SEVENTH FRAMEWORK PROGRAMME OF THE EUROPEAN UNION

Support for training and career development of researchers
(Marie Curie)
Networks for Initial Training (ITN)

FP7- PEOPLE-2011-ITN

Evaluation protocols

Work-package: 1
Deliverable: D1.2
Output of task: T1.2
Date: 20-12-2013

Delivered by:
AQUILAB (Partner 1)

Project Number 290148

Software for the Use of Multi-Modality images
in External Radiotherapy

Sommaire

1 Preface

The radiotherapy treatment workflow is extremely complex, with a large number of steps involved. Many of these steps are made easier using specialized software tools which can partially automate the processes, or provide the user with additional perspectives on the data to augment their decision making abilities.

Before being integrated into the SUMMER software prototype, each software tool must be evaluated individually, to ensure its quality. Software evaluation can be technical or functional. In this document, we provide tentative evaluation protocols that have been designed to test each software tool. For each protocol, information related to the experiments and metrics used to evaluate the software too are provided, as well as details about the material (i.e. datasets and observers) needed to perform the experiments.

As a conclusion, an overview of the needs in terms of material (e.g. datasets, reference standard building and observers) gathering, distribution and sharing is given.

2 Technical evaluation

According to the deliverable D1.1. 'Clinical application definition', two clinical applications are primarily targeted in the SUMMER project: brain and lung cases. Software tools for brain and lung image processing, registration and visualization are therefore developed by SUMMER researchers, and will be tested, according to the following (tentative) evaluation protocols.

2.1 Brain case

2.1.1 Quantification of MR spectroscopy datasets

Within the workpackage 5 'Multi-modality image processing', the researchers from ICR (partner 4) shall develop and test new algorithm to contour target volume on Magnetic resonance spectroscopy (MRS) images.

MRS is a non-invasive analytical technique that has been used to study metabolic changes in brain tumors. More specifically, it uses hydrogen protons (1H) signals to determine the relative concentrations of target brain metabolites. Each metabolite has a different peak in the spectrum, which appears at a known frequency.

The quality of spectra across an MRSI data set can vary greatly; it is thus important to automatically provide uncertainty estimates for the quantitative results. This need is further increased for clinical MRSI applications where a diagnosis may be based upon the results. Here, different techniques to characterize/evaluate the accuracies of estimated MRSI parameters are presented.

2.1.1.1 Materials

The following set of data will be used for evaluation:

- **Healthy volunteers datasets**
  Based on previously published literature [Safriel2005, Baker2008], 1H MRS performed on healthy volunteers is considered as ideal for comparing and validating novel spectrum quantification methods.
  The following datasets will be acquired on healthy volunteers:
  - 1 data set Siemens (1.5T, long TE=135 ms)
  - 1 data set Siemens (3T, short TE=30 ms)
  - 1 data set Philips (3T, short TE=135 ms)
  - 1 data set GE (1.5T, long TE=135 ms)

- **Synthetic datasets**
  In vivo data on healthy volunteers has the disadvantage that the true metabolite concentrations are unknown, making it inappropriate for measuring accuracy.
  On the other hand, phantoms containing solutions of metabolites at known concentration can be used to measure accuracy. However, such phantoms are not easily accessible and results can be trivial, as important baseline effects caused by macromolecules are not models within a phantom. In addition, metabolites in solution can chemically degrade over time, making results difficult to interpret.

Synthetic datasets, created by Monte-Carlo simulation method, is a good alternative to phantom studies as true metabolite quantities are known to numerical accuracy. In addition, other spectral properties such as noise, line shape, and baseline effects can be easily modeled and controlled. Thus, a minimum of 100 realizations of brain MRSI synthetic data will be generated, divided as follows to be representative of clinical datasets variability:

- Set1: linear combination of simulated 6 metabolite and 2 lipide profiles; no disturbance components such as baseline, noise or water are added to the simulated signals.
  Metabolite profiles: myo-inositol (Myo), creatine (Cr), phosphocholine (PCh), glutamate (Glu), NAA, lactate (Lac)
  Lipid profiles: at 1.3 ppm (Lip1) and lipid at 0.9 ppm (Lip2).
- Set 2=Set 1 with water resonance at 4.7 ppm
- Set 3=Set 1 with high SNR
- Set 4=Set 1 with low SNR
- Set 5=Set 1 with baseline distortion
- Set 6=Set 1 with water, baseline and high noise

- **Clinical datasets**
  Usually, it is pathological brain tissues (i.e. non-healthy) which are of interest. This data type is generally of poorer quality and can often contain high levels of lipids and other molecules not seen in healthy brain tissue. It is therefore important to test any new method on a large set of clinical data before it can be regarded as robust to noisy, abnormal, and artifactual spectra.
  To this purpose, we will perform clinical study on:
  - 60 data sets Siemens (1.5T, long TE=135 ms)
  - 1 data set Philips (2D, 3T, long TE=135 ms)

2.1.1.2 Evaluation design

In order to test the robustness and accuracy of a quantification scheme we propose the following set of experiments:

- **Experiment 1: accuracy on parameter estimation**
  This experiment will be performed using the synthetic datasets of Set1.
  For each metabolite in the simulated signal, the true amplitudes will be compared with the estimated amplitudes by means of a performance measure described in next section.
- **Experiment 2: influence of water, baseline and noise on the estimated amplitudes**
  This experiment will be performed using the synthetic datasets of Set2 to Set6.
  For each metabolite in the simulated signal, the true amplitudes will be compared with the estimated amplitudes by means of a performance measure described in next section.
- **Experiment 3: In vivo MRSI**
  This experiment will be performed using the healthy volunteers and clinical datasets.
  The goal of this experiment is to show that results are in accordance with the literature, for the specific acquisition procedure and settings.

2.1.1.3 Evaluation metrics

The following quantitative evaluation metrics will be computed:

- For **experiments 1&2** performed on synthetic datasets, for which reference standard is available
  - **Root mean square error (RMSE)**

$$RMSE = \sqrt{\frac{\sum_{l=1}^{L} (y - \hat{y}_l)^2}{L}}$$

A value closer to 0 indicates a fit that is more useful for prediction.

  - **Relative root mean square error (RRMSE)**

$$RRMSE = 100\sqrt{\frac{1}{L}\sum_{l=1}^{L} \frac{(y - \hat{y}_l)^2}{y^2}}$$

- **Performance measure (PM)**
  For each metabolite *k* in the simulated signal *l*, the true parameter $y_{k,l}$ can be compared with the estimated values $\hat{y}_{k,l}$ by means of a performance measure defined as:

$$PM_k = 100\sqrt{\frac{\sum_{l=1}^{L}(y_{k,l}-\hat{y}_{k,l})^2}{\sum_{l=1}^{L}y_{k,l}^2}}$$

  where L is the number of simulated signals. A low PM reflects a high performance, and is a percentage measure of the difference between estimated and true amplitudes.

- For **experiment 3** on clinical datasets, for which no reference standard is available

  - **Measure of the fit quality Q**
    The fit quality Q is defined as the standard deviation of the frequency domain residual between 0.2 and 4.0 ppm, divided by the standard deviation of the spectral noise. This definition is proposed in [Wilson2011] and is similar to that of [Slotboom2009].

    The fit quality Q has the attribute that it will be:
    - less than unity where over-fitting has occurred
    - equal to unity where the fit is perfect
    - greater than unity when the signal has not been completely modeled.

    Q cannot be used to identify baseline problems and has the counter intuitive property that adding random noise to a fit will improve the fit quality (unless quality = 0). However, it is useful for assessing fit quality for large numbers of spectra, where manual inspection is not feasible.

