Standardization and harmonization in radiomics

Mathieu Hatt, PhD, HDR - CR INSERM

mathieu.hatt@inserm.fr

Laboratoire du Traitement de l’Information Médicale

LaTIM, UMR INSERM-UBO 1101, Brest

Varenna, June 27th 2019
Most radiomics models are not properly validated

Main limitations for routine clinical practice:
1. Heterogeneity of workflow/features, lack of common practice → Studies are not reproducible → Meta-analysis is impossible ✓ Need for standardization (e.g. IBSI initiative)

2. Very few results on multicentric large (prospective) datasets ✓ Need for harmonization
Standardization of features
- Calibrate code with benchmarks

Transparency and quality of methods
- Follow nomenclature and reporting guidelines of the IBSI
- Rely on robust statistics and machine learning
- Sharing data, code, models

Validation of models
- Independent testing of models
- Design of prospective studies
- Randomized clinical trials

Standardization
IBSI

• 25 institutions from 8 countries (USA, Germany, The Netherlands, France, Canada, UK, Italy, Switzerland)

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019

**Standardization**

**IBSI**

- First two phases of IBSI

**Phase I**
- Feature calculation using simple phantom

**Phase II**
- Image processing using CT data of lung tumor

Total number of biomarkers tested: 487
- 2D, 2.5D, 3D …
- Merging vs Averaging of directional texture matrices, etc.


Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Partial volume effect (PVE) correction

Wavelet-based noise reduction

Intensity nonuniformity correction

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Common interpolation types: nearest neighbour, trilinear, tricubic convolution, tricubic spline.

Consider grey level rounding for CT after interpolation (HU are integers).

Consider partial volume fraction of 0.5 for ROI interpolation (other than nearest neighbour).

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Image discretisation reduces the original intensity dynamic range into a lower range of integer values.

It is a crucial step to notably reduce the noise dependence of textures.

The lowest bin ALWAYS has to be set to a value of 1!
FIXED BIN NUMBER (FBN)

→ The total number of bins is fixed for all ROIs.
→ Bin size (i.e. width) is dependent on the range of intensities in the ROI.
→ Does NOT conserve the relationship between image intensity and physical meaning (if any for definite intensities).
→ Introduces a normalization effect that can be beneficial for arbitrary intensity units and when contrast is considered important.
→ Feature values CAN be compared across different ROIs.

FIXED BIN SIZE (FBS)

→ The bin size (i.e. width) is fixed for all ROIs.
→ Total number of bins is dependent on the maximum intensity (if fixed min value).
→ Does conserve the relationship between image intensity and physical meaning IF AND ONLY IF the minimum value of the first bin is fixed for all ROIs (i.e., do not use the minimum value of the ROI; use lower bound of re-segmentation range).
→ thus, algorithm not to be used for arbitrary intensity units.
→ Feature values (most of them) CANNOT be compared across different ROIs.

FBN
Number of bins = 16

FBS
Bin size = 0.5 SUV

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Currently not included in the IBSI benchmark: histogram equalization: defined as the process of re-distributing image intensities so that the resulting bins contain a similar number of voxels $\rightarrow \textit{contrast increased}$ by flattening the histogram of image intensities.

Haralick et al., IEEE TSMC SMC-3, 1973)
Why choose a single set of texture extraction parameters for all our features?

An optimal set of parameters may exist for a given application → DIFFERENT SET FOR DIFFERENT TEXTURES!!

“Texture optimization”: compute texture features using all possible parameters, and identify the most relevant ones in the machine learning part.

### Standardization

#### IBSI

*Table: Imaging intensity units, Re-segmentation range, FBN, FBS*

<table>
<thead>
<tr>
<th>Imaging intensity units(^{(1)})</th>
<th>Re-segmentation range</th>
<th>FBN(^{(2)})</th>
<th>FBS(^{(3)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>definite</td>
<td>([a, b]) → CT</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>([a, \infty)) → PET, MRI-ADC</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>arbitrary</td>
<td>Raw MRI, filters, etc.</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

-> **Fixed bin number** (FBN) discretisation uses the actual range of intensities in the analysed ROI (re-segmented or not), and not the re-segmentation range itself (when defined).

