cluster shade	1|0 (o)	2.22	1.55	2.88	9.19	4.71	17.84
RLM: Short runs emph.	2|0 (o)	0.37	0.04	0.79	1.44	1.04	2.19
RLM: Long runs emph.	2|0 (o)	0.60	0.18	1.02	1.82	1.19	2.76
RLM: Low grey level run emph.	2|1 (o)	2.00	1.62	2.35	7.36	5.05	10.49
RLM: High grey level run emph.	2|1 (o)	1.40	1.03	1.76	4.07	2.81	5.79
RLM: Short run low grey level emph.	2|0 (o)	2.22	1.78	2.65	9.21	5.90	14.16
RLM: Short run high grey level emph.	2|0 (o)	1.63	1.19	2.05	5.08	3.29	7.80
RLM: Long run low grey level emph.	2|1 (o)	2.04	1.61	2.43	7.65	5.01	11.4
RLM: Long run high grey level emph.	2|0 (o)	1.64	1.20	2.07	5.18	3.33	7.93
RLM: Grey level non-uniformity	1|0 (o)	1.10	0.44	1.75	3.01	1.56	5.74
RLM: Grey level non-uniformity norm.	1|0 (o)	1.31	0.78	1.86	3.72	2.18	6.41
RLM: Run length non-uniformity	1|0 (o)	1.49	0.85	2.13	4.43	2.34	8.44
RLM: Run length non-uniformity norm.	1|0 (o)	0.09	0.00	0.38	1.09	1.00	1.47
RLM: Run percentage	2|1 (o)	0.08	0.00	0.32	1.08	1.00	1.38
RLM: Grey level variance	1|0 (o)	2.19	1.64	2.74	8.91	5.15	15.45
RLM: Run length variance	1|0 (o)	1.83	1.30	2.37	6.26	3.67	10.75
SZM: Small zone emph.	2|0 (o)	0.98	0.54	1.41	2.66	1.72	4.11
SZM: Large zone emph.	2|0 (o)	1.71	1.27	2.13	5.50	3.58	8.39
SZM: Low grey level emph.	2|1 (o)	1.97	1.59	2.33	7.14	4.91	10.26
SZM: High grey level emph.	2|1 (o)	1.55	1.18	1.90	4.71	3.27	6.71
SZM: Small zone low grey level emph.	2|0 (o)	2.19	1.75	2.63	8.92	5.78	13.85
SZM: Small zone high grey level emph.	2|1 (o)	1.70	1.33	2.05	5.45	3.78	7.80
SZM: Large zone low grey level emph.	2|1 (o)	2.29	1.94	2.64	9.84	6.94	13.99
SZM: Large zone high grey level emph.	2|0 (o)	2.03	1.58	2.47	7.59	4.87	11.76
SZM: Grey level non-uniformity	1|0 (o)	1.29	0.64	1.92	3.63	1.89	6.82
SZM: Grey level non uniformity norm.	1|0 (o)	1.22	0.68	1.75	3.39	1.97	5.75
SZM: Zone size non-uniformity	1|1 (o)	1.46	1.02	1.88	4.29	2.77	6.59
SZM: Zone size non-uniformity norm.	1|0 (o)	1.03	0.50	1.55	2.80	1.65	4.73
SZM: Zone percentage	2|1 (o)	1.02	0.66	1.37	2.77	1.93	3.92
SZM: Grey level variance	1|0 (o)	2.41	1.87	2.95	11.13	6.50	19.02
SZM: Zone size variance	1|0 (o)	2.12	1.57	2.64	8.31	4.82	14.06
NGTDM: Coarseness	2|1 (o)	1.69	1.28	2.09	5.42	3.59	8.10
NGTDM: Contrast	2|1 (o)	1.90	1.48	2.31	6.68	4.41	10.03
NGTDM: Busyness	2|1 (o)	1.83	1.42	2.24	6.24	4.12	9.36
NGTDM: Complexity	2|0 (o)	1.80	1.39	2.23	6.07	4.03	9.31
NGTDM: Strength	2|0 (o)	1.65	1.22	2.08	5.22	3.38	8.00
NGLDM: Low dependence emph.	1|0 (o)	1.25	0.62	1.90	3.49	1.86	6.69
NGLDM: High dependence emph.	1|0 (o)	1.74	1.11	2.38	5.69	3.03	10.85
NGLDM: Dep. count non-uniformity	1|0 (o)	1.45	0.79	2.10	4.26	2.21	8.20
NGLDM: Dependence count entropy	1|0 (o)	1.12	0.49	1.75	3.08	1.63	5.75
NGLDM: Dependence count energy	1|0 (o)	1.44	0.79	2.10	4.23	2.20	8.15

