



www.visceral.eu

Result meta-analysis

Deliverable number	<i>D4.5</i>
Dissemination level	<i>Public</i>
Delivery date	<i>18 May 2015</i>
Status	<i>Final</i>
Author(s)	<i>Abdel Aziz Taha, Allan Hanbury, Marianne Winterstein, Anna Walleyo, Markus Krenn and Oscar Jiménez</i>



This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Grant Agreement Number: 318068

Executive Summary

We analyze the results of the segmentation of the Anatomy1 and Anatomy2 benchmarks to infer more knowledge about the evaluation metrics, namely the metric sensitivities and some problems in relation to validating fuzzy segmentations. In particular, we take a deeper look at metrics by means of the following: (I) Analysis of the correlation between the 21 metrics that have been implemented in the evaluation software. (II) Analysis of the rankings produced by the metrics, compared with two manual rankings done by two different radiologists. (III) Performing the metric selection method [2] on the segmentations to rank the metrics according to their suitability for the segmentations, and judging this ranking using the manual rankings. (IV) Evaluating the same segmentations against fuzzy variants of the ground truth that are generated synthetically; then analyzing how the metric rankings differ.

The analysis on correlation among metrics shows that there are two main groups of metrics in terms of correlations, where the metrics in each group strongly correlate with each other, but have weak correlation with the metrics in the other group. After a deeper look into the metrics in each group, we found that in the first group, metrics do not take true negative voxels into consideration in contrast to the metrics in the second group, where true negatives are considered. Another observation is that the correlation between metrics is affected by the overlap between the segmentations compared, that is, if the Dice score is high the correlation between distance based metrics and overlap based metrics are high. On the contrary, when the overlap is low, distance based metrics are not more correlated with overlap based metrics. More detail about this analysis is in Section 2.

Analysis based on comparing manual expert rankings of segmentations with ranking produced by metrics shows: (i) Ranking single segmentations using metrics is less reasonable, since this type of ranking is sensible to small differences that are potentially irrelevant and are ignored in manual ranking. For this reason, ranking at single segmentation level has in general a weak to moderate correlation with manual ranking. (ii) Metric ranking at system level, given there are more than one segmentation produced by each system, is more reasonable, if used in combination with statistical testing to decide whether two systems differ in their performance. (iii) The result of ranking at system level shows that the four metrics selected for evaluating the segmentation benchmarks (Anatomy 1 and 2), namely the Dice coefficient (DICE), the average distance (AVD), the interclass correlation (ICC), and the adjusted Rand index (ARI) have a strong correlation with both of the manual expert rankings except for one metric in one of the rankings, namely the average distance (AVD) in Ranking 1. These four metrics have been selected based on a correlation study, done before the benchmarks, where an automatic metric selection method [2] was performed on brain tumor segmentations [8]. More detail about this analysis is in Section 3.

The metric selection method proposed in [2] was performed on the segmentations that have been ranked manually by the two radiologists. The metrics are then ranked by sorting them in ascending order according their sum of bias, which indicates their suitability for evaluating this set of segmentations, where the metric with the least bias is the most suitable. Now, this ranking of the metrics is compared with their ranking according their correlation with the manual ranking in order to test the efficiency of the selection method. The results show moderate correlation between the automatic metric selection and the selection depending on the manual ranking.

Analysis on fuzzy segmentation was done using (i) fuzzy and binary variants of the silver corpus, obtained by fusing the automatic segmentations of Anatomy 1 and 2, (ii) fuzzy segmentations provided by one of the participants in Anatomy 2, and (iii) synthetic fuzzy ground truth segmentations produced by performing smoothing filters on the original ground truth segmentation. In one experiment, each binary image in the silver corpus was compared with its corresponding fuzzy variant, using all metrics. The aim was to measure the invariant of metrics against fuzzification. This is to know which metric(s) should be used when fuzzy ground truth is used to evaluate binary segmentation and the opposite. In another experiment, systems are first ranked by using the binary ground truth (the official ranking of the benchmark), and the same systems were ranked using the fuzzy variant of the ground truth, i.e. the synthetic volume from (ii). The difference between the rankings in the two cases was observed for each

D4.5 Result meta-analysis

of the seven organs considered. Results show that using fuzzy instead of binary ground truth has a considerable impact on the ranking, given that the differences in the performance of systems are small. More detail about this analysis is in Section 4.