  - **Random visual inspection**
    A random number (e.g. 100) of spectra will be selected; each fit will visually be inspected by a specialist to identify any fitting problems that may not be possible to detect using Q.

  - **Determination of uncertainties**
    Characterizing the accuracies of calculated parameter estimates is an important part of quantitative analysis, and reliable uncertainty calculation is mandatory for assessing or comparing individual results.
    - **Cramer-Rao (CR) bounds**
      which represent lower bounds on the variance of the fitted parameters [Golding1998, Cavassila2000, Cavassila2001, Ratiney2004, Jiru2006, Espinasse2012]
    - **Confidence interval estimates**
      which represent intervals about the estimated parameter values that can be expected to contain the actual value with a specified degree of confidence [Young2000].
    - **measurement of the residual (Chi^2)**
      determined as the summed squared difference between the data and the spectral model generated from the fit result.

2.1.1.4 References

S. Cavassila *et al.*, "Cramer-Rao bound expressions for parametric estimation of overlapping peaks: influence of prior knowledge", J Magn Reso, 2000
S. Cavassila *et al.*, "Cramer-Rao bounds: an evaluation tool for quantification", NMR in Biomedicine, 2001
E.H. Baker *et al.*, "Regional apparent metabolite concentrations in young adult brain measured by (1)H MR spectroscopy at 3 Tesla", J Magn Reson Imaging, 2008.
T. Espinasse and P. Rochet, "A Cramér-Rao inequality for non differentiable models", Comptes Rendus Mathematique, 2012
E.M. Golding *et al.*, "Mathematical assessment of the precision of parameters in measuring resonance spectra", J Magn Reson 1998
F. Jiru *et al.*, "Error images for spectroscopic imaging", ESMRMB 2006

H. Ratiney *et al.*, "Extended Cramér-Rao lower bounds: background accommodation", Proc Intl Soc Magn Reson Med 2004.
Y. Safriel *et al.*, "Reference Values for Long Echo Time MR Spectroscopy in Healthy Adults", AJNR, 2005.
J. Slotboom *et al.*, "Reliability testing of in vivo magnetic resonance spectroscopy (MRS) signals and signal artifact reduction by order statistic filtering". Meas Sci Technol 2009
M. Wilson *et al.*, "A Constrained Least-Squares Approach to the Automated Quantitation of in vivo 1H Magnetic Resonance Spectroscopy Data", Magnetic Resonance in Medicine, 2011.
K. Young *et al.*, "Confidence images for MR spectroscopic imaging", Magnetic Resonance in Medicine, 2000

2.1.2 Functional MRI analysis

Within the workpackage 5 'Multi-modality image processing', the researchers from FSL (partner 5) shall develop and test new algorithm to detect regions at risk in the brain using functional magnetic resonance imaging (fMRI).

fMRI is a functional neuroimaging procedure that measures brain activity by detecting associated changes in blood flow.

Task fMRI is used to identify regions linked to critical functions, such as speaking, moving, sensing, or planning; such information are useful to plan for brain surgery and/or radiation therapy (RT). The goal of task fMRI data analysis is to detect correlations between brain activation areas and tasks that the subject is asked to perform during the scan.

Besides, it is now recognized that brain functions are distributed across networks in the brain and deficits occurring are not necessarily caused by lesions in eloquent areas, but instead, are due to the disturbance in the connections between these network regions. These disturbances in the functional connections can be studied with resting state fMRI (R-fMRI). R-fMRI studies offer a way for localizing functional networks in the absence of a task.

Here, information derived from fMRI and/or R-fMRI analysis shall be used to define some regions at risks (e.g. regions corresponding to critical functions), which shall serve to better define a CTV.

2.1.2.1 Materials

The following set of data will be used for evaluation:

- **Healthy volunteers datasets**
  Minimum of 10 datasets with
  - task fMRI (various task)
  - resting-state fMRI
- **Clinical datasets**
  Minimum of 10 datasets with
  - task fMRI (various task)
  - resting-state fMRI

2.1.2.2 Evaluation design

The following experiments will be carried out:

- **Experiment 1: task paradigm design validation**
  This experiment will be performed using the healthy volunteer datasets with task fMRI.
  For each task in the paradigm design, the activation area will be checked by experts, and compared with reported task activation areas published in the literature.

- **Experiment 2: task fMRI analysis**
  This experiment will be performed using the healthy volunteer and clinical datasets with task fMRI. Activation area will be identified using two methods: FEAT, and independent component analysis (ICA). Results derived from the two methods will be compared in order to determine the best method for analysis of brain tumor patients with fMRI.

- **Experiment 3: resting-state fMRI analysis**
  This experiment will be performed using the healthy volunteer and clinical datasets with both task and resting-state fMRI.

Functional networks will be identified using two methods: independent component analysis (ICA), and seed-based analysis. Results derived from the two methods will be compared in order to determine the best method for analysis of brain tumor patients.
Furthermore, resting state networks will be compared to task regions intra-subject, to study differences in the activation areas.
Last, Identified networks in healthy volunteers will be compared to reported functional networks in the literature and changes of these connections in tumor patients will be studied.

- **Experiment 4: fMRI based clinical target volume delineation**
  This experiment will be performed using the clinical datasets with both task and resting-state fMRI. The objective is to compare the target volume segmentation defined regardless of fMRI information with the target volume segmentation defined considering the fMRI information, and its consequences on dose planning.

2.1.2.3 Evaluation metrics

The following quantitative evaluation metrics will be computed:

- **For experiment 1**
  Qualitative evaluation will be performed, by mean of visual inspection of activation areas by a consensus of 3 experts.
  Task activation areas in healthy volunteers after successful registration to MNI-standard template can be compared to be in accordance with published literature, as reported for example in with Brede database, accessible at http://hendrix.imm.dtu.dk/services/brededatabase/.

- **For experiment 2**
  Different post-processing analysis steps (motion correction and co-registration) will be evaluated by computing the **overlap**, before and after the corrections intra-subject.
  Accuracy of task activation will be evaluated using two additional correction techniques: B0-fieldmapping [Wang2013] and physiological signal correction; original results will be compared with these corrected activation areas.

- **For experiment 3**
  Different post-processing analysis steps (motion correction and co-registration) will be evaluated by computing the overlap before and after the corrections intra-subject. Accuracy of network locations will be evaluated using B0-fieldmap correction technique [Wang2013] and comparison to original results will be done.

- **For experiment 4**
  For this experiment, metrics remains to be further discussed with experts. Tentatively:
  At least 2 experts should make dose planning before knowledge on activation locations and afterwards including fMRI information. Differences between these plans should be then compared.
  The following metrics could potentially be used, according to literature, to compare fMRI and non-fMRI based target volume delineation:
  - Iso-dose distribution
  - Dose Volume Histogram (DVH) analysis and Normal Tissue Complication Probability (NTCP) model
  - Volume of PTV receiving more than 80% of the prescribed dose

  In order to investigate the following points:
  - Is the dose given to the tumor still adequate?
  - What is the impact on the dose delivered to OARs?