-> **Fixed bin size** (FBS) discretisation uses the lower bound of the re-segmentation range as the minimum set value. When the re-segmentation range is not or cannot be defined (e.g. arbitrary intensity units), the use of the FBS algorithm is not recommended.
Standardization

IBSI


Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
5 × 4 × 4 (x,y,z) voxels

2.0 × 2.0 × 2.0 mm³ in size

Shaded voxels (blue): outside mask

Grey levels: 1, 3, 4, 6

Image processing: None required


Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Standardization
IBSI

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
Calibrate your code/software with IBSI benchmarks/tests.

Follow IBSI nomenclature and reporting guidelines (provide all computation details).

Always compare radiomics against tumour volume and existing prognostic factors (e.g. clinical info).

Share data, programming code, models.

Follow the FAIR guidelines in your research: all research objects need to be Findable, Accessible, Interoperable and Reusable.

Perform 3D radiomics analysis, unless the problem suggests otherwise.

Align grid centers for the interpolation process. Provide enough 3D support. Beware of slice spacing vs thickness.

Perform full merging of texture matrices, unless the problem suggests otherwise.

Do not use the FBS discretisation algorithm for arbitrary intensities (raw MRI, filters, etc.) or if a re-segmentation range is not defined.

Optimize texture features using many different extraction parameters.

Zwanenburg, et al. Standardized image biomarkers for high-throughput extraction of features from images, under review 2019
CONTACTS

• Alex Zwanenburg, coordinator of the IBSI:
  alexander.zwanenburg@nct-dresden.de

USEFUL LINKS

• Official IBSI website: https://theibsi.github.io/
• IBSI Google group: https://groups.google.com/forum/#!forum/the_ibsi
• FAIR guidelines: http://doi.org/10.1038/sdata.2016.18
Radiomics analyses rely on several sequential steps:
- Each step is associated with pitfalls → uncertainties

**Clinical data**

1. Acquiring / collecting images
2. Pre-processing (denoising, registration…)
3. Tumor or organ delineation
4. "Handcrafted" features extraction
5. Building and validating predictive models

Other -omics (genomics, transcriptomics…)
Issues with multicentric data
Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: PET

Issues with multicentric data
Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: PET

C = direct $2 \times 2 \times 2$ mm$^3$ reconstruction
A and B = $4 \times 4 \times 4$ mm$^3$ reconstruction
Resampled with:
A: nearest neighbors (no interpolation)
B: quintic b-splines (interpolation)

Issues with multicentric data
Robustness of radiomics features

 Dependency on reconstruction and acquisition settings: PET

<table>
<thead>
<tr>
<th>Image #</th>
<th>Acq. Mode</th>
<th>Grid-Size</th>
<th>Recon. Alg</th>
<th>Iter. number</th>
<th>Post-filter width (mm)</th>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2D</td>
<td>128×128</td>
<td>OSEM</td>
<td>2</td>
<td>3</td>
<td>2D-128-OSEM2-3mm</td>
</tr>
<tr>
<td>2</td>
<td>2D</td>
<td>128×128</td>
<td>OSEM</td>
<td>2</td>
<td>5</td>
<td>2D-128-OSEM2-5mm</td>
</tr>
<tr>
<td>3</td>
<td>2D</td>
<td>128×128</td>
<td>OSEM</td>
<td>4</td>
<td>5</td>
<td>2D-128-OSEM4-5mm</td>
</tr>
<tr>
<td>4</td>
<td>2D</td>
<td>256×256</td>
<td>OSEM</td>
<td>2</td>
<td>3</td>
<td>2D-256-OSEM2-3mm</td>
</tr>
<tr>
<td>5</td>
<td>2D</td>
<td>256×256</td>
<td>OSEM</td>
<td>2</td>
<td>5</td>
<td>2D-256-OSEM2-5mm</td>
</tr>
<tr>
<td>6</td>
<td>3D</td>
<td>128×128</td>
<td>ITER</td>
<td>2</td>
<td>3</td>
<td>3D-128-ITER2-3mm</td>
</tr>
<tr>
<td>7</td>
<td>3D</td>
<td>128×128</td>
<td>ITER</td>
<td>2</td>
<td>6</td>
<td>3D-128-ITER2-6mm</td>
</tr>
<tr>
<td>8</td>
<td>3D</td>
<td>128×128</td>
<td>ITER</td>
<td>4</td>
<td>6</td>
<td>3D-128-ITER4-6mm</td>
</tr>
<tr>
<td>9</td>
<td>3D</td>
<td>256×256</td>
<td>ITER</td>
<td>2</td>
<td>3</td>
<td>3D-256-ITER2-3mm</td>
</tr>
<tr>
<td>10</td>
<td>3D</td>
<td>256×256</td>
<td>ITER</td>
<td>2</td>
<td>6</td>
<td>3D-256-ITER2-6mm</td>
</tr>
</tbody>
</table>

Acq. Mode = acquisition mode; Recon. Alg = reconstruction algorithm; Iter = iteration.