Table of Contents

1	Introduction	5
1.1	Validation of medical segmentation.....	5
1.2	Metrics for evaluating medical volumes.....	5
2	Analysis of metric correlation	7
2.1	Correlation between metrics	7
2.2	Influence of overlap on correlation.....	9
2.3	Guidelines for metric selection.....	11
3	Analysis based on manual rankings	11
3.1	Description of the manual rankings.....	12
3.1.1	Data set.....	12
3.1.2	Manual rankings	12
3.2	Correlation between metric and manual rankings at segmentation level.....	13
3.3	Correlation between manual and metric ranking at system level.....	14
3.4	Automatic metric selection	14
3.5	Discussion of the manual ranking analysis	16
4	Fuzzy segmentation and fuzzy metrics.....	17
4.1	Fuzzy segmentations.....	17
4.2	Fuzzy metrics	18
4.3	Analysis	18
4.3.1	Metric sensitivity against fuzzification	18
4.3.2	Ranking systems using binary/fuzzy ground truth	18
5	Conclusion.....	23
6	References	23

List of Abbreviations

MRI	Magnetic Resonance Imaging
CT	Computed Tomography
GT	ground truth
thd	threshold
TP	true positive
TN	true negative
FP	false positive
FN	false negative

1 Introduction

1.1 Validation of medical segmentation

Segmentation methods with high precision and high reproducibility are a main goal for assisting in surgical planning because they directly impact the results, e.g. the detection and monitoring of tumor progress. Accurately recognizing the change patterns is of great value for early diagnosis and efficient monitoring of diseases. Therefore, assessing the accuracy and the quality of segmentation algorithms is of great importance, which is a matter of the evaluation methodology. Segmentation evaluation is the task of comparing two segmentations by measuring the distance or similarity between them, where one is the segmentation to be evaluated and the other is the corresponding ground truth segmentation.

Thus, the knowledge about the metrics in terms of their strength, weakness, sensitivities, bias, and the aspects they measure, is essential for taking the decision about which metrics are to be used in the evaluation.

Medical segmentations are often fuzzy meaning that voxels have a grade of membership, e.g. the silver corpus in the VISCERAL project is such a case, where the segmentations are the result of averaging different segmentations of the same structure annotated by different annotators. Here, segmentations can be thought of as probabilities of voxels belonging to particular classes. One way of evaluating fuzzy segmentations is to cut the probabilities at a particular threshold to get binary representations that can be evaluated as crisp segmentations. However, thresholding is just a workaround that provides a coarse estimation and it is not always satisfactory. Furthermore, there is still the challenge of selecting the threshold because the evaluation results depend on the selection.

1.2 Metrics for evaluating medical volumes

In this section, we describe the metrics that have been selected for validating medical segmentation, and have been implemented in the EvaluateSegmentation tool for evaluating medical image segmentation, which is available as an open source project under the following link <https://github.com/codalab/EvaluateSegmentation>. The metrics implemented in this tool are presented in Table 1.

These metrics were selected based on a literature review of papers in which medical volume segmentations are evaluated. Only metrics with at least two references (papers) of use are considered. An overview of these metrics is available in Table 1. Depending on the relations between the metrics, their nature and their definition, we group them into five categories, namely:

- spatial overlap based,
- pair-counting based,
- information theoretic based,
- probabilistic based, and
- spatial distance based.

The aim of this grouping is to enable a reasonable selection when a subset of metrics is to be used, i.e. selecting metrics from different groups to avoid biased results.