2.1.2.4 References

H. Liu *et al.*, "Task-free presurgical mapping using functional magnetic resonance imaging intrinsic activity: Laboratory investigation", J Neurosurg 111(4):746 – 754, 2009.
R. Garcia-Alvarez et al., "Repeatability of Functional MRI for conformal Avoidance Radiotherapy Planning", Journal of Magnetic Resonance Imaging. 23: 108-114, 2006.
S. Kokkonen et al., "Preoperative localization of the sensorimotor area using independent component analysis of resting-state fMRI", Magnetic Resonance Imaging, 2009.
A. Kovács *et al.*, "Integrating functional MRI information into conventional 3D radiotherapy planning of CNS tumors. Is

it worth it?", J Neurooncol, 105(3):629-37, 2011.
M. Otten *et al.*, "Motor deficits correlate with resting state motor network connectivity in patients with brain tumors", Brain 135; 1017-1026, 2012.
D. Zhang *et al.*, "Preoperative Sensorimotor Mapping in Brain Tumor Patients using Spontaneous Fluctuations in Neuronal Activity Imaged with fMRI: Initial Experience", Neurosurgery. 65(6):226-236, 2009.
H Wang, J Balter and Y Cao. "Patient-induced susceptibility effect on geometric distortion of clinical brain MRI for radiation treatment planning on a 3T scanner", Phys. Med. Biol. 58; 465–477, 2013.

2.1.3 Diffusion MRI analysis

Within the workpackage 5 'Multi-modality image processing', the researchers from FSL (partner 5) shall develop and test new algorithm to detect regions at risk in the brain, using diffusion magnetic resonance imaging, also referred to as diffusion tensor imaging (DTI).
DTI is a magnetic resonance imaging technique that enables the measurement of the restricted diffusion of water in tissue, in order to produce neural tract images. DTI is primarily used to perform tractography within white matter. Localizing critical tracts in relation to a tumor (infiltration, deflection) is useful, during surgical planning for instance, to provide information about the proximity and relative position of critical tracts and the tumor. Similarly, this information could be of interest during RT planning, in order to better define target volume and deliver less dose to critical tracks.
Here, we develop and evaluate a new DTI tractography algorithm.
According to literature [Tomoyuki2012, Peroni2012], the fibers of interest in image guided RT are:

- Corticospinal tracts (one of the pyramidal tracts), related to motor activity
- arcuate fasciculus, related to language
- optic radiation, related to visual activity.

2.1.3.1 Materials

The following set of data will be used for evaluation:

- **Phantom dataset**
  The evaluation will be perform using the *evaluation framework* set up for the fiber tracking challenge 'FiberCup', which was launched during a workshop that was held during the 12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2009. The goal of this challenge is to compare the performances of the state-of-the-art models and fiber tracking techniques on a common MR phantom dataset.
  The diffusion phantom has been designed following a 3 steps strategy:
  - designing a specific fibre configuration
  - designing a media to hold the fibres strongly tightened and a container to receive the phantom
  - choosing an adequate recipe to fill the container with a MRI compatible solution.

  Details about the mechanical conceptions are available here.

  Diffusion-weighted MR data of the phantoms were acquired on the 3T Tim Trio MRI systems, equipped with a whole body gradient coil (40 mT/m, 200 T/m/s), and using a 12-channel receiver head coil, in combination with the whole body transmitter coil of the MRI system. Two datasets were acquired at two different spatial resolutions:
  - 3mm isotropic resolution (image size: 64x64x3) with 3 b-values (650, 1500, 2000 s/mm2) made of two repetitions.
  - 6mm isotropic resolution (image size: 64x64x1) with 3 b-values (650, 1500, 2650 s/mm2) made of only one repetition.

  The diffusion sensitization was applied along a set of 64 orientations, uniformly distributed over the sphere.
  The phantom dataset is publicly available on the challenge website.

- **Clinical datasets from a Grand Challenge in Medical Imaging**
  The evaluation will be perform using the evaluation framework set up for the DTI Tractography Challenge, which is being held each year since 2011 as a workshop held at MICCAI conferences.
  The datasets for the challenge will be given at the MICCAI'2014 conference. According to previous years, it can be expected that 5 cases will be made available, with its ground truth, i.e. fiber tracks.

- **Clinical datasets**
  The following DTI shall be acquired:
  - 5 healthy subjects
  - 15 tumor case patients, with pre- and post- operative DTI
    - 5 patients with tumor near motor area
    - 5 patients with tumor near language area
    - 5 patients with tumor near optical area

  with 64 orientations as diffusion sensitization.
  A ground truth, i.e. fiber tracks, built by a consensus of neuroradiologists experts, must be available.

2.1.3.2 Evaluation design

The following experiments will be carried out:

- **Experiment 1: 'Fiber cup'**

This experiment will be carried out on the phantom dataset.
The objectives of the challenge are: 1) to provide an MR phantom containing a plethora of crossing, kissing, splitting and bending fiber configurations, and 2) to offer a set of quantitative criteria for algorithm performance evaluation with online results. We will be run our algorithms on the phantom dataset and return our results along with a 2-page paper summarizing our method for quantitative evaluation.

- **Experiment 2: DTI Tractography Challenge 2014**

The goal of this challenge is to compare the performances of fiber tracking algorithms in the reconstruction of peritumoral anatomy and corticospinal tracts trajectory, on pre-operative and post-operative diffusion MR data from patients presenting with a tumor in or near the motor system.

- **Experiment 3:**

This experiment will be performed using the clinical datasets.
The goal of this experiment is to compare the performances of our fiber tracking algorithms in the reconstruction of 1) corticospinal tracts, 2) arcuate fasciculus tracts, and 3) optic radiation tracts trajectory, on pre-operative and post-operative diffusion MR data, from patients presenting with a tumor in or near the motor, the language or the visual system respectively.

2.1.3.3 Evaluation metrics

- **Experiment 1:**
  Given a certain sampling of the fiber tracts, the following metrics will be used for evaluation:
  - spatial metric
  - tangent metric
  - curve metric

  for which the following quantities will be reported:
  - mean
  - stdev
  - min
  - max

  Performance of our algorithm will be compared to the phantom ground truth.

- **Experiments 2&3 :**
  - **Quantitative evaluation**

    The quantitative evaluation of tract reconstruction in each hemisphere, for patients and controls, will be performed using the following five metrics:
    - Dice's coefficient for volumetric overlap
    - point distance between bundles (mean, stdev, min, max)
    - fiber profiles on Fractional Anisotropy (FA)
    - fiber profiles of Mean Diffusivity (MD)

- STAPLE sensitivity and specificity score

- **Qualitative evaluation**

Here, a consensus of min. 3 observers, i.e. experts in neuroradiology, will be needed.

The qualitative assessment of **tract reconstruction** in each hemisphere, for patients and controls, will be based on:

- anatomical correctness of the tract
- presence of false-positive tracts
- presence of false-negative tracts

The qualitative assessment of **the depiction of the spatial relation** between the tumor and the tract, will be based on:

- correct depiction of the distance between the tract and the lesion
- demonstration of tract displacement
- demonstration of tumor infiltration

2.1.3.4 References

P. Fillard et al., "Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom", NeuroImage, 56(1):220-234, 2011.
M. Peroni et al., "VMAT inverse planning including DTI tractography fiber bundles as organs at risk: a feasibility study", Workshop on Image-Guidance and Multimodal Dose Planning in Radiation Therapy, MICCAI, Nice, France, 2012.
C. Poupon *et al.*, "New diffusion phantoms dedicated to the study and validation of high-angular-resolution diffusion imaging (HARDI) models", Magn Reson Med.,60(6):1276-83, 2008.
K. Tomoyuki et al., "Outcomes of diffusion tensor tractography integrated stereotactic radiosurgery", International Journal of Radiation Oncology Biology Physics, 82(2):799-802, 2012.

2.1.4 Rigid registration: application to brain case

Within the workpackage 3 'Multi-modality image registration', the researchers from MUW (partner 2) shall develop and test new algorithm to rigidly register multi-modality images for brain case.