Gavalis, et al. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* 2010
Issues with multicentric data
Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: PET

Gavalis, et al. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010
Issues with multicentric data
Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: CT

<table>
<thead>
<tr>
<th>CT Scanner</th>
<th>KVP</th>
<th>mAs</th>
<th>Scan Type</th>
<th>Pitch</th>
<th>Rotation time (Sec)</th>
<th>Reconstruction Kernel</th>
<th>Detector Configuration (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE Discovery STE (GE1)</td>
<td>120</td>
<td>250*</td>
<td>Helical</td>
<td>0.984</td>
<td>1.0</td>
<td>Standard</td>
<td>Det. Coverage = 40</td>
</tr>
<tr>
<td>GE Lightspeed 32 pro (GE2)</td>
<td>120</td>
<td>250*</td>
<td>Helical</td>
<td>0.984</td>
<td>1.0</td>
<td>Standard</td>
<td>Det. Coverage = 40</td>
</tr>
<tr>
<td>Philips Big Bore (P1)</td>
<td>120</td>
<td>250</td>
<td>Helical</td>
<td>1.024</td>
<td>1.0</td>
<td>Standard (B)</td>
<td>16 × 0.75</td>
</tr>
<tr>
<td>Philips Brilliance 64 (P2)</td>
<td>120</td>
<td>250</td>
<td>Helical</td>
<td>1.024</td>
<td>1.0</td>
<td>Standard (B)</td>
<td>64 × 0.625</td>
</tr>
<tr>
<td>Siemens Definition AS (S1)</td>
<td>120</td>
<td>250</td>
<td>Helical</td>
<td>1.0</td>
<td>1.0</td>
<td>I31f-2</td>
<td>64 × 0.625</td>
</tr>
<tr>
<td>Siemens Sensation 64 (S2)</td>
<td>120</td>
<td>250</td>
<td>Helical</td>
<td>1.0</td>
<td>1.0</td>
<td>B31f</td>
<td>64 × 0.625</td>
</tr>
<tr>
<td>Siemens Sensation 40 (S3)</td>
<td>120</td>
<td>250</td>
<td>Helical</td>
<td>1.0</td>
<td>1.0</td>
<td>B31f</td>
<td>40 × 0.625</td>
</tr>
<tr>
<td>Siemens Sensation 16 (S4)</td>
<td>120</td>
<td>250</td>
<td>Helical</td>
<td>1.0</td>
<td>1.0</td>
<td>B31f</td>
<td>16 × 0.75</td>
</tr>
</tbody>
</table>

Issues with multicentric data
Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: CT

Issues with multicentric data

Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: MRI

Issues with multicentric data

Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: MRI

Issues with multicentric data

Robustness of radiomics features

- Dependency on reconstruction and acquisition settings: MRI

- Dependency on reconstruction and acquisition settings

Multicentric data for validating radiomics models!

Issues with multicentric data

Robustness of radiomics features

Feature 1
Feature 2

Model
Retrain model on entire dataset

Center 1
Center 2

Responders
Non responders
Methods for harmonization

Imaging protocols

- Harmonizing acquisition and reconstruction protocols
  - Relies on existing imaging guidelines (QIN, EANM guidelines…)
  - Only possible for prospective data collection
  - Can be quite difficult to implement
  - Current guidelines only concerns basic metrics (e.g., SUVmax)
  - Some differences will never go away (e.g. scanner model)

Methods for harmonization

Processing images

- Filtering/processing the images
  - Possible for both retrospective and prospective collected images
  - Potential loss of useful information (degraded images)
  - Technically challenging as there could be hundreds of possible methods to consider
  - Likely insufficient to suppress all differences

Image 1

Image 2

Brainweb template

GAN

Transformed image 1

Transformed image 2

Hognon, et al. Standardization of Multicentric Image Datasets with Generative Adversarial Networks. IEEE MIC 2019
Selecting robust features

- Use only robust features
  - Technically easy
  - Can be done for both retrospective and prospective data
  - Loss of potentially useful variables (discarded beforehand)
  - Compromise between informative and robust difficult to determine

<table>
<thead>
<tr>
<th>Method</th>
<th>Development</th>
<th>Internal Validation</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAIR All features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1C &amp; FLAIR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAIR Robust features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1C &amp; FLAIR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Methods for harmonization
Harmonization of radiomic features

