

# VISCERAL Anatomy3 Benchmark Guidelines for Participation v2.0

## Document History

v1.0 - 20141118 - Released version of document

v2.0 - 20151906 - Updated document for continuation of Anatomy3

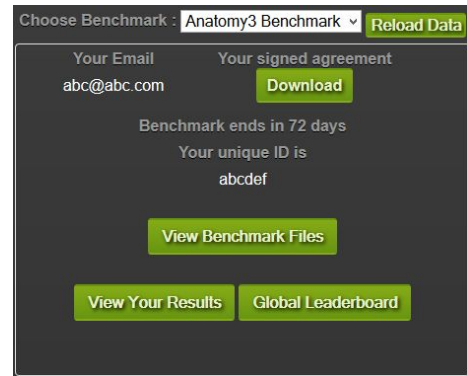
## 1. Introduction

### 1.1 Registration

The first step in participation is registration. This is done online on the following page:

<http://visceral.eu:8080/register/Registration.xhtml>

During the registration process, participants will be required to sign and upload a participation agreement. Once the participant is registered and the participation agreement has been accepted by the organisers, the account will be activated. Logging into the registration system will then reveal the *participant dashboard*.



### 1.2 Virtual Machine

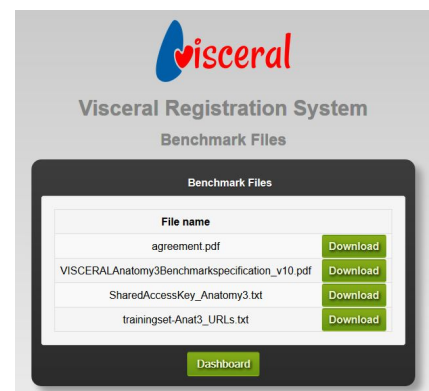
The medical imaging data is stored on the Microsoft Azure Cloud. When participants register successfully, they will receive a virtual machine (VM) in the Microsoft Azure cloud (Windows or Linux VMs are available), provided and financed with the support of Microsoft. The information for accessing the VM will appear on the participant dashboard in the registration system (after a delay of up to a week).



Note that each participating group should only register for the VISCERAL Benchmark once. After successful registration, the person completing the registration will get root access to the assigned VM, and will be able to create logins for colleagues.

Please shut down the VM if it will not be used for a long time (e.g. weekends) using the “Stop VM” button on the participant dashboard in the registration system. It can again be started using the “Start VM” button.

Documents about the Benchmark, including information on using the VM and a list of volumes can be found by pressing [View Benchmark Files](#). The evaluation software that will be used for calculating the evaluation metrics (see Section 6) is installed on the VM.



### 1.3 Training Data

The training data can be accessed from the VM and can also be downloaded via ftp (see section 1.3.1). The *Data Key* for accessing the data is provided in the file *SharedAccessKey\_Anatomy3.txt* in the Benchmark Files list (see bottom of previous page). Please only access the data on the cloud from within the assigned VM — accessing the data from outside the cloud results in additional costs for the organisers.

A list of all files in the training data set can be downloaded from the participant dashboard in the registration system by clicking on .

The image file URLs are constructed as:

cURL+filename+saKey

**cURL:** container URL,

`http://visceralstorage1.blob.core.windows.net/trainingset/`

**filename:** PatientID\_ModalityCounter\_ModalityName\_RegionName.nii.gz

**saKey:** shared access key, e.g.

`?sr=c&si=readonly&sig=Z6909Vz8TU0RxawtASpmpWZnT%2FhF2OgJOI7iEt60mis%3D`

#### 1.3.1 Training Data downloads outside the cloud

For full training set downloads, an FTP access will be provided upon request. Please send an email for further information to [oscar.jimenez@hevs.ch](mailto:oscar.jimenez@hevs.ch) CC: [ivan.eggel@hevs.ch](mailto:ivan.eggel@hevs.ch)

### 1.4 Output storage

All the VMs have a temporary storage space that is deleted with every reboot. The full path in which output files are to be stored is the temporary drive (D:) in Windows VM and /mnt/resource in Linux VM. The participant's algorithms must use only the temporary storage from their VM for any output files from their execution. All the files required for the execution of the algorithm themselves must not be stored in the temporary folders.