Metric	Symbol	Category
Dice (=F1-Measure)	DICE	Spatial overlap based
Jaccard index	JAC	Spatial overlap based
True positive rate (Sensitivity, Recall)	TPR	Spatial overlap based
True negative rate (Specificity)	TNR	Spatial overlap based
False positive rate (=1-Specificity, FPR)	FPR	Spatial overlap based
positive predictive value (Precision)	PPR	Spatial overlap based
Accuracy	ACU	Spatial overlap based
F-Measure (F1-Measure=Dice)	FMS	Spatial overlap based
Volumetric Similarity	VS	Spatial overlap based
Global Consistency Error	GCE	Spatial overlap based
Rand Index	RI	Pair counting based
Adjusted Rand Index	ARI	Pair counting based
Mutual Information	MI	Information theoretic based
Variation of Information	VOI	Information theoretic based
Interclass correlation	ICC	Probabilistic based
Probabilistic Distance	PBD	Probabilistic based
Cohens KAP	KAP	Probabilistic based
Area under ROC curve	AUC	Probabilistic based
Hausdorff distance	HD	Spatial distance based
Average distance	AVD	Spatial distance based
Mahalanobis Distance	MHD	Spatial distance based

Table 1: Metrics implemented in the evaluation SW (EvaluateSegmentation).

For evaluation of medical image segmentation, four metrics were selected from the 21 metrics presented in Table 1. The following criteria were considered

1. The metrics were selected so that they cover as many different categories as possible. One metric was selected from each of the following categories: (i) spatial overlap based metrics, (ii) distance-based metrics, (iii) probabilistic based metrics, and (iv) pair-counting-based metrics with chance adjustment.
2. Two of the metrics are capable of comparing fuzzy segmentations, i.e. have fuzzy as well as crisp definition. For the other two metrics, fuzzy comparison is calculated indirectly by cutting the voxel values at 0.5 threshold.
3. From those metrics that meet the criteria above, metrics were selected that have the most correlation with the rest of the metrics in each category.

Depending on these criteria, the following metrics have been considered for validating segmentations in all the segmentation benchmarks of the VISCERAL project: the Dice coefficient (DICE), the average distance (AVD), the interclass correlation (ICC), and the adjusted Rand index (ARI)

In this document, we provide analysis about evaluation metrics based on the results of the segmentations in the Anatomy1 and Anatomy2 Benchmarks. In Section 2, we analyze the correlation between the 21 metrics presented in Table 1 and discuss the properties of each metric. In Section 3, we present an analysis based on comparison between rankings produced by the metrics and manual rankings made by radiologists. Finally, in Section 4, we validate a subset of the segmentations of Anatomy2 against synthetic fuzzy variants of the ground truth and discuss the results.

2 Analysis of metric correlation

Metrics differ in their properties and thus in their suitability for different tasks and different data. Selecting a suitable metric is not a trivial task. In this section we present a metric analysis that was initially presented in [1]. In particular, it provides analysis of the correlation between the metrics presented in Table 1 to infer information about their properties and capabilities for discovering different types of error. Based on this analysis, we provide a guideline for selecting a suitable metric, given a data set and a task.

2.1 Correlation between metrics

One way to analyze metrics is to examine the correlation between rankings produced by them. Figure 1 shows the result of a correlation analysis between the rankings produced by 16 of the metrics presented in Table 1 when applied to a data set of 4833 automatic MRI and CT segmentations. In this data set, all medical volumes provided by all the participants of the VISCERAL project in its two initial challenges, namely Anatomy1 and Anatomy2, were included. Each medical volume is a segmentation of only one of 20 anatomy structures varying from organs like the lung, liver, and kidney to bone structures like the vertebra, glands like the thyroid, and arteries like the aorta. More details on these structures are available in [3]. Note that the Jaccard (JAC) and F-Measure (FMS) were excluded because they provide the same ranking as the Dice coefficient (DICE), a fact that follows from the equivalence relations between them [1]. Also FPR and FNR were excluded because of their relations to TNR and TPR respectively. In a first step, volume segmentations were ranked using each of the metrics to get 16 rankings in total. Then, the pairwise Pearson's correlation coefficients were calculated. Note that analyzing the correlation between rankings instead of metric values solves the problem that some of the metrics are similarities and some others are distances and avoids the necessity to convert distances to similarities as well as to normalize metrics to a common range. Each cell in Figure 1 represents the Pearson's correlation coefficients between the rankings produced by the corresponding metrics. The darkness of the cells represent the strength of the correlation.