2.1.4.1 Materials

The following set of data will be used for evaluation:

- **Pig datasets**
  The dataset is from a pig, called Herbert, for which ground truth is available by mean of available fiducial markers.
  MR-T1, T2, EPI, PD, CT, CBCT and calibrated X-Ray images are publicly available here.
  Gold standard using fiducial markers is escribed in [Paviro and Birkfellner 2011]

- **Clinical datasets**
  A min. of 15 glioma patient datasets, with following image modalities:
  - pCT
  - MR T1
  - MR T2
  - MR FLAIR
  - fMRI and/or MRS

  shall be used, and follow-up studies shall be available.

  For quantitative evaluation, **reference standard** is required.
  In default of a ground truth transformation, many investigators focus on a limited set of anatomical features for which homologies can be established between the image pairs.
  - Based on **anatomical landmarks**, annotated on both registered image independently.
    A general validation technique is based on identification of the anatomical landmarks in clinical data.
    The landmark definition (approx. 50) will be further discussed with the **two observers** (i.e. experts in neuroradiology). Landmark points will be spread over the entire volume.

2.1.4.2 Evaluation design

The following experiments will be carried out:

- **Experiment 1:** on the pig dataset.
  The accuracy of our rigid registration method will be primarily investigated on a pig dataset, for which ground truth is available.

- **Experiment 2:** on clinical datasets
  The accuracy of our rigid registration method will be investigated, on representative clinical datasets, for which no ground truth is available. Besides, a cross-validation with rigid registration methods of commercially available workstations (e.g. Analyzer, BrainLab and/or Eclipse) will be performed. Cross-validation will be done in collaboration with TU Delft (partner 7), as part of a user-testing.

2.1.4.3 Evaluation metrics

The following evaluation metrics will be used to assess rigid registration accuracy:

- for **experiment 1**, on pig dataset
  The pair**-wise point's distance error** will be computed between fiducial markers present on original and registered images.
  The following quantities will be computed:
  - mean
  - stdev
  - min
  - max

- for **experiment 2**, on clinical datasets
  - **qualitative metrics**
    Visual inspection will be performed.
    At least 2 observers will be needed.
    The results will be analyzed by means of:
    - Kappa statistic
    - agreement rate
    - area of receiver-operating-characteristic (ROC) curves

  - **quantitative metrics**
    At least 2 observers for defining anatomical landmarks on brain datasets will be needed, to check intra-/inter-observer variability.

    - **Landmark-based**
      In default of a ground truth transformation, many investigators focus on a limited set of anatomical features for which homologies can be established between the image pairs. A general validation technique is based on identification of the anatomical landmarks in clinical data. The location of the landmark points is labeled and the **pair-wise point's distance error** of corresponding points is calculated. However, this metric depends on the accuracy of identification of homologous landmarks in two different images by the experts.

      The following quantities will be computed:
      - mean
      - stdev
      - min
      - max

2.1.4.4 References

S.A. Paviro and W. Birkfellner et al. "Validation for 2D/3D registration. I: A new gold standard data set", Med Phys.; 38(3):1481-90, 2011.

W. C. Lavely , C. Scarfone , H. Cevikalp , R. Li , D. W. Byrne , A. J. Cmelak , B. Dawant , R. R. Price , D. E. Hallahan and J. M. Fitzpatrick "Phantom validation of coregistration of PET and CT for image-guided radiotherapy", Med. Phys., vol. 31, pp.1083 -1092, 2004.
D. J. Hawkes "Algorithms for radiological image registration and their clinical application", J. Anat., vol. 193, pp.347 -61, 1998.
J. M. Fitzpatrick , D. L. Hill , Y. Shyr , J. West , C. Studholme and C. R. Maurer, Jr. "Visual assessment of the accuracy of retrospective registration of MR and CT images of the brain", IEEE Trans. Med. Imag., vol. 17, no. 4, pp.571 -85, 1998.
C. Studholme , D. L. Hill and D. J. Hawkes "Automated 3-D registration of MR and CT images of the head", Med. Image Anal., vol. 1, pp.163 -175, 1996.

2.2 Lung case

2.2.1 Segmentation of organs at risks on 3D CT images

Within the workpackage 5 'Multi-modality image processing, the researchers from AQUILAB (partner 1) shall develop and test new algorithm to segment the organs at risk (OAR) on 3D lung CT images.

2.2.1.1 Materials

The following set of data will be used for evaluation:

- **Clinical datasets**
  In order to demonstrate the robustness of our method, around **50 chest CT** datasets shall be included, in which a minimum of 3 OAR shall be delineated (i.e. heart, liver, aorta, lung,..)
  The CT scans may be acquired with different vendor machines and present different field of view.

2.2.1.2 Evaluation design

The following experiments will be carried out:

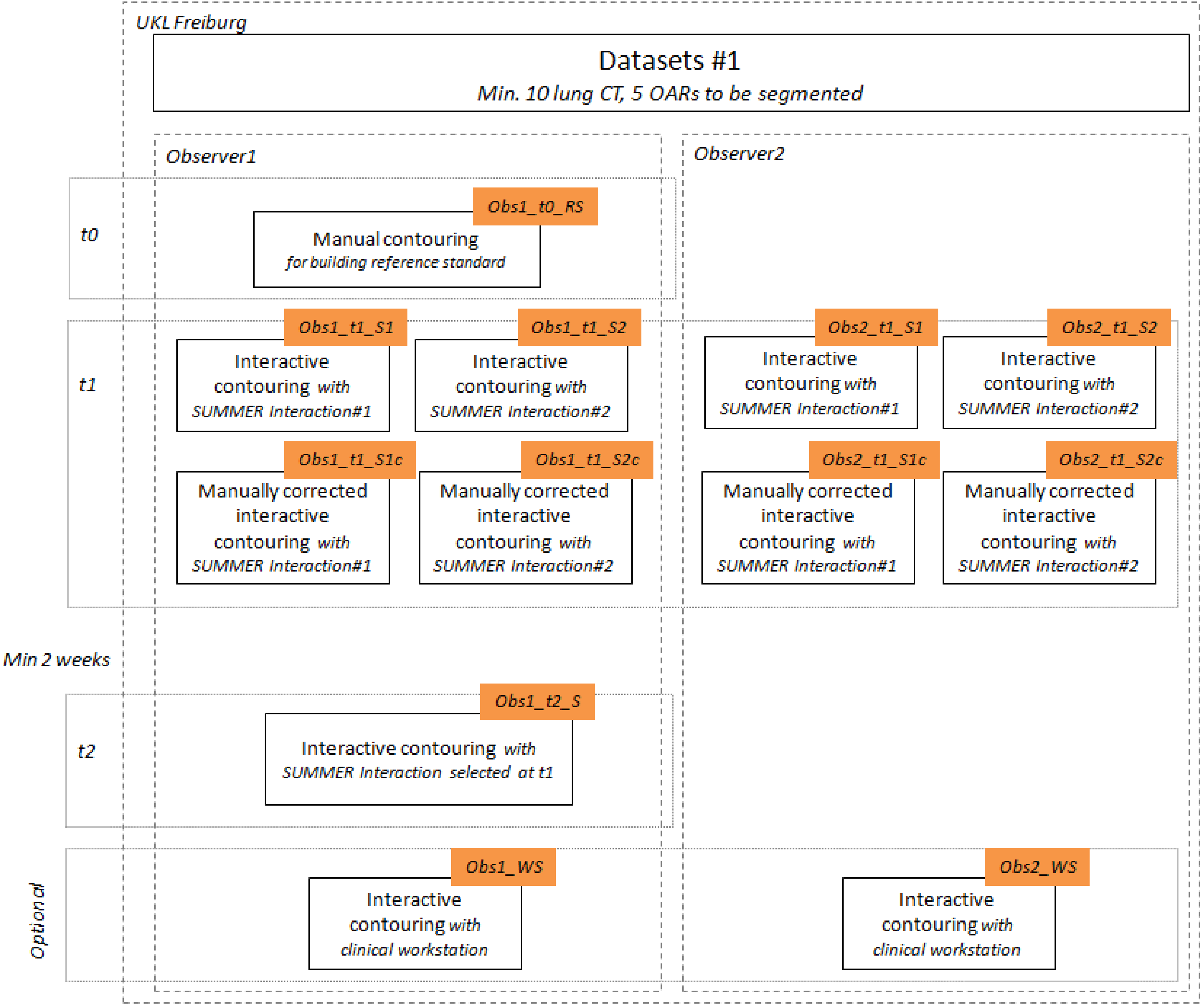
UKL Freiburg
Datasets #1
Min. 10 lung CT, 5 OARs to be segmented
Observer1
Observer2
t0
Obs1_t0_RS
Manual contouring
for building reference standard
t1
Obs1_t1_S1
Interactive contouring with SUMMER Interaction#1
Obs1_t1_S2
Interactive contouring with SUMMER Interaction#2
Obs2_t1_S1
Interactive contouring with SUMMER Interaction#1
Obs2_t1_S2
Interactive contouring with SUMMER Interaction#2
Obs1_t1_S1c
Manually corrected interactive contouring with SUMMER Interaction#1
Obs1_t1_S2c
Manually corrected interactive contouring with SUMMER Interaction#2
Obs2_t1_S1c
Manually corrected interactive contouring with SUMMER Interaction#1
Obs2_t1_S2c
Manually corrected interactive contouring with SUMMER Interaction#2
Min 2 weeks
t2
Obs1_t2_S
Interactive contouring with SUMMER Interaction selected at t1
Optional
Obs1_WS
Interactive contouring with clinical workstation
Obs2_WS
Interactive contouring with clinical workstation