### 1.5 Result publication

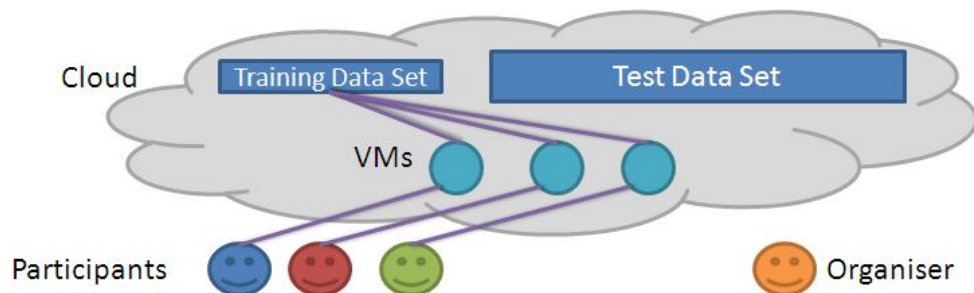
When results obtained using the VISCERAL Anatomy3 Benchmark Resources are published, please do the following:

- Ensure that the results published in the paper are also published on the VISCERAL online Leaderboard.
- Link the publication from <http://bibsonomy.org> with the tag *visceral-anatomy3*.
- Reference the following paper:  
Georg Langs, Henning Müller, Bjoern H. Menze and Allan Hanbury, VISCERAL: towards large data in medical imaging - challenges and directions. Proc. MICCAI 2012 Workshop on Medical Content-based Retrieval for Clinical Decision Support (MCBR-CDS), 2012, Springer LNCS 7723, pages 92–98, Nice, France.

## 2. Benchmark Organisation

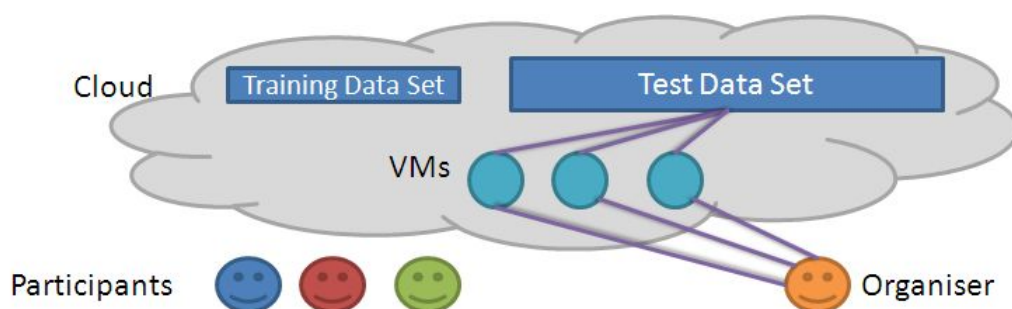
The segmentation tasks require that the organs be detected and then segmented. The benchmark runs as a series of training and testing phases. Participants can choose to publish results of the testing phases on a publicly accessible leaderboard.

### 2.1 Training phase





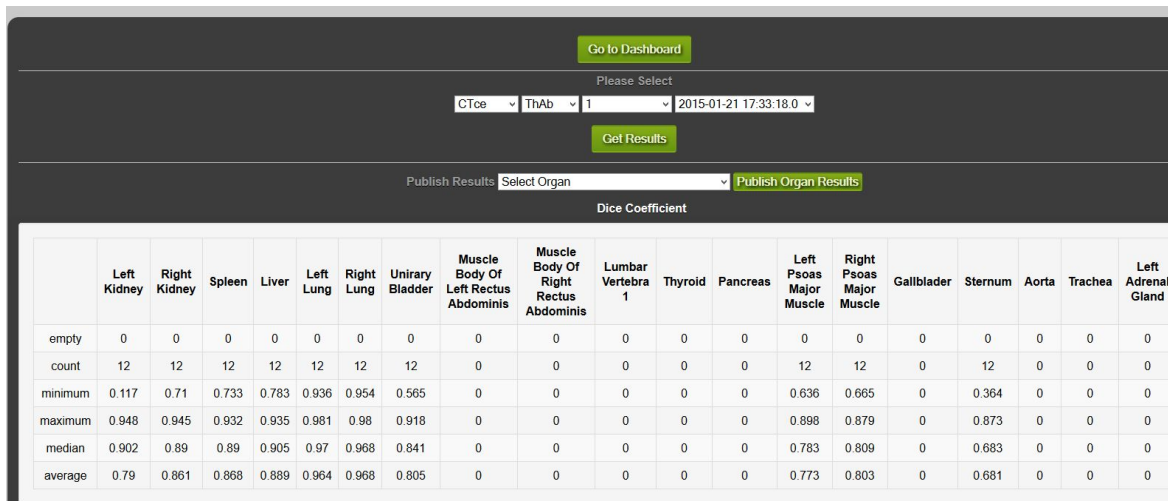
The participants each have their own VM in the cloud, linked to an annotated training data set of the same structure as the test data set. Software for carrying out the benchmark tasks must be placed into the VMs by the participants. The software must be at least executable binaries and all libraries and other support required to execute the software. Source code is not required and must be removed from the VM before submitting it if the organizers should not see it (although even if source code is there, the organizers will not copy it to anywhere outside of the VM that it is in. Participants requiring additional security with respect to code or binaries can request to sign a non-disclosure agreement with the organizers). The software must satisfy all specifications in this document. The test data set is not accessible to the participants.