Metrics in Figure 1 can be divided into three groups based on the correlation between the rankings produced by them, one group is at the top left (Group 1) including ARI, KAP, ICC, DICE, AVD, MHD, PBD, and VS and another group is at the bottom right (Group 2) including TNR, RI, GCE, and VOI. The metrics in each of these groups strongly correlate with each other, but have no correlation with metrics in the other group. The remaining metrics (Group 3) including MI, AUC, TPR, and HD have medium correlation between each other and the other groups. A deeper consideration in the metric definitions shows that Group 1 and Group 2 classify the metrics according to whether they consider or not the true negatives (background voxels) in their definitions. While all metrics in Group 2 include the true negatives in their definitions, none of the metrics in Group 1 does this. Note that the adjusted Rand index and the KAP measures principally include the true negatives in their definitions, but both of them perform chance adjustment, which eliminates the impact of the true negatives, i.e. avoids that the influence of the background dominates the result [4]. Also note that the average distance (AVD) and the

D4.5 Result meta-analysis

Mahalanobis distance (MHD) in Group 1 do not consider the true negatives, since they are based on the distances between the foreground voxels (non-zero voxels). Considering the true negatives in the evaluation has a large impact on the result, since the background (normally the largest part of the segmentation) contributes to the agreement. Figure 2 illustrates, by means of a real example, how metrics based on the true negatives change the resulting rankings when the true negatives are reduced by selecting a smaller bounding cube [6]. Such metrics are biased against the ratio between the total number of foreground voxels and the number of the background voxels, which is denoted as the class imbalance. This leads to segmentations with large segments being penalized and those with small ones being rewarded, a case that is common in medical image segmentation e.g. when the quality of two segmentations is to be compared, where one of them is larger, and the other one is smaller than the ground truth segmentation. Vinh et. al [7] stated that such metrics need chance adjustment, since they do not meet the constant baseline property.