Fig. 1 – OARs segmentation evaluation protocol – Experiment 1

- **Experiment 1:**
  This experiment will be performed on a subset (min. 10) of the clinical datasets.
  The goal of this experiment is 1) to determine which one of two user-interactions is the optimal and shall be used to provide input to our segmentation algorithm, 2) test the accuracy of the proposed method using the optimal user-interaction.

Performances of our minimally-interactive segmentation algorithm will be compared to 1) the reference standard, i.e. manual contouring by an expert, and 2) optionally, to state-of-the art algorithms available on commercial workstations. Besides, intra- and inter- observer variability will be computed.

- **Experiment 2:**
  This experiment will be performed on the clinical datasets (min. 50), and is an extension of Experiment 1.
  The goal of this experiment is to investigate the accuracy of the proposed method (i.e. using the optimal user-interaction selected in Experiment 1) on a larger set of datasets, representative of clinical practice.

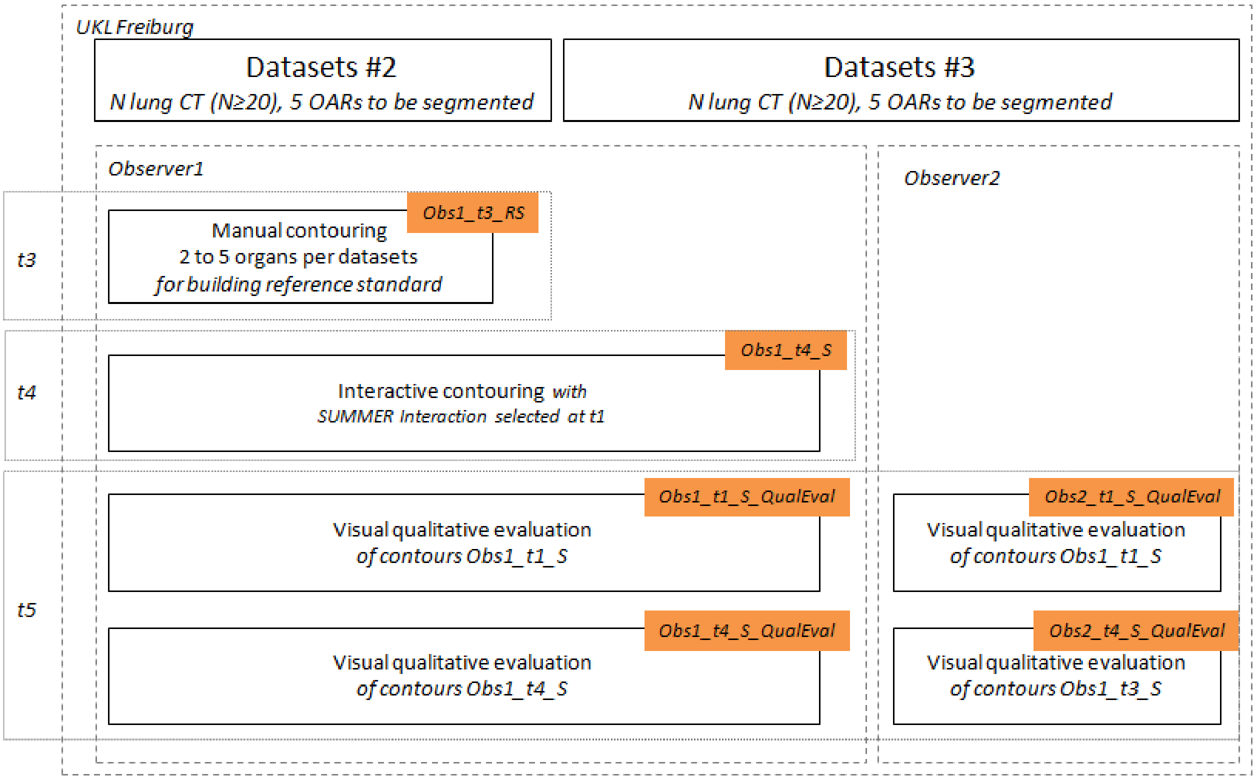
UKL Freiburg
Datasets #2
N lung CT (N≥20), 5 OARs to be segmented
Datasets #3
N lung CT (N≥20), 5 OARs to be segmented
Observer1
Observer2
t3
Obs1_t3_RS
Manual contouring
2 to 5 organs per datasets
for building reference standard
t4
Obs1_t4_S
Interactive contouring with
SUMMER Interaction selected at t1
t5
Obs1_t1_S_QualEval
Visual qualitative evaluation
of contours Obs1_t1_S
Obs1_t4_S_QualEval
Visual qualitative evaluation
of contours Obs1_t4_S
Obs2_t1_S_QualEval
Visual qualitative evaluation
of contours Obs1_t1_S
Obs2_t4_S_QualEval
Visual qualitative evaluation
of contours Obs1_t3_S

Fig. 2 – OARs segmentation evaluation protocol – Experiment 2

Performances of our minimally-interactive segmentation algorithm will be compared to 1) the reference standard, i.e. manual contouring by an expert, and 2) optionally, to state-of-the art algorithms available on commercial workstations.

2.2.1.3 Evaluation metrics

The following evaluation metrics will be used to assess OARs segmentation accuracy and robustness:

- **Quantitative metrics**
  Manual contour being tedious, sample data size will be discussed with the observers.
  - Shape: Dice similarity metric.
  - Volume: volume difference (mean+/- stdev)
    the difference between the automatic volume (Vauto) and the manual volume, i.e. reference volume (Vref), is calculated
  - Alignment: COG (mean+/-stdev)
    the three coordinates (x,y,z) of the centres of gravity (COG) of the automatic and reference structures were compared:
    Dx = |xauto - xref|; Dy = |yauto - yref|, Dz = |zauto - zref|.