### 2.2 Testing phase



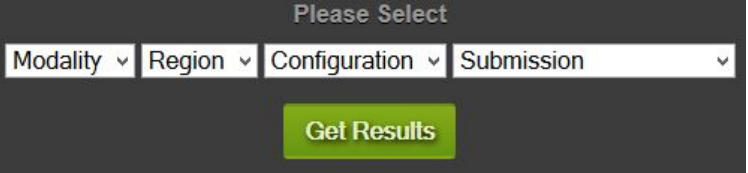
- **Once every seven days**, participants can press the **Submit VM** button to evaluate their algorithms. This lets the VISCERAL system know that it can take over the VM from the participants, execute the software installed on the VMs on the unseen test data set and evaluate the results.
- The testing should only be done once all software is installed on the VM as required by the specifications. If there is an error during the execution, the VM will be returned turned off **the following working day**.
- No feedback on the execution of the results will be provided.

- The following procedure will take place during the testing phase:
  - The temporary folder will be deleted when receiving the VM.
  - A batch-script will initially be run on a training volume to test that the output files correspond to those defined in these specifications.
  - If no output file is obtained that can be evaluated, the VM will be returned turned off **the following working day**.
  - Once the evaluation is complete (it can take up to some days, depending on how many evaluations are being done concurrently), the results will be displayed in the participant dashboard (shown below), which can be accessed by pressing .
  - The temporary folder will be deleted before returning the VM.
  - Any file stored outside the temporary folder of the VM during the execution will be deleted before returning the VM.
- **Note that the results are displayed only in this dashboard, and are only available to the person having the login credentials.** The option to publish results on the public leaderboard is available on the result display  - see Section 2.3 for more details.



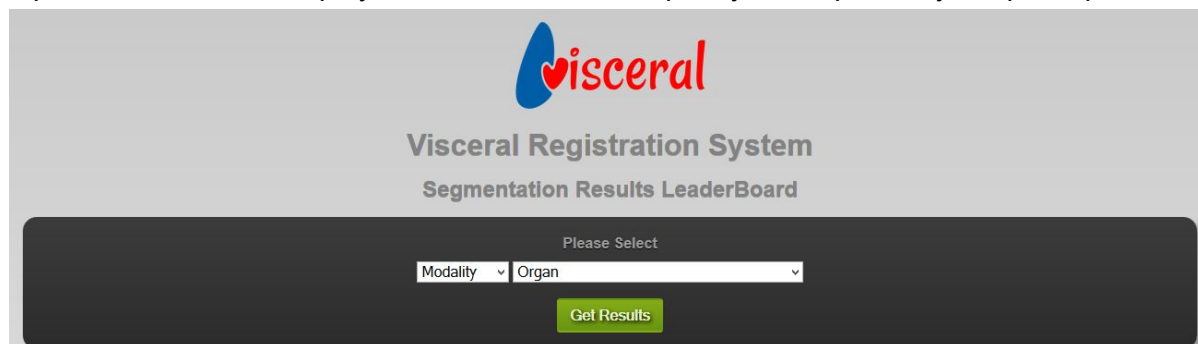
The screenshot shows a dashboard with a 'Go to Dashboard' button at the top. Below it, there are dropdown menus for 'CToe', 'ThAb', '1', and '2015-01-21 17:33:18.0'. A 'Get Results' button is present. Further down, there is a 'Publish Results' section with a 'Select Organ' dropdown and a 'Publish Organ Results' button. The main content is a table titled 'Dice Coefficient' with 20 columns representing different organs and rows for 'empty', 'count', 'minimum', 'maximum', 'median', and 'average'.