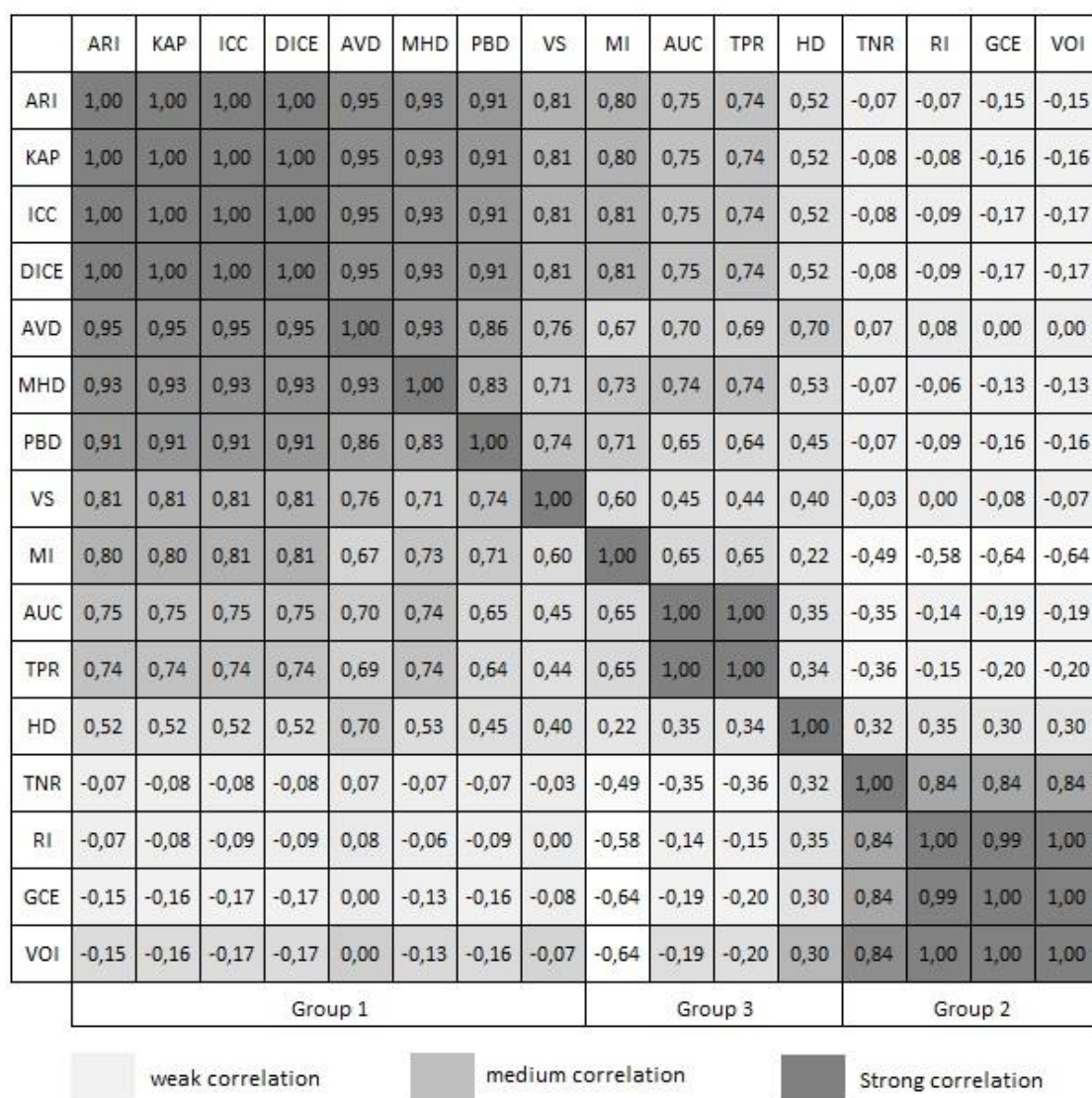


Figure 1: the correlation between the rankings produced by 16 different metrics. The pair-wise Pearson's correlation coefficients between the rankings of 4833 medical volume segmentations produced by 16 metrics. The darkness of each cell represents the strength of the correlation.


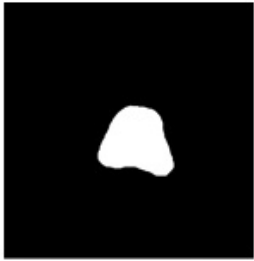

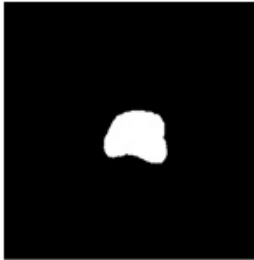


	Ground truth		Segmentation	
A		↔		DICE 0,939
				AVG 0,204
B		↔		DICE 0,839
				AVG 1,149
Á		↔		DICE 0,939
				AVG 0,204
				VS 0,953
				<u>RI</u> <u>0,986</u>
				<u>GCE</u> <u>0,013</u>
				<u>TNR</u> <u>0,994</u>
				DICE 0,839
				AVG 1,149
				VS 0,855
				<u>RI</u> <u>0,970</u>
				<u>GCE</u> <u>0,026</u>
				<u>TNR</u> <u>0,999</u>
				DICE 0,939
				AVG 0,204
				VS 0,953
				<u>RI</u> <u>0,878</u>
				<u>GCE</u> <u>0,124</u>
				<u>TNR</u> <u>0,897</u>

Figure 2: the effect of decreasing the true negatives (background) on the ranking. Both of the segmentations in A and B is compared with the same ground truth. All metrics assess that the segmentation in A is more similar to the ground truth than in B. In Á, the segmentation and ground truth are the same as in A, but after reducing the true negatives by selecting a smaller bounding cube. The metrics RI, GCE, and TNR change their rankings as a result of reducing the true negatives. Note that some of the metrics are similarities and others are distances.