Suppl. Table 9 | Estimated coefficients for image biomarkers based on the within-subject coefficient of variation. Evidence is denoted as # high-quality studies | # low-quality studies and cumulative evidence (o: weak; + moderate; ++: strong). E: expected value; LCI: lower bound of 95% credibility interval; UCI: upper bound of 95% credibility interval; eE: value of exp(E); eLCI: value of exp(LCI); eUCI: value of exp(UCI).

factor	evidence	E	LCI	UCI	eE	eLCI	eUCI
morph: Volume	3|3 (+)	0.070	0.004	0.183	1.072	1.004	1.201
morph: Surface area	3|1 (+)	0.083	0.005	0.219	1.087	1.005	1.245
morph: Surface to volume ratio	3|1 (+)	0.137	0.02	0.274	1.147	1.020	1.315
morph: Compactness 1	2|1 (o)	0.111	0.010	0.262	1.118	1.010	1.299
morph: Compactness 2	2|1 (o)	0.171	0.028	0.333	1.187	1.029	1.395
morph: Spherical disproportion	2|2 (+)	0.200	0.059	0.343	1.222	1.061	1.409
morph: Sphericity	3|2 (+)	0.213	0.078	0.342	1.237	1.081	1.408
morph: Maximum 3D diameter	1|0 (o)	0.110	0.004	0.299	1.116	1.004	1.349
morph: Integrated intensity	1|3 (o)	0.057	0.003	0.187	1.058	1.003	1.205
morph: Moran's I index	0|1 (o)	0.143	0.008	0.418	1.154	1.008	1.518
morph: Geary's C measure	0|1 (o)	0.130	0.007	0.392	1.138	1.007	1.479
LI: Local intensity peak	2|1 (o)	0.057	0.003	0.190	1.058	1.003	1.209
IS: Mean	3|5 (++)	0.055	0.003	0.150	1.057	1.003	1.162
IS: Standard deviation	3|4 (++)	0.062	0.003	0.165	1.064	1.003	1.180
IS: Variance	3|1 (+)	0.130	0.017	0.272	1.139	1.017	1.313
IS: Skewness	2|4 (+)	0.189	0.065	0.310	1.208	1.067	1.364
IS: Kurtosis	2|4 (+)	0.277	0.158	0.396	1.319	1.171	1.486
IS: Median	2|1 (o)	0.080	0.005	0.225	1.083	1.005	1.253
IS: Minimum	2|2 (+)	0.183	0.044	0.326	1.201	1.045	1.385
IS: Maximum	3|6 (++)	0.055	0.003	0.148	1.057	1.003	1.159
IS: Range	1|1 (o)	0.074	0.004	0.228	1.077	1.004	1.256
IS: Mean absolute deviation	2|1 (o)	0.086	0.005	0.223	1.090	1.005	1.250
IS: Median absolute deviation	0|1 (o)	0.113	0.005	0.375	1.120	1.005	1.455
IS: Coefficient of variation	0|4 (o)	0.096	0.007	0.250	1.100	1.007	1.284
IS: Energy	3|1 (+)	0.050	0.002	0.166	1.051	1.002	1.180
IS: Root mean square	2|1 (o)	0.065	0.003	0.205	1.067	1.003	1.228
IH: Mean	1|0 (o)	0.230	0.021	0.538	1.259	1.021	1.712
IH: Variance	1|0 (o)	0.515	0.202	0.827	1.673	1.223	2.286
