Validation of Tumor Delineation Algorithms
Dedicated to PET Imaging

M. Hatt

LaTIM, INSERM U650 Brest - France
Introduction

PET volume delineation/segmentation

- A large number of proposed methodologies
- semi-automated threshold-based approaches (fixed or adaptive thresholding)
- Others: based on methodologies developed in the wider segmentation field
  - Various levels of complexity and automation/user interaction
  - Various approaches and modeling
- However, no methodology is widely used in the clinical practice or research
Unreliable methodologies like fixed threshold are still being used in clinical studies!

Most methodologies lack rigorous validation for at least one or several requirements.

In spite of all these issues, new methods are still being proposed!

The objectives of this talk are therefore:

- to describe what is expected from a delineation methodology for PET
- to propose a validation framework
PET delineation requirements

- Noise
  (variability of values within homogeneous regions)
- Blur
  (limited spatial resolution)
- Spatial sampling
  (voxel dimensions)
- Uptake heterogeneities
  within the tumor
Outline

- PET delineation requirements
  - automation
  - accuracy
  - robustness
  - reproducibility and repeatability

- Validation framework and tools
  - physical phantoms
  - realistic simulated data
  - clinical data
  - measures of performance
Automation can be defined as the amount of significant user interaction.

Why is it important?
- the less automated, the more dependent on user interaction
  - the more time consuming
  - the less reproducible

Does the user have to
- detect the tumor(s) ?
- isolate the tumor(s) ?
- define a region of interest for additional measurements?
- choose or set parameters, and if yes, how many ?

Most approaches can therefore be classified as
- Semi-automatic
- Fully automatic
PET delineation requirements

Accuracy can be defined as the ability of the approach to recover the “true” object from the PET image, as far as its position, shape and volume are concerned.

Why is it important?

- radiotherapy planning: complete target coverage and tissues sparring + dose painting & boosting
- oncology: staging, prognostic and predictive factors, therapy response assessment & tumor monitoring

Accuracy can be estimated using various measures.

Estimation requires a ground-truth or surrogate of truth

- Simulated data
- Physical phantoms
- Histopathology

- Manual delineation is NOT a surrogate of truth, especially in PET!
These are often of limited interest for accuracy evaluation:
- usually only spherical and homogeneous objects
Validation framework

Accuracy: physical phantoms

- Anthropomorphic thorax phantom
- Wall-less small radioactive lesions using moulded wax tumors

- More realistic but not easily available and still only “spherical-like” and homogeneous uptake distribution

DWG Montgomery et al, Med Phys 2007
Validation framework

- Less realistic but very easily and quickly done
  - using any image processing software, draw ground-truth, then add noise and blur using 3D PSF filter to obtain test PET image
Validation framework

- Monte Carlo simulators dedicated to PET:
  - PET-SORTEO (*Simulation Of Realistic Tridimensional Emitting Objects*)
  - SIMSET (*Simulation System for Emission Tomography*)
  - GATE (*Geant4 Application for Tomography Emission*)

Validation framework

Accuracy: simulated data

- It is possible to adjust the anatomy of the phantom to any patient dataset

Courtesy of William Paul Segars, Johns Hopkins Outpatient Center
Validation framework

Accuracy: simulated data

PET

Ground-truth

Rhinoceros™

NURBS Tumor
(Non-Uniform Rational Basis Splines)

NCAT phantom
(NURBS)

PET scanner model

Patient image

Errors computation

Segmentation result

Segmentation

Simulated tumor

Tumor extraction

Simulated image

Validation framework

Accuracy: simulated data

Clinical image

Simulated image
Voxel ground-truth

Validation framework

Accuracy: simulated data

Clinical image

Simulated image
Voxel ground-truth

Validation framework

Accuracy: simulated data

Some examples of simulated lesions based on clinical ones.

- **Small, homogeneous**
  - Clinical
  - Simulated

- **Large, strongly heterogeneous**
  - Clinical
  - Simulated

- **Medium, slightly heterogeneous**
  - Clinical
  - Simulated
Validation framework

Accuracy: clinical data

- Clinical data:
  - Manual delineation is **NOT** a surrogate of truth due to high inter and intra observer variability
  - CT volume is **NOT** a surrogate of truth either! If it was, why would we bother delineating the PET signal?

- Solution: histopathology
  - Non-trivial task
  - Multiple sources of error
Validation framework

Accuracy: clinical data
Validation framework

Accuracy: clinical data
Validation framework

Accuracy: clinical data
Validation framework

Accuracy: clinical data

With ground-truth, accuracy can be evaluated:

- Volume errors: \( VE = \frac{V_{\text{seg}} - V_{\text{GT}}}{V_{\text{GT}}} \)
- Classification errors: \( EC = \frac{PCE + NCE}{V_0} \)

Volume errors should be used only as a first approximation and for non-complex shaped objects, classification errors are more representative.

\[ VE = \frac{2 + (-2)}{29} = 0 \% \]
\[ CE = \frac{2 + 2}{29} = 8.89 \% \]

PET delineation requirements

Robustness can be defined as the ability of the approach to achieve similar and satisfactory performance versus changes in images characteristics.