2.2 Influence of overlap on correlation

Obviously, the correlation between rankings produced by overlap based metrics and rankings produced by distance based metrics cannot hold in all cases because when the overlap between segments is zero, all overlap based metrics are zero regardless of how far the segments are from each other, on the contrary distance based metrics still provide values dependent on the spatial distance between the segments. This motivated us to examine how the correlation described in Section 2.1 behaves when only segmentations with overlap values in particular ranges are considered.

Figure 3 shows the Pearson's correlation between the DICE and each of the other metrics when the measured DICE is in a particular range. One important observation is that the correlation between DICE and the distance based metrics (AVD, HD, and MHD) decreases with decreasing overlap, i.e. with increasing false positives and false negatives. This is intuitive because overlap based metrics, in contrast to distance based metrics, don't consider the positions of voxels that are not in the overlap region (false positives and false negatives), which means they provide the same value independent of the distance between the voxels. It follows that increasing the false positives and/or false negatives (decreasing overlap) means increasing the probability of divergent correlation.

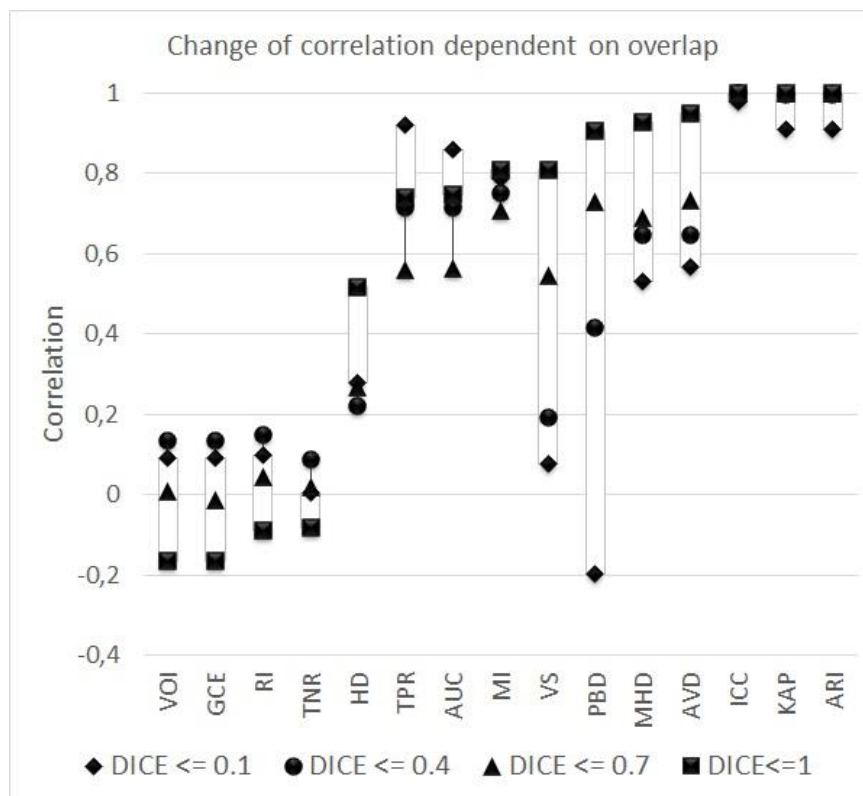


Figure 3: the effect of overlap on the correlation between rankings produced by different metrics. The positions and heights of the bars show how metrics correlate with DICE and how this correlation depends on the overlap between the compared segmentations. Four different overlap ranges are considered.