- Harmonizing variables *a posteriori* (e.g., normalization\(^1\), ComBat\(^2\))
  - Allows using all the available information (no discarded feature)
  - Technically easy
  - Can be done for both retrospective and prospective data
  - Requires to have a sample from each center

Chatterjee, et al. *Creating robust predictive radiomic models for data from independent institutions using normalization*. *IEEE TRPMS* 2019 (special issue on machine learning)

Methods for harmonization

Examples: normalization

- Normalization is standard technique in ML
  - Rescaling: features values are rescaled between 0 and 1
  - Standardization: features distributions are made to be zero mean and unit standard deviation
- Proposed approach:
  1. Create balanced sets (training and testing) through subsampling
  2. Normalize features separately for each dataset
- Radiomics trained in 94 patients (institution 1), tested in 63 (institution 2)

<table>
<thead>
<tr>
<th>FIGO</th>
<th>Mean (before)</th>
<th>Mean (after)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.64</td>
<td>0.72</td>
</tr>
<tr>
<td>F-score</td>
<td>0.48</td>
<td>0.71</td>
</tr>
<tr>
<td>MCC</td>
<td>0.34</td>
<td>0.44</td>
</tr>
</tbody>
</table>

ComBat[1] estimates $y_i$ and $\delta_i$ using Empirical Bayes estimates $y_i^*$ and $\delta_i^*$:

$$y_{ij}^{\text{ComBat}} = \frac{y_{ij} - \hat{\alpha} - x_{ij}\hat{\beta} - y_i^*}{\delta_i^*} + \hat{\alpha} + x_{ij}\hat{\beta}$$
Methods for harmonization

Examples: ComBat

Total: 190 patients from 3 centers
112 (Brest, France), 50 (Nantes, France), 28 (McGill, Canada)

- 112 (Brest, France)
  - Training / internal validation
  - 70 / 42

- External validation
  - 50 (Nantes, France)
  - 28 (McGill, Canada)

FDG PET, MRI T2, CE and ADC maps from DWI MRI

- Same PET/CT scanners in Brest and Nantes but different protocols
- Different PET/CT scanner and protocol in McGill
- Different MRI scanners and protocols in all 3 centers

Methods for harmonization

Examples: ComBat


Methods for harmonization

Examples: ComBat

Prediction of loco-regional control (LRC)

Conclusions
Recent advances in harmonization techniques for radiomics

**Multicentric radiomics studies**
- are challenging, with numerous pitfalls...
- ... but are necessary and there are solutions!

- Follow guidelines and recommendations
- Implement normalization/harmonization of features
Thanks for your attention
Methods for harmonization
Examples: ComBat

Improving ComBat

- The reference is chosen (e.g., the dataset for which confidence in curation/quality is the highest)
- Bootstrap is added in the estimation for higher robustness
Methods for harmonization

Examples: ComBat

- Improving ComBat
  - The reference is chosen (e.g., the dataset for which confidence in curation/quality is the highest)
  - Bootstrap is added in the estimation for higher robustness

- 3 pipelines: Random Forest (RF) and Support Vector Machines (SVM) (both with embedded feature selection), and multivariate regression (MR) with LASSO for feature selection.
- No modification on Brest; Nantes and McGill aligned to Brest → Features do not lose their meaning and keep realistic values

<table>
<thead>
<tr>
<th>ML method</th>
<th>Original</th>
<th>ComBat</th>
<th>Bootstrapped -ComBat</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>0.76</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>SVM</td>
<td>0.78</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>RF</td>
<td>0.80</td>
<td>0.91</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Using ComBat in case of high heterogeneity

- A CE-CT radiomics model predictive of non-response to induction chemotherapy in laryngeal cancers
- **87** patients from **5** centres, **27** different sets of CT scanners, acquisition and reconstruction settings → not possible to directly apply ComBat!
- 2 ‘batch labels’ were obtained through unsupervised hierarchical clustering in order to apply ComBat.
- Training/validation split + SMOTE for imbalanced data (15% of non responders)

<table>
<thead>
<tr>
<th>Performance</th>
<th>Original</th>
<th>ComBat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>SE</td>
</tr>
<tr>
<td>Training</td>
<td>0.61</td>
<td>60%</td>
</tr>
<tr>
<td>Validation</td>
<td>0.52</td>
<td>-</td>
</tr>
</tbody>
</table>