IH: Skewness	1|0 (o)	0.258	0.026	0.568	1.294	1.026	1.765
IH: Kurtosis	1|0 (o)	0.412	0.104	0.721	1.510	1.110	2.057
IH: Median	1|0 (o)	0.161	0.010	0.456	1.175	1.010	1.577
IH: Minimum	1|0 (o)	0.252	0.024	0.562	1.287	1.025	1.754
IH: Maximum	1|0 (o)	0.098	0.004	0.347	1.103	1.004	1.415
IH: Range	1|0 (o)	0.235	0.019	0.541	1.265	1.019	1.718
IH: Mean absolute deviation	1|0 (o)	0.550	0.234	0.860	1.733	1.264	2.363
IH: Entropy	3|4 (++)	0.207	0.097	0.321	1.230	1.102	1.378
IH: Uniformity	3|1 (+)	0.261	0.115	0.405	1.299	1.122	1.499
IVH: Vol. fraction at 10% intensity	1|1 (o)	0.228	0.034	0.449	1.256	1.035	1.566
IVH: Vol. fraction at 90% intensity	1|1 (o)	0.342	0.162	0.522	1.408	1.176	1.686
IVH: Intensity at 10% volume	1|1 (o)	0.117	0.008	0.288	1.124	1.008	1.333
IVH: Intensity at 90% volume	1|1 (o)	0.134	0.012	0.311	1.143	1.012	1.365
IVH: Vol. fraction diff. (10-90%)	1|1 (o)	0.263	0.083	0.440	1.300	1.086	1.552
IVH: Intensity diff. (10-90%)	1|1 (o)	0.070	0.003	0.228	1.072	1.003	1.257
IVH: Area under the IVH curve	1|3 (o)	0.174	0.028	0.336	1.191	1.028	1.399
CM: Joint maximum	4|0 (+)	0.156	0.009	0.450	1.168	1.009	1.569
CM: Joint variance	5|1 (++)	0.205	0.042	0.387	1.227	1.043	1.472
CM: Joint entropy	5|5 (++)	0.155	0.040	0.274	1.167	1.040	1.315
CM: Difference variance	1|0 (o)	1.142	0.059	3.272	3.133	1.061	26.365
CM: Difference entropy	4|1 (+)	0.098	0.005	0.289	1.103	1.005	1.335
CM: Sum average	5|1 (++)	0.095	0.007	0.259	1.100	1.007	1.295
CM: Sum variance	4|1 (+)	0.088	0.004	0.276	1.092	1.004	1.318
CM: Sum entropy	4|1 (+)	0.109	0.006	0.306	1.115	1.006	1.358
CM: Angular second moment	5|3 (++)	0.146	0.020	0.287	1.157	1.020	1.332
CM: Contrast	5|3 (++)	0.106	0.010	0.245	1.112	1.010	1.278
CM: Dissimilarity	5|5 (++)	0.085	0.007	0.203	1.089	1.007	1.225
CM: Inverse difference	5|3 (++)	0.063	0.004	0.171	1.065	1.004	1.187
CM: Inverse diff. norm.	4|1 (+)	0.283	0.063	0.506	1.327	1.065	1.659
CM: Inverse diff. moment	4|4 (++)	0.057	0.003	0.165	1.059	1.003	1.180
CM: Inverse diff. moment norm.	4|1 (+)	0.296	0.080	0.526	1.345	1.083	1.692
CM: Inverse variance	4|1 (+)	0.073	0.004	0.249	1.075	1.004	1.282
CM: Correlation	5|3 (++)	0.214	0.075	0.360	1.238	1.078	1.433
CM: Autocorrelation	5|1 (++)	0.099	0.006	0.265	1.104	1.007	1.304