Another observation is the strongly divergent correlation between volumetric similarity (VS) and DICE. This divergence is intuitive since the VS only compares the volume of the segment(s) in the automatic segmentation with this in the ground truth, which implicitly assumes that the segments are optimally aligned. Obviously, this assumption only makes sense when the overlap is high. Actually, the VS can have its maximum value (one) even when the overlap is zero. However, the smaller the overlap, the higher is the probability that two segments that are similar in volume are not aligned, which explains the strong divergence in correlation when the overlap is low. Now, since one can assume that the probability of wrongly aligned segments is higher when the segments are small and vice versa (the degree of freedom for the segment location is higher when the segment is small), it follows that the VS is not recommended for segmentations with small segments.

Finally, the highest divergence in the correlation is observed with the probabilistic distance (PBD). This is caused by the fact that PBD, in contrast to DICE, over-penalizes false positives and false negatives. This can be explained by means of the definition of the PBD in [1]: differences in the voxel values in the compared segmentations have a double impact on the result because they increase the numerator and decrease the denominator at the same time, causing the distance to increase rapidly. Actually, the PBD even reaches infinity when the overlap reaches zero. PBD behaves the opposite of the VS regarding the sensitivity to the alignment, i.e. it strongly penalizes alignment errors, which makes it suitable for tasks where the alignment is of more interest than the volume and the contour.

2.3 Guidelines for metric selection

Different metrics have sensitivities to different properties of the segmentations, and thus they can discover different types of error. Taha et al. [2] provide a formal method for choosing the most suitable metric, given a set of segmentations to be evaluated and a segmentation task. In this section, we conclude the discussion in Sections 2.1 and 2.2 to provide the following guidelines for choosing a suitable metric:

1. When the objective is to evaluate the general alignment of the segments, especially when the segments are small (the overlap is likely small or zero), it is recommended to use distance based metrics rather than overlap based metrics. The volumetric similarity (VS) is not suitable in this case.
2. Distance based metrics are recommended when the contour of the segmentation, i.e. the accuracy at the boundary, is of importance [5]. This follows from being the only category of metrics that takes into consideration the spatial position of false negatives and false positives.
3. The Hausdorff distance is sensitive to outliers and thus not recommended to be used when outliers are likely. However, methods for handling the outliers, such as the quantile method [9], could solve the problem, otherwise the average distance (AVG) and the overlap based metrics as well as probabilistic based metrics are known to be stable against outliers.
4. Probabilistic distance (PBD) and overlap based metrics are recommended when the alignment of the segments is of interest rather than the overall segmentation accuracy [10].
5. Metrics considering the true negatives in their definitions have sensitivity to segment size. They reward segmentations with small segments and penalize those with large segments [6]. Therefore, they tend to generally penalize algorithms that aim to maximize recall and reward algorithms that aim to maximize precision. Such metrics should be avoided in general, especially when the objective is to reward recall, e.g. segmentations having the goal of tumor removal.
6. When more than one metric are to be combined, the aim should be to select them from different categories (Table 1) as well as to avoid selecting metrics that are strongly correlated (Figure 1). This is to avoid biased evaluation.
7. When the segmentations have a high class imbalance, e.g. segmentations with small segments, it is recommended to use metrics with chance adjustment, e.g. the KAP measure (KAP) and the adjusted rand index (ARI) [4].

3 Analysis based on manual rankings

In this section, we provide an analysis of the metrics based on two manual rankings of segmentations, done by two medical experts. Manual rankings provide a references for judging metrics and evaluation methods. That is, when evaluating segmentations by comparing them with the corresponding ground truth using distance or similarity metrics, one gets scores denoting how similar or how far the segmentations are from the ground truth. However, since different metrics provide different scores, there is a need of another level of ground truth that judges, which metric provides scores that are more correlated with the expert rankings than other metrics.

Another aim of this analysis is to validate the selection of the subset of four metrics from Table 1 used for evaluation of medical image segmentation in the Anatomy1 and Anatomy2 Benchmarks of the VISCERAL project.