- **Qualitative metrics**

- **Visual inspection**

  Visual inspection being faster and less tedious than manual contouring, a larger sample data size can be used

  The segmentation quality will be assessed visually by observers, using grade classification. The segmentation quality grade classification is based on regional segmentation accuracy, as proposed in [Abadi2009]. It ranges from 1 – very accurate - to 5 - fail -. This grade scale can be adapted to the OARs considered.

Grade	*Description*
1	Very accurate: Deviation up to 1mm
2	Most regions accurate: 1 or 2 regions may deviate up to 3mm
3	Most regions accurate: 1 region may deviate up to 1cm or more than 2 regions may deviate up to 3mm
4	A significant region (up to 50%) has not been segmented or has been incorrectly segmented
5	Segmentation failed

Table 1. Segmentation quality grades classification.

The results will be analyzed by means of:

- Kappa statistic
- agreement rate
- area of receiver-operating-characteristic (ROC) curves

- **User case study**

  For **experiment 1** exclusively, a user-case study may be performed, with the following study question: "Which user interaction is the most appropriate to segment OARs on lung CT images using the proposed minimally interactive approach", and study case "use of the SUMMER OAR segmentation prototype by two experts".

2.2.1.4 References

La Macchia et al., "Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer", Rad. Oncol., 7:160, 2012.
S. Abadi et al. "Feasibility of automatic assessment of four-chamber cardiac function with MDCT: initial clinical application and validation". Eur J of Radiol, 74 (1), PP. 175-181. 2009.
H.A. Kirisli et al., "Evaluation of a multi-atlas based method for segmentation of cardiac CTA data: a large-scale, multi-center and multi-vendor study", Medical Physics, 2010.

2.2.2 Segmentation of tumor target volume on 4D PET images

Within the workpackage 5 'Multi-modality image processing, the researchers from UKL (partner 3) shall develop and test new algorithm to segment the tumor gross target volume (GTV) on 4D PET/CT images.

Respiratory motion poses a challenge in nuclide imaging, where data often must be acquired over many respiratory cycles to obtain adequate statistics. Intra-scanning organ deformation results in lesion motion, thereby spreading the radiotracer activity over an increased volume, distorting apparent tumor shape and location, and reducing signal-to noise ratio (SNR).
4D PET/CT provides useful information primarily in all those clinical situations where a mobility of the target due to respiration is expected. In such cases, its use provides an objective measure of the motion. The difficulty in PET image segmentation is compounded by the low spatial resolution and high noise characteristics of PET images.
Several image segmentation approaches have been proposed and use in the clinical setting. Validation of accuracy (i.e. fidelity to the truth) and precision (i.e. reproducibility / repeatability) are a very crucial step for any clinical use of a computer algorithm.

2.2.2.1 Materials

The following sets of data will be used for evaluation:

- **Phantom datasets**
  The phantom that will be used for this study will be created following the criteria:
  - it covers, at least, the more significant anatomy in CT image
  - it has compartments that can be filled independently for, at least, the organs that present higher activity concentration in PET lung tumor image
  - it gives the possibility to study different target to background ratios (TBR)
  - it gives the possibility to study respiratory Motion.

  Phantom images will be acquired to create a database with variations in terms of:
  - Different values for organ activity distribution
  - Variations in the organ with highest activity concentration
    - Liver
    - heart
  - Variations in the relative activity between the tumor and the organ with highest activity concentration
    - Different TBR
    - Different tumor position
  - Different tumor size and shape
  - Different respiratory movement

- **Clinical datasets**
  Depending on the outcome of the initial study on phantoms, and advancement of the project, clinical evaluation on patient datasets could be considered. Further brainstorming about building ground truth for GTV delineation will then be performed.

2.2.2.2 Evaluation design

The following experiments will be performed to assess the accuracy and precision of the developed GTV segmentation method for 4D PET/CT images.

- **Experiment 1:** on phantom datasets
  Results obtained with the algorithm will be compared with GTV delineation ground truth (i.e. phantom specifications).

- **Experiment 2:** on clinical datasets
  Results obtained with the algorithm will be compared to ground truth built by experts.

2.2.2.3 Evaluation metrics

The following evaluation metrics will be used to assess the GTV segmentation method accuracy:

- **Analytical**
  To examine and asses the segmentation algorithm itself by analyzing its principles and properties. The following points will be evaluatied:
  - Processing strategy
  - Processing complexity
  - Memory consumption
  - Processing time

- **Empirical**
  - **Goodness methods**
    Some desirable properties are measured as “goodness” parameters.
    Here, the **intra-region uniformity**, as expressed in [Zhang96] and [Huang95] will be computed.

  - **Discrepancy methods**
    The methods in this group take into account the difference (measured by various discrepancy parameters) between the actually segmented and reference images, i.e. these methods try to determine how far the actually segmented image is from the reference

image (gold standard). Segmentation will be evaluated in terms of **accuracy**, to reflect the precision of the segmentation to the gold standard

- based on the feature values of segmented objects:
  - Relative Ultimate Measured Accuracy (RUMA)
  - Radial distance between contours: mean, stdev, max.
- Based on mis-segmented pixels
  - DICE similarity coefficient [Zou04]
  - Multiclasses [Zaidi10]
  - Probability error

2.2.2.4 References

R. Boellaard et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version1.0. Eur J Nucl Med Mol Imaging, 37(1):181–200, 2010.
Q. Huang et al. Quantitative methods of evaluating image segmentation. Proceedings, International Conference on, Volume: 3, 1995.
Zaidi et al. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging,37(1), 2010.
YJ. Zhang et al. A survey on evaluation methods for image segmentation. Patter Recognition, 29(8), 1996.
K. Zou et al. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. Acad Radiol. 11(2): 178–189., 2004

2.2.3 Deformable registration

Within the workpackage 3 'Multi-modality image registration', the researchers from MUW (partner 2) shall develop and test new algorithm to non-rigidly register multi-modality images for lung case, and particularly planning CT with treatment CBCT.

2.2.3.1 Materials

The following sets of data will be used for evaluation:

- **Clinical datasets**
  A minimum of 20 datasets with planning CT and treatment CBCT images will be used.
  The images can be acquired at different medical institutions, thus increasing the variety of the datasets.
  The field of view (FOV) of both modalities shall be similar (required for building reference standard).

  For quantitative evaluation, **reference standard** is required.
  In default of a ground truth transformation, many investigators focus on a limited set of anatomical features for which homologies can be established between the image pairs.
  - Based on **OARs contours**, annotated on both CT and CBCT independently.
    The contours will be obtained by fully manual annotation performed by **an observer**.
  - Based on **anatomical landmarks**, annotated on both CT and CBCT independently.
    A general validation technique is based on identification of the anatomical landmarks in clinical data.
    The landmark definition (approx. 50) will be further discussed with the **two observers** (i.e. experts experienced in lung CT/CBCT registration). Landmark points will be spread over the entire volume.

2.2.3.2 Evaluation design

This following experiment will be performed on the clinical datasets. Its goal is to investigate the performance of the deformable registration in terms of accuracy.
Two observers, i.e. experts in lung who are familiar with registration of pCT-CBCT, will be required to perform the suggested study presented in Figure 3.