Why is it important?
- within-center variability: size and shape of lesion, contrast, inter- & intra-patients metabolism variability…
- multi-center trials: images coming from different scanner models, acquisition and reconstruction protocols…

Significant changes to take into account:
- Acquisition duration (noise levels)
- Voxel size (spatial sampling)
- Contrast
- Shape and size of object
- Heterogeneity within the regions

Estimation of robustness requires several acquisitions of the same known objects with variable parameters (scanner model, duration, voxel size…)
- Simulated data
- Physical phantoms
Manual delineation is sensitive to contrast, window thresholding for image visualization...

Semi or fully automated methods may be sensitive to spatial sampling, contrast, noise levels, object shape and size...

Let’s illustrate with the fixed thresholding
Validation framework

Robustness

Noise sensitive

Size sensitive

Contrast sensitive

(22 mm / 5 min / 8:1) 42% +14% VE 14% CE

(22 mm / 1 min / 8:1) 42% -11% VE 15% CE

(17 mm / 5 min / 8:1) 42% +80% VE 80% CE

(22 mm / 1 min / 4:1) 42% +8% VE 11% CE

Other parameters like acquisition protocols, scanner model, reconstruction parameters might also have a significant impact
Validation framework

Example: voxel size
18F-FDG PET acquisition; same scanner; same acquisition and data, two different voxel sizes in the reconstructions: 2x2x2 mm$^3$ or 4x4x4 mm$^3$
Even for the same scanner model, significant variability of the “optimal threshold” with respect to acquisition protocols across clinical centers.
Multiple acquisitions of the same object under various imaging conditions

- 4:1 contrast
- 8:1 contrast

Philips GEMINI

- Voxels 2x2x2 mm³
- Voxels 4x4x4 mm³

Philips GEMINI TF

- Voxels 2x2x2 mm³
- Voxels 4x4x4 mm³
Multiple acquisitions of the same object under various imaging conditions

**Validation framework**

- Robustness: data

GE Discovery LS

- 4:1 contrast
- Voxels 2x2x2 mm³

- 8:1 contrast
- Voxels 4x4x4 mm³

Siemens Biograph

- Voxels 2x2x2 mm³
- Voxels 4x4x4 mm³
Robustness with respect to object size

PET delineation requirements

- Repeatability can be defined as the ability of the approach to achieve similar results (independently of the accuracy) when applied multiple times to a single image.

- Why is it important?
  - It is the main advantage of automated method versus manual delineation!
  - Good consistency is necessary for the user to trust the tool.

- Estimation of repeatability is easy: repeat segmentation multiple times.

- Reproducibility can be defined as the ability of the approach to achieve similar results with respect to intrinsic reproducibility of PET scan acquisitions.

- Why is it important?
  - It is mandatory to define confidence intervals of tumor volume measurements, especially for tumor monitoring and therapy assessment.

- Estimation of reproducibility requires at least two scans of the same patient within a short period of time (<a few days) without any treatment.
Validation framework

Repeatability: evaluation

- Use of any PET image (clinical and/or simulated)
- Repeat several times the delineation and study consistency of measurements
- Deterministic methods (like fixed thresholding) have perfect repeatability
- Manual delineation is usually associated with high variability (~20-30%)
- Automated methodologies using iterative estimation or stochastic procedure might result in measurements variability and this should be quantified

M. Hatt et al., Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements, Journal of Nuclear Medicine 2010; in press
Validation framework

Repeatability: evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>17 esophageal lesions</th>
<th>Mean variability (%)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAB</td>
<td></td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>FCM</td>
<td></td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Fixed threshold</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adaptive threshold</td>
<td></td>
<td>2.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Manual delineation</td>
<td></td>
<td>14.1</td>
<td>12.2</td>
</tr>
<tr>
<td>(expert 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual delineation</td>
<td></td>
<td>16.4</td>
<td>11.3</td>
</tr>
<tr>
<td>(expert 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

M. Hatt et al., Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements, Journal of Nuclear Medicine 2010; in press
Validation framework

Reproducibility: data

- Use of two PET scans performed at a few days interval without any treatment
- Assumption: no physiological changes occurred between the two scans

M. Hatt *et al.*, *Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements*, *Journal of Nuclear Medicine* 2010; in press
Validation framework

M. Hatt et al. Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements, *Journal of Nuclear Medicine* 2010; in press
Validation framework

Reproducibility: evaluation

M. Hatt et al., Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements, *Journal of Nuclear Medicine* 2010; in press
Validation framework

**Adaptive threshold tumor volume measurements**

- **SUV**
- **URL**
- **95% CI**
- **Mean**
- **LRL**

**Datasets**
- ○ 18F-FDG esophageal lesions
- □ 18F-FLT breast lesions

M. Hatt *et al.*, *Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements*, *Journal of Nuclear Medicine* 2010; in press
Validation framework

Reproducibility: evaluation

M. Hatt et al., Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements, *Journal of Nuclear Medicine* 2010; in press
Conclusion

- PET delineation methodologies can only be widely used if rigorously validated for accuracy, robustness and repeatability/reproducibility.

- Such evaluation is time consuming and requires complex data to be acquired or generated and then analysed.

- Future work should consist in better evaluating existing methodologies using standardized evaluation protocols and datasets, rather than producing new delineation methods that will not be used for lack of validation.

- Actual clinical impact of accurate, robust and reproducible delineation should be clinically investigated for:
  - radiotherapy treatment planning
  - diagnosis, prognosis
  - therapy assessment and patient monitoring.
Thank you for your attention