	Left Kidney	Right Kidney	Spleen	Liver	Left Lung	Right Lung	Urinary Bladder	Muscle Body Of Left Rectus Abdominis	Muscle Body Of Right Rectus Abdominis	Lumbar Vertebra 1	Thyroid	Pancreas	Left Psoas Major Muscle	Right Psoas Major Muscle	Gallbladder	Sternum	Aorta	Trachea	Left Adrenal Gland
empty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
count	12	12	12	12	12	12	12	0	0	0	0	0	12	12	0	12	0	0	0
minimum	0.117	0.71	0.733	0.783	0.936	0.954	0.565	0	0	0	0	0	0.636	0.665	0	0.364	0	0	0
maximum	0.948	0.945	0.932	0.935	0.981	0.98	0.918	0	0	0	0	0	0.898	0.879	0	0.873	0	0	0
median	0.902	0.89	0.89	0.905	0.97	0.968	0.841	0	0	0	0	0	0.783	0.809	0	0.683	0	0	0
average	0.79	0.861	0.868	0.889	0.964	0.968	0.805	0	0	0	0	0	0.773	0.803	0	0.681	0	0	0

- To see the results of a specific run, select them using the drop-down lists at the top of the page and press *Get Results*. Note that the configuration option is not used in the Anatomy3 Benchmark, and will always have a value of 1.
 
- Delays caused by the participant's executable (crashes, error messages ...) during the automatic evaluation phase might influence the total number of volumes evaluated.
- Participants should also make sure that the output segmentations are compatible with the evaluation tool before submitting their VM.

## 2.3 Making results public

A public leaderboard displays all results that are explicitly made public by the participants.



The public leaderboard displays the maximum DICE value obtained over all submissions per organ. Detailed results on each organ can be displayed by selecting modality and organ from the drop-down lists at the top and pressing *Get Results*.

Whether or not to make a results public is at the sole discretion of each participant. Making results public is done on the basis of individual organs and individual submissions. The facility to publish results is available on the results display of each participant. To publish the results of an organ, select the submission and get its results, then select the organ to publish and press the *Publish Organ Results* button.



**When publishing papers about VISCERAL Anatomy3 results, please ensure that the results published are visible on the public leaderboard.**

## 3. Further Information

Detailed information on the file formats used can be found in VISCERAL Deliverable 2.2.1<sup>1</sup>. Further information on the dataset and the modalities can be found in VISCERAL Deliverable 2.3.1<sup>2</sup>. This document also contains, in Section 3, the decisions taken regarding the manual annotation of the images.

All participants and organisers are automatically registered to the [participants-anatomy3\\_benchmark@visceral.eu](mailto:participants-anatomy3_benchmark@visceral.eu) mailing list, and can post on the mailing list. Use this list to communicate only among participants and the organisers, to ask questions, draw attention to problems or share hints and tips.

A LinkedIn group has been set-up for discussion about the Benchmark. Ask questions and make comments on this group:

<http://www.linkedin.com/groups/VISCERAL-Benchmark-Discussion-5089631>

<sup>1</sup> VISCERAL Deliverable 2.2.1:

<http://www.visceral.eu/assets/Uploads/Deliverables/VISCERAL-D2-2-1.pdf>

<sup>2</sup> VISCERAL Deliverable 2.3.1:

<http://www.visceral.eu/assets/Uploads/Deliverables/VISCERAL-D2.3.1.pdf>

## 4. Evaluation metrics

The following metrics have been selected for segmentation evaluation:

- Dice coefficient
- Average distance
- Adjusted Rand Index
- Interclass Correlation

## 5. Program and File Format Conventions

After the submission deadline, for those VMs submitted by participants (“Submit VM” button), the organisers will run the participants’ programs installed in the VMs on the unseen test data. Please ensure that all of the following naming and calling conventions and file format conventions are followed, to ensure that this works smoothly.

### 5.1 Program naming and Calling

One executable file (can be a script or compiled program) with the name

```
execute_standard or execute_standard.{extension}
```

must be in the home directory of the *azureuser* user (for those using Linux VMs), or on the Windows Desktop of the *azureuser* user (for those using Windows VMs) of the virtual machine. The *azureuser* user will be the only username available when the participant is assigned the virtual machine after registration.