CM: Cluster tendency	4|1 (+)	0.083	0.004	0.268	1.087	1.004	1.307
CM: Cluster shade	4|1 (+)	0.153	0.012	0.366	1.165	1.012	1.442
CM: Cluster prominence	4|1 (+)	0.089	0.006	0.281	1.093	1.006	1.324
CM: Information correlation 1	4|1 (+)	0.194	0.020	0.411	1.214	1.020	1.508
CM: Information correlation 2	4|3 (++)	0.098	0.010	0.227	1.103	1.010	1.254
RLM: Short runs emph.	5|2 (++)	0.087	0.007	0.208	1.091	1.007	1.232
RLM: Long runs emph.	5|2 (++)	0.101	0.009	0.229	1.106	1.009	1.258
RLM: Low grey level run emph.	5|1 (++)	0.200	0.039	0.386	1.222	1.040	1.471
RLM: High grey level run emph.	5|2 (++)	0.123	0.011	0.279	1.131	1.011	1.322
RLM: Short run low grey level emph.	5|1 (++)	0.192	0.029	0.380	1.212	1.030	1.462
RLM: Short run high grey level emph.	5|1 (++)	0.114	0.007	0.286	1.121	1.007	1.331
RLM: Long run low grey level emph.	5|1 (++)	0.190	0.032	0.371	1.210	1.032	1.449
RLM: Long run high grey level emph.	5|1 (++)	0.093	0.006	0.254	1.097	1.006	1.289
RLM: Grey level non-uniformity	4|1 (+)	0.073	0.003	0.254	1.075	1.003	1.289
RLM: Grey level non-uniformity norm.	1|0 (o)	0.676	0.375	0.992	1.966	1.455	2.697
RLM: Run length non-uniformity	4|1 (+)	0.082	0.004	0.271	1.085	1.004	1.312
RLM: Run length non-uniformity norm.	1|0 (o)	0.264	0.026	0.580	1.302	1.026	1.786
RLM: Run percentage	5|2 (++)	0.064	0.004	0.186	1.066	1.004	1.204
RLM: Grey level variance	3|0 (+)	0.284	0.039	0.593	1.328	1.040	1.810
RLM: Run length variance	3|0 (+)	0.668	0.355	0.988	1.951	1.427	2.687
SZM: Small zone emph.	5|2 (++)	0.344	0.187	0.508	1.411	1.206	1.662
SZM: Large zone emph.	5|2 (++)	0.287	0.130	0.443	1.333	1.138	1.558
SZM: Low grey level emph.	5|2 (++)	0.333	0.167	0.496	1.395	1.182	1.642
SZM: High grey level emph.	5|2 (++)	0.198	0.039	0.357	1.219	1.040	1.429
SZM: Small zone low grey level emph.	5|2 (++)	0.373	0.215	0.538	1.451	1.240	1.712
SZM: Small zone high grey level emph.	5|2 (++)	0.279	0.118	0.433	1.321	1.125	1.542
SZM: Large zone low grey level emph.	5|2 (++)	0.219	0.064	0.375	1.244	1.066	1.455
SZM: Large zone high grey level emph.	5|3 (++)	0.198	0.058	0.340	1.219	1.060	1.405
SZM: Grey level non-uniformity	2|1 (o)	0.116	0.005	0.388	1.122	1.005	1.474
SZM: Grey level non uniformity norm.	1|0 (o)	0.671	0.351	0.997	1.956	1.421	2.710
SZM: Zone size non-uniformity	2|1 (o)	0.163	0.008	0.449	1.177	1.008	1.566
SZM: Zone size non-uniformity norm.	1|0 (o)	0.669	0.360	0.999	1.953	1.433	2.717
SZM: Zone percentage	5|3 (++)	0.212	0.072	0.357	1.236	1.075	1.429
SZM: Grey level variance	5|2 (++)	0.167	0.025	0.324	1.182	1.025	1.383
SZM: Zone size variance	5|2 (++)	0.158	0.025	0.316	1.172	1.025	1.372
NGTDM: Coarseness	3|2 (+)	0.213	0.049	0.396	1.238	1.050	1.485
NGTDM: Contrast	3|1 (+)	0.230	0.037	0.458	1.259	1.038	1.581
NGTDM: Busyness	3|0 (+)	0.590	0.289	0.911	1.804	1.335	2.488
NGTDM: Complexity	3|1 (+)	0.247	0.040	0.469	1.280	1.041	1.598
NGTDM: Strength	3|1 (+)	0.363	0.147	0.588	1.437	1.158	1.800

Suppl. Table 10 | Estimated coefficients for image biomarkers based on the intraclass correlation coefficient. Evidence is denoted as # high-quality studies | # low-quality studies and cumulative evidence (o: weak; + moderate; ++: strong). E: expected value; LCI: lower bound of 95% credibility interval; UCI: upper bound of 95% credibility interval; eE: value of exp(E); eLCI: value of exp(LCI); eUCI: value of exp(UCI).