In Section 3.1, we describe the manual rankings. We then analyze the correlation between the manual ranking and the rankings produced by metrics: in Section 3.2, the ranking is done at segmentation level, while in Section 3.3, the ranking is done at system level. Finally in Section 0, we discuss the results of the manual ranking analysis.

3.1 Description of the manual rankings

3.1.1 Data set

To provide a manual ranking, 483 segmentations were selected by medical experts from the output of the Anatomy2 participating algorithms. This segmentation set has the following properties:

- The segmentations correspond to six organs/structures, namely liver, pancreas, urinary bladder, aorta, left lung, and right kidney. These structures were selected by medical experts so that different aspects are covered, like size and shape, etc.
- The segmentations corresponds to 110 different medical cases, where a case is defined as a combination of a ground truth volume and a structure (organ).
- The segmentations were produced by seven participating algorithms. However, different medical cases were segmented by different number of algorithms. This means for some medical cases, seven segmentations are available, but for other medical case there are less than seven. For the ranking analysis, only those medical cases were considered for which at least three segmentations are available. These are only 92 medical cases.

3.1.2 Manual rankings

The segmentations described above have been ranked by two different radiologists separately, resulting in two different rankings, which we will call Manual Ranking 1 (MRK1) and Manual Ranking 2 (MRK2). The ranking was performed in a double blind way.

The following subjective scoring system was observed:

Score	Ranking criteria
1	Severe deviation to other organs, no connection with expected organ segmentation.
2	Evident crossing of organ border, organ parts missing from segmentation
3	Irregular segmentation with respect to segmentations guidelines from Deliverable 2.3.1.
4	Minor deviations from segmentation guidelines.
5	Optimal segmentation, organ borders and segmentation guidelines from VISCERAL Deliverable 2.3.1 respected

For each ground truth segmentation, the corresponding automatic segmentations were considered as one group, within which these segmentations are ranked. The ranking was created using a point-based system, where different qualities are rated using points, i.e. the existence of particular qualities is rated by adding pre-defined numbers of points depending on the relevance of these qualities from a medical point of view. The absence of the quality is rated by subtracting a number of points and the rank is the sum of points achieved.

As a consequence of this ranking system, rankings have not to start with one and end with the number of objects being ranked (which is however the case with metric rankings). Also, different objects may have the same rank (which is not common with metric rankings). For example, it is common with manual ranking that five segmentations are ranked with 1, 2, 2, 2, 3. This is not common in rankings produced by metrics, since equal metric values are very unlikely.

In order to test how the two manual rankers agree between each other, the Pearson's correlation between the two manual rankings was measured. **The correlation between the manual rankings, RNK1 and RNK1, is 0.62.** This is a moderate correlation, which means that there is a non-ignorable discrepancy between the rankers.

3.2 Correlation between metric and manual rankings at segmentation level

We analyze the correlation between rankings of groups of segmentations produced by each of the metrics in Table 1 and the rankings of the same segmentations based on the manual rankings (MRK1 and MRK2). This analysis is to infer which metrics have the most correlation with the manual ranking.

The rankings in this experiment are at segmentation level, which means that individual segmentations corresponding to the same ground truth are ranked. To this end, the segmentations were grouped so that each group consists of a medical case (volume) and the corresponding segmentations. The segmentations in each group are then ranked using each of the metrics by comparing each of the segmentations with its corresponding ground truth. The segmentation with the lowest match is given the lowest rank and that with the best match was given the highest rank. This is in order to get a ranking that is comparable with the manual ranking done based on the point system, as described in Section 3.1.2.

Table 2 shows the correlations between each of the metrics presented in Table 1 and each of the two manual rankings, MRK1 and MRK2. The metrics are sorted according to the correlation with MRK1. Note the highest correlation value (0.64) is a moderate correlation, and many of the metrics have weak correlation. This is expected, since ranking at segmentation level using the