#### Parameters

This executable file must take the following set of parameters, which are explained below:

```
-i [URL file to segment]
-o [output path]
-l [URL centroids for files to segment]
-r [RadLexIDs to segment]
```

The image file URLs are constructed as:

cURL+filename+saKey

**cURL:** container URL,

`http://visceralstorage1.blob.core.windows.net/trainingset/`

**filename:** `PatientID_ModalityCounter_ModalityName_RegionName.nii.gz`

**saKey:** shared access key, e.g.

`?sr=c&si=readonly&sig=Z6909Vz8TU0RxawtASpmpWZnT%2FhF2OgJOI7iEt60mis%3D`

A full list of URLs of all training set files can be downloaded from the registration system.

Standard organ segmentation	
Parameter	Explanation
-i [URL file to segment] <b>REQUIRED</b>	The URL of a file to segment + shared access key. The file will be in NIFTI.GZ format. The filename contains info on: PatientID, ModalityID, ModalityAbbreviation  cURL+PatientID_ModalityCounter_ModalityName_RegionName.nii.gz+saKey
-o [output path] <b>REQUIRED</b>	The full path in which output files are to be stored in: - Windows VM: temporary drive (D:) - Linux VM: /mnt/resource
-r [RadLexIDs to segment] <b>OPTIONAL</b>	The program should segment the organs given by the sequence of one or more RadLexID taken from the following table: - Kidneys (two, most often) Left Kidney (RadLexID 29663) Right Kidney (RadLexID 29662) - Spleen (RadLexID 86) - Liver (RadLexID 58) - Lungs (two) Left Lung (RadLexID 1326) Right Lung (RadLexID 1302) - Urinary bladder (RadLexID 237) - Rectus muscle (two) Muscle body of left rectus abdominis (RadLexID 40358) Muscle body of right rectus abdominis (RadLexID 40357) - Lumbar Vertebra 1 (RadLexID 29193) - Thyroid (RadLexID 7578) - Pancreas (RadLexID 170) - Psoas muscle (two) Left psoas major muscle (RadLexID 32249) Right psoas major muscle (RadLexID 32249) - Gallbladder (without ductus) ( RadLexID 187) - Sternum (RadLexID 2473) - Aorta (RadLexID 480) - Trachea (RadLexID 1247) - Adrenal glands (two) Left Adrenal Gland (RadLexID 30325) Right Adrenal Gland (RadLexID 30324)

## Organs to segment

It is not necessary that a program be able to segment all of the organs in the list. The program should accept all of the organ IDs in the above table, and provide no output for organ IDs that are not supported. Evaluation will only be done on the organ types that a program segments.

It is permitted to use different algorithms for different organs, but the executable file should take care of calling the correct algorithm for a given organ ID.

## Output Files

Output	Explanation
segmentation files	<p>Segmentation file is one file per segmented organ. The output file must be written to the [output path]. The file is a NIFTI.GZ file that contains binary values:  0 ... background  1 ... organ of interest.  [0,1] ... fuzzy answers are possible, too, i.e. one value in the range 0 to 1 per voxel.</p> <p>The filename has to follow the convention:</p> <p><b>PatientID_ModalityCounter_ModalityName_RegionName_RadLexID_ParticipantID_1.nii.gz</b></p> <p>Note that the first part of the filename is identical to the volume file name the segmentation is performed on:  <b>PatientID_ModalityCounter_ModalityName_RegionName</b></p> <p><b>RadLexID</b> ... identifies the anatomical structure  <b>ParticipantID</b> ... is a unique identifier for every participant (assigned by the registration system, visible on the participant information page after login).</p> <p><b>Note that the filename must end with <code>_1.nii.gz</code></b></p>

## Example

**Segmentation output:** During the evaluation phase, the VISCERAL system will call the program of the participant with participantID `5xd9t` as follows:

```
execute_standard -i
http://visceralstorage1.blob.core.windows.net/testset/CT1.nii.gz?sr=c&si=read
only&sig=Z6909Vz8TU0RxawtASpmpWznT%2FhF2OgJOI7iEt60mis%3D
-o /mnt/resource/output_p1/ -r 1247 58
```

meaning that it should segment the trachea and liver in the file `CT1.nii.gz` and write the files:

```
/mnt/resource/output_p1/CT1_1247_5xd9t_1.nii.gz
```

and

```
/mnt/resource/output_p1/CT1_58_5xd9t_1.nii.gz
```

## 6. Evaluation Software

To evaluate segmentations against the ground truth, the program `EvaluateSegmentation` is provided. This software is available on the the virtual machine assigned to each participant. Participants should make sure that the output segmentations and landmark files are compatible with the tool before submitting their VM. The latest version of `EvaluateSegmentation` is always available for download here:

<https://github.com/codalab/EvaluateSegmentation>

### Description

`EvaluateSegmentation` is a command that compares two volumes (a test segmentation and a ground truth segmentation) using 22 different metrics that were selected as a result of comprehensive research into the metrics used in the medical volume segmentations. `EvaluateSegmentation` provides the following measures:



## Similarity

1. Dice Coefficient
2. Jaccard Coefficient
3. Area under ROC Curve (one system state)
4. Cohen Kappa
5. Rand Index
6. Adjusted Rand Index
7. Interclass Correlation
8. Volumetric Similarity Coefficient
9. Mutual Information

## Distance

10. Hausdorff Distance
11. Average Distance
12. Mahanabolis Distance
13. Variation of Information
14. Global Consistency Error
15. Coefficient of Variation
16. Probabilistic Distance

## Classic measures

17. Sensitivity (Recall, true positive rate)
18. Specificity (true negative rate)
19. Precision
20. F-Measure
21. Accuracy
22. Fallout (false positive rate)

## Supported Images

`EvaluateSegmentation` supports all 3D file formats that are supported by ITK, e.g. `.nii`, `.mha`, etc. The two Images should however have the same dimensions (the same number of voxels). There should be only one label in an image, where a voxel value can be either zero (background) or a value between zero and one  $[0,1]$  that denotes the fuzzy membership or the probability that the corresponding voxel belongs to the label.

## Syntax

```
USAGE: EvaluateSegmentation truthURL segmentPath [-thd threshold]
[-xml xmlpath] [-use all|DICE,JACRD, ...]
```

where:

`truthURL`: `cURL+path to truth image+saKey`

`segmentPath`: `path to image being evaluated`

`-th threshold`: `before evaluation convert fuzzy images to binary using the given threshold`

`-xml xmlpath`: `path to xml file where results should be saved`

`-help`: `more information`

`-use metriclist`: `this option can be used to specify which metrics should be used.`

Metriclist consists of the codes of the desired metrics separated by commas. For those metrics that accept parameters, it is possible to pass these parameters by writing them between two @ characters, e.g. -use MUTINF,FMEASR@0.5@. This option tells the tool to calculate the mutual information and the F-Measure at beta=0.5. Possible codes for metriclist are:

- all: calculate all available metrics (default)
- DICE: calculate Dice Coefficient
- JACRD: calculate Jaccard Coefficient
- GCOERR: calculate Global Consistency Error
- VOLSMTY: calculate Volumetric Similarity Coefficient
- KAPPA: calculate Cohen Kappa
- AUC: calculate Area under ROC Curve (one system state)
- RNDIND: calculate Rand Index
- ADJRIND: calculate Adjusted Rand Index
- ICCORR: calculate Interclass Correlation
- MUTINF: calculate Mutual Information
- FALLOUT: calculate Fallout (false positive rate)
- COEFVAR: calculate Coefficient of Variation
- AVGDIST: calculate Average Distance
- HDRFDST: calculate Hausdorff Distance HDRFDST@0.95@ means use 0.95 quantile

to

- avoid outliers. Default is quantile of 1 which means exact Hausdorff distance
- VARINFO: calculate Variation of Information
- PROBDST: calculate Probabilistic Distance
- MAHLNBS: calculate Mahanabolis Distance
- SNSVTY: calculate Sensitivity (Recall, true positive rate)
- SPCFTY: calculate Specificity (true negative rate)
- PRCISON: calculate Precision
- FMEASR: calculate F-Measure FMEASR@0.5@ means use 0.5 as a value for beta in the F-Measure
- ACURCY: calculate Accuracy

## Examples

**Example 1:** EvaluateSegmentation truth.nii segment.nii -use RNDIND,HDRFDST@0.96@,FMEASR@0.5@ -xml result.xml

This example shows how to compare two NIFTI images providing the Rand Index, Hausdorff distance and the F-Measure and save the results in result.xml.

The values between two @ symbols are parameters to the specific measures in this case the quantile value used with Hausdorff distance to avoid outliers. The second value (0.5) is the beta value used with the F-Measure.

**Example 2:** EvaluateSegmentation truth.nii segment.nii -use all -th 0.5

This example compares two images using all available metrics. Before comparing the images, they are converted to binary images using a threshold of 0.5, that is voxels with values in  $[0,0.5)$  are considered as background and those with values in  $[0.5,1]$  are assigned the label with a membership of 1.