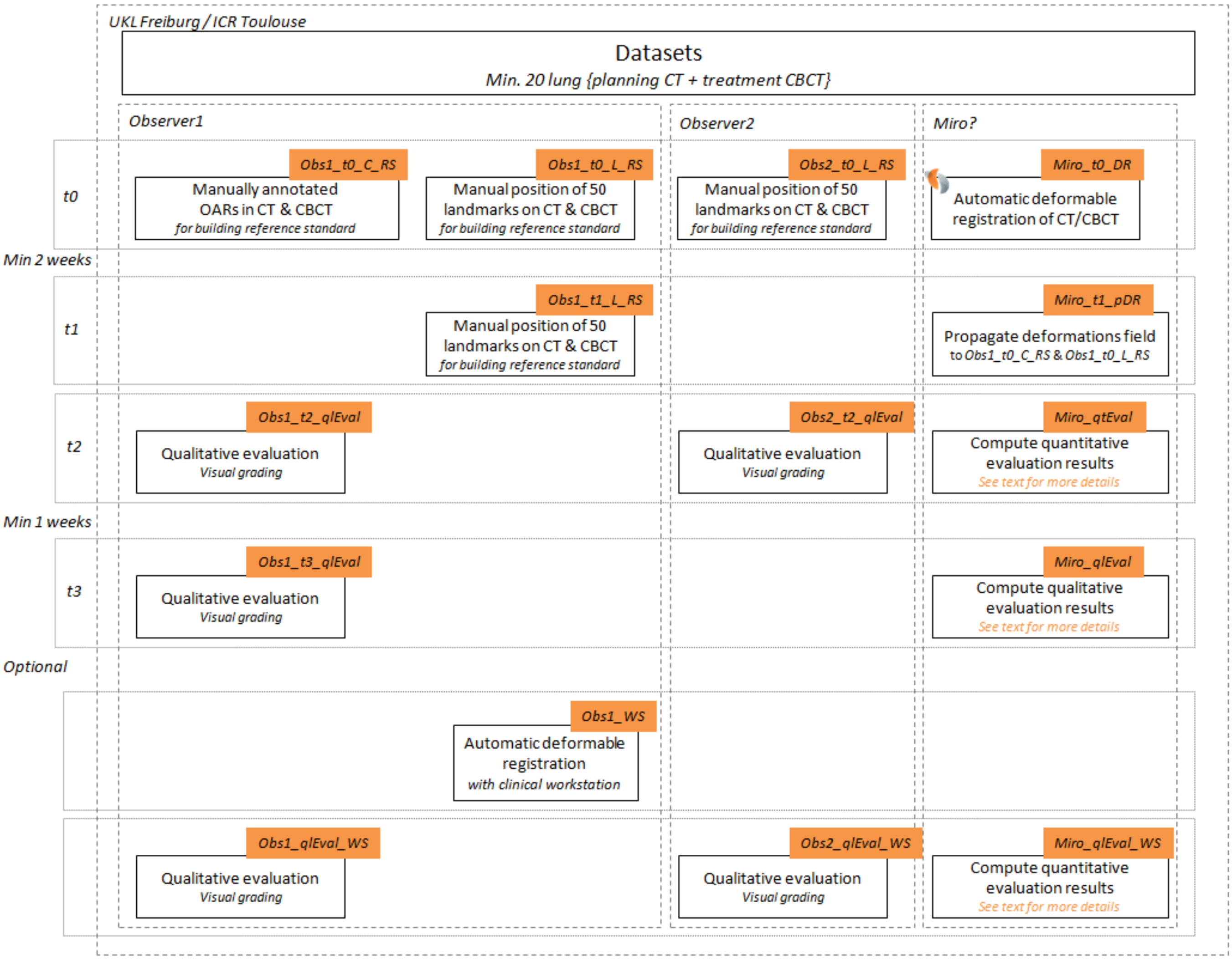
UKL Freiburg / ICR Toulouse
Datasets
Min. 20 lung {planning CT + treatment CBCT}
Observer1
Observer2
Miro?
t0
Obs1_t0_C_RS
Manually annotated OARs in CT & CBCT
for building reference standard
Obs1_t0_L_RS
Manual position of 50 landmarks on CT & CBCT
for building reference standard
Obs2_t0_L_RS
Manual position of 50 landmarks on CT & CBCT
for building reference standard
Miro_t0_DR
Automatic deformable registration of CT/CBCT
Min 2 weeks
t1
Obs1_t1_L_RS
Manual position of 50 landmarks on CT & CBCT
for building reference standard
Miro_t1_pDR
Propagate deformations field
to Obs1_t0_C_RS & Obs1_t0_L_RS
t2
Obs1_t2_qlEval
Qualitative evaluation
Visual grading
Obs2_t2_qlEval
Qualitative evaluation
Visual grading
Miro_qtEval
Compute quantitative evaluation results
See text for more details
Min 1 weeks
t3
Obs1_t3_qlEval
Qualitative evaluation
Visual grading
Miro_qlEval
Compute qualitative evaluation results
See text for more details
Optional
Obs1_WS
Automatic deformable registration
with clinical workstation
Obs1_qlEval_WS
Qualitative evaluation
Visual grading
Obs2_qlEval_WS
Qualitative evaluation
Visual grading
Miro_qlEval_WS
Compute quantitative evaluation results
See text for more details

Fig. 3 – Deformable registration of pCT-CBCT – Evaluation study design (tentative)

Performances of the deformable registration algorithm will be compared to 1) the reference standard , i.e. expert's manual contouring and landmark definition, and 2) optionally, to state-of-the art algorithms available on commercial workstations.

2.2.3.3 Evaluation metrics

The following evaluation metrics will be used to assess the non-rigid deformation accuracy:

- **qualitative metrics**

The segmentation quality will be assessed visually by observers, using:

- **Grade classification** The segmentation quality grade classification will be based on regional registration accuracy.

  For each OAR present in the common field of view, as well as tumor, the observer will assign one of the following grade: poor, satisfactory, or good. Specification for each grade will be further discussed with experts before performing evaluation.

  At least 2 observers (if done independently, and results of observers compared) or at least 3 observers (consensus; odd number required) will be needed.

  The results will be analyzed by means of:

  - Kappa statistic
  - agreement rate
  - area of receiver-operating-characteristic (ROC) curves

- **Binary test: best/worth**
  Given two (blinded) registrations, the observer must decide on which registrations is better/worth than the other one. This qualitative metric will be used for comparing registrations obtained using our method with those of commercially available workstations.

- **quantitative metrics**
  - **Landmark-based**
    The **pair-wise point's distance error** of corresponding landmark points are calculated. Note that this metric depends on the accuracy of identification of homologous landmarks in two different images by the experts.

    The following quantities will be computed:
    - mean
    - stdev
    - min
    - max

    At least 2 observers for defining anatomical landmarks on lung datasets will be needed, to check intra-/inter-observer variability.

    - **Contour-based**
      - Shape: Dice similarity metric.
      - Volume: volume difference (mean+/- stdev)
        the difference between the automatic volume (Vauto) and the manual volume, i.e. reference volume (Vref), is calculated
      - Alignment: COG (mean+/-stdev)
        the three coordinates (x,y,z) of the centres of gravity (COG) of the automatic and reference structures were compared: Dx = |xauto - xref|; Dy = |yauto - yref|, Dz = |zauto - zref|.

2.2.3.4 References

N. Hardcastle et al., "A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy", Radiation Oncology, 7:90, 2012.

Z. Wu et al., "Evaluation of deformable registration of patient lung 4DCT with sub-anatomical region segmentations", Med. Phys. 35, 775, 2008.

3 Functional evaluation

Within the workpackage 4 'Adaptive visualization', the researchers from VRVis (partner 6) shall develop a novel application specific visualization framework that substantially advances the state of the art in interactive visualization of multi-modal, spatiotemporal and multi-variate image in the context of radiotherapy planning.

The integrated visualization concepts will allow physicians to easily explore the available (registered) datasets and to combine relevant information of several images into one comprehensive and meaningful task specific image that optimizes the visualization of heterogeneous tumour tissue and allows easy exploration and comparison of follow-up studies.

3.1 Material

The following datasets are required:

- Phantom dataset
    - 4D lung PET/CT(DICOM)

- Patient datasets:
    - Brain case
        - MR T1, MR T2, MR Gadolinium, MR FLAIR, MRSI (DICOM)
        - fMRI and associated MRI, DTI (DICOM)
    - Lung case
        - 4D lung PET/CT (DICOM), with associated contours (RTSTRUCT) and dose (RTDose)

The above mentioned datasets are those required to perform the technical evaluation of each SUMMER processing tool; please refer to the respective evaluation protocols for estimation of number of datasets.

- Processed datasets
    - Brain case
        - Functional activations data
        - Tractography
        - OARs segmentation data (binary volumes, RTSTRUCT)
        - Rigid transformations (rotation matrix, translation vector) and associated deformation field (DVF volume)
        - MRSI metabolite maps
    - Lung case
        - GTV segmentation on 4D PET
        - OARs segmentation data (binary volumes, RTSTRUCT)

3.2 Evaluation

The multi-modality visualization will be evaluated qualitatively as well as quantitatively. The evaluation will be focused on use-cases developed with individual partners, i.e. according to each specific use-case we assess and quantify different variables.

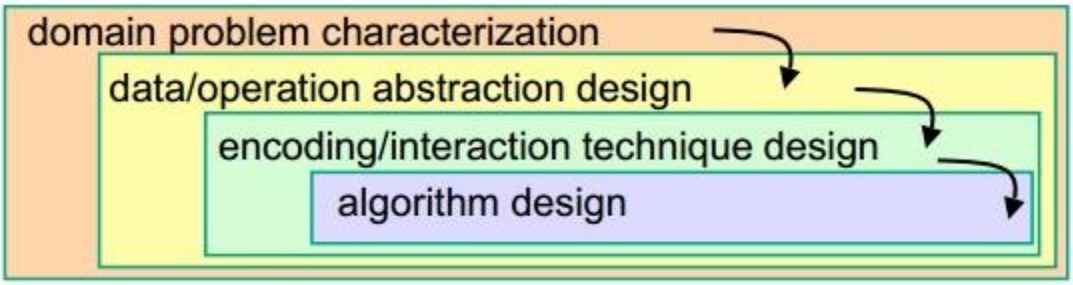
domain problem characterization
data/operation abstraction design
encoding/interaction technique design
algorithm design

Fig. 4 – Evaluation overview, based on [Munzner2009]

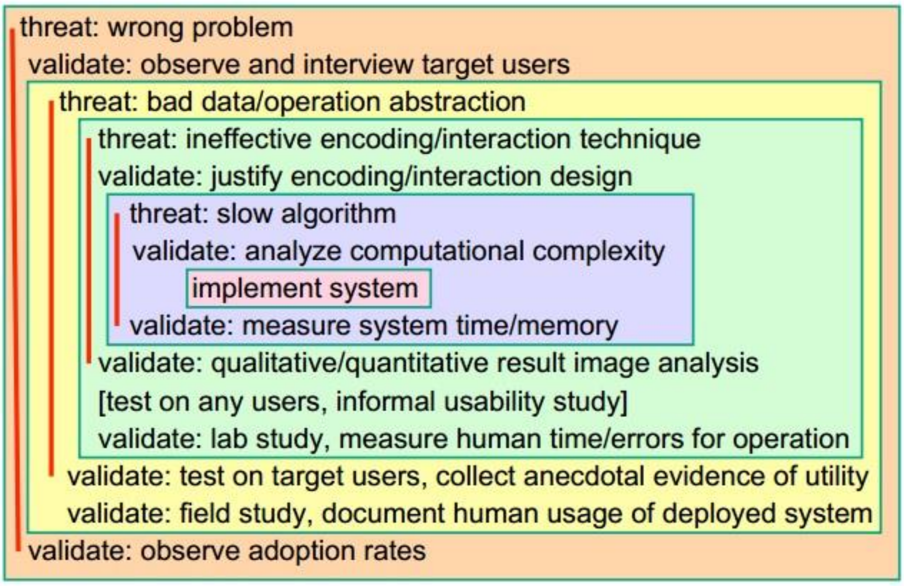
threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Fig. 5 – Threats and validation in the nested model of [Munzner2009]

According to [Munzner2009], it is of good practice to break down the evaluation of each use-case keeping in mind four key levels of abstraction:

1. **Domain Problem and Data Characterization**
   - know the domain in which we are working on (Medical field, Adaptative Radiotherapy)
   - understand the problem of the target audience (Medical staff, use visualization of complex medical information and uncertainty to better delineate the cancer target volumes)
   - the adoption rate of the system (renderer, ComVis-MITK)
2. **Operation and Data Type Abstraction**
   - generic tasks (display a set of values, render multi-modality images, align images according to registration, contouring, etc.)
   - transform raw data into data types (numerical tables, field of values, etc.)
   - the overall impact regarding the experience of the medical staff
3. **Visual Encoding and Interaction Design**
   - how the system should behave (is the target domain user capable of performing the operations? effectiveness of the contouring of the target volumes?)
   - quality of resulting images (the impact it may have on the contouring quality of the target volumes)
4. **Algorithm Design**
   - implementation of algorithms (fusion methods, calculating lung volume, interpolation, rendering 4D images, etc.)
   - technical evaluation of algorithms

For each these four main points, a set of threats and validations has to be set up and adjusted depending on the use-case at hand.

The following points will be evaluated:

- **Qualitatively**
  - Inquire and measure satisfaction of physicians with their contours by making use of the ComVis - MITK system. (Point 2, 3)
  - By making use of a multi-modality renderer, evaluate if ambiguities and uncertainties of images are reduced. (Point 3)
  - Evaluate if a group of doctors reach agreement about achieved contours of a specific case in less time. (Point 1,2)
  - Access quality and quantity of information displayed to the doctor associated to perception's evaluation. (Point 2,3)

  Prototype testing will be done by the developers as a proof of concept (i.e. refer to technical evaluation in Section 2). Intensive user-testing shall be done by TU Delft (partner7).

- **Quantitatively**
  - **Rendering speed** (Point 4)
    The rendering speed of the system will be evaluated in comparison to other systems (e.g. vtk, voreen, MeVisLab). The comparison system may vary depending on which system supports the use case at hand.
  - **Memory consumption** (Point 4)
  - **Correctness of visualization**
    - The rendering algorithms (e.g. raycasting) can be evaluated using similar methods as in [Etiene2013] (Point 4)
    - Measure the **time** needed to perform **validation** (i.e. visual inspection + optional manual refinement) using the new visualization techniques and compare to time required using other systems (depending on use-case, e.g. RT-View, Eclipse, BrainLab, Slicer) (Point 3)
    - Investigate whether the new multi-modal visualization techniques lead to **similar contours** done/approved by doctors, i.e., have less differences in area, volume and perimeters, as respect to those obtained using other systems (depending on use-case, e.g. RT-View, Eclipse, BrainLab, Slicer). (Point 3)

3.3 References

T. Munzner. A Nested Model for Visualization Design and Valida- tion. IEEE Transactions on Visualization and Computer Graphics, 15(6):921–928, nov.-dec. 2009.

T. Etiene, et al., “Verifying Volume Rendering Using Discretization Error Analysis”, IEEE TVCG, 2013.