References

1. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6:e1000097.

2. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6:e1000100.

3. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? Eur J Nucl Med Mol Imaging. 2017;44:151–65.

4. Lovinfosse P, Visvikis D, Hustinx R, Hatt M. FDG PET radiomics: a review of the methodological aspects. Clin Transl Imaging. 2018;6:379–91.

5. Reuzé S, Schernberg A, Orlhac F, Sun R, Chargari C, Dercle L, et al. Radiomics in Nuclear Medicine Applied to Radiation Therapy: Methods, Pitfalls, and Challenges. Int J Radiat Oncol Biol Phys. 2018;102:1117–42.

6. 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.

7. Zwanenburg A, Leger S, Vallières M, Löck S, for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative [Internet]. arXiv [cs.CV]. 2016. Available from: http://arxiv.org/abs/1612.07003

8. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol. 2019;130:2–9.

9. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44:1050–62.

10. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol. 2017;56:1544–53.

11. Lodge MA. Repeatability of SUV in Oncologic 18F-FDG PET. J Nucl Med. 2017;58:523–32.

12. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/

13. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. Journal of Statistical Software, Articles. Columbia Univ., New York, NY (United States); Harvard Univ., Cambridge, MA …; 2017;76:1–32.

14. Stan Development Team. RStan: the R interface to Stan [Internet]. 2018. Available from: http://mc-stan.org/

15. Lasnon C, Majdoub M, Lavigne B, Do P, Madelaine J, Visvikis D, et al. 18F-FDG PET/CT heterogeneity quantification through textural features in the era of harmonisation programs: a focus on lung cancer. Eur J Nucl Med Mol Imaging. 2016;43:2324–35.

16. Cortes-Rodicio J, Sanchez-Merino G, Garcia-Fidalgo MA, Tobalina-Larrea I. Identification of low variability textural features for heterogeneity quantification of 18F-FDG PET/CT imaging. Rev Esp Med Nucl Imagen Mol. 2016;35:379–84.

17. 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;18:32–48.

18. Bailly C, Bodet-Milin C, Couespel S, Necib H, Kraeber-Bodéré F, Ansquer C, et al. Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. PLoS One. 2016;11:e0159984.

19. Bashir U, Azad G, Siddique MM, Dhillon S, Patel N, Bassett P, et al. The effects of segmentation algorithms on the measurement of 18F-FDG PET texture parameters in non-small cell lung cancer. EJNMMI Res. 2017;7:60.

20. Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, et al. Quantifying the robustness of [18F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. Phys Med. 2018;49:105–11.

21. Bogowicz M, Leijenaar RTH, Tanadini-Lang S, Riesterer O, Pruschy M, Studer G, et al. 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.

22. Carles M, Torres-Espallardo I, Alberich-Bayarri A, Olivas C, Bello P, Nestle U, et al. Evaluation of PET texture features with heterogeneous phantoms: complementarity and effect of motion and segmentation method. Phys Med Biol. 2017;62:652–68.

23. 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;63:065013.

24. Desseroit M-C, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, et al. 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.

25. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. Eur Radiol. 2015;25:2805–12.

26. Forgacs A, Pall Jonsson H, Dahlbom M, Daver F, D DiFranco M, Opposits G, et al. A Study on the Basic Criteria for Selecting Heterogeneity Parameters of F18-FDG PET Images. PLoS One. 2016;11:e0164113.

27. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49:1012–6.

28. Gallivanone F, Interlenghi M, D'Ambrosio D, Trifirò G, Castiglioni I. Parameters Influencing PET Imaging Features: A Phantom Study with Irregular and Heterogeneous Synthetic Lesions. Contrast Media Mol Imaging. 2018;2018:5324517.

29. Grootjans W, Tixier F, van der Vos CS, Vriens D, Le Rest CC, Bussink J, et al. The Impact of Optimal Respiratory Gating and Image Noise on Evaluation of Intratumor Heterogeneity on 18F-FDG PET Imaging of Lung Cancer. J Nucl Med. 2016;57:1692–8.

30. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur J Nucl Med Mol Imaging. 2013;40:1662–71.

31. Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med. 2015;56:38–44.

32. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta

Oncol. 2013;52:1391–7.

33. 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;5:11075.

34. Lovat E, Siddique M, Goh V, Ferner RE, Cook GJR, Warbey VS. The effect of post-injection 18F-FDG PET scanning time on texture analysis of peripheral nerve sheath tumours in neurofibromatosis-1. EJNMMI Res. 2017;7:35.

35. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of Radiomic Features in [11C]Choline and [18F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization. Mol Imaging Biol. 2016;18:935–45.

36. 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;28:3245–54.

37. Manabe O, Ohira H, Hirata K, Hayashi S, Naya M, Tsujino I, et al. Use of 18F-FDG PET/CT texture analysis to diagnose cardiac sarcoidosis. Eur J Nucl Med Mol Imaging. 2019;46:1240–7.

38. Mu W, Chen Z, Liang Y, Shen W, Yang F, Dai R, et al. Staging of cervical cancer based on tumor heterogeneity characterized by texture features on (18)F-FDG PET images. Phys Med Biol. 2015;60:5123–39.

39. Nyflot M, Bowen SR, Yang F, Byrd D, Sandison GA, Kinahan PE. Quantitative Radiomics: Effects of Stochastic Variability on PET Textural Features and Implications for Clinical Trials. Int J Radiat Oncol Biol Phys. 2015;93:E566–7.

40. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. Transl Oncol. 2015;8:524–34.

41. Oliver JA, Budzevich M, Hunt D, Moros EG, Latifi K, Dilling TJ, et al. Sensitivity of Image Features to Noise in Conventional and Respiratory-Gated PET/CT Images of Lung Cancer: Uncorrelated Noise Effects. Technol Cancer Res Treat. 2017;16:595–608.

42. Orlhac F, Soussan M, Maisonobe J-A, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med. 2014;55:414–22.

43. Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-Derived Textural Indices Reflect Tissue-Specific Uptake Pattern in Non-Small Cell Lung Cancer. PLoS One. 2015;10:e0145063.

44. Orlhac F, Nioche C, Soussan M, Buvat I. Understanding changes in tumor textural indices in PET: a comparison between visual assessment and index values in simulated and patient data. J Nucl Med. 2017;58:387–92.

45. Presotto L, Bettinardi V, De Bernardi E, Belli ML, Cattaneo GM, Broggi S, et al. PET textural features stability and pattern discrimination power for radiomics analysis: An “ad-hoc” phantoms study. Phys Med. 2018;50:66–74.

46. Reuzé S, Orlhac F, Chargari C, Nioche C, Limkin E, Riet F, et al. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. Oncotarget. 2017;8:43169–79.

47. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. Eur Radiol. 2017;27:4498–509.

48. 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.

49. 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.

50. Tixier F, Vriens D, Cheze-Le Rest C, Hatt M, Disselhorst JA, Oyen WJG, et al. Comparison of Tumor Uptake Heterogeneity Characterization Between Static and Parametric 18F-FDG PET Images in Non-Small Cell Lung Cancer. J Nucl Med. 2016;57:1033–9.

51. van Velden FHP, Nissen IA, Jongsma F, Velasquez LM, Hayes W, Lammertsma AA, et al. 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.

52. van Velden FHP, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. 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.

53. Willaime JMY, 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.

54. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW Jr, et al. Early-Stage Non-Small Cell Lung Cancer: Quantitative Imaging Characteristics of (18)F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis. Radiology. 2016;281:270–8.

55. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. J Nucl Med. 2015;56:1667–73.

56. Yip S, McCall K, Aristophanous M, Chen AB, Aerts HJWL, Berbeco R. Comparison of texture features derived from static and respiratory-gated PET images in non-small cell lung cancer. PLoS One. 2014;9:e115510.

57. Yip SSF, Parmar C, Kim J, Huynh E, Mak RH, Aerts HJWL. Impact of experimental design on PET radiomics in predicting somatic mutation status. Eur J Radiol. 2017;97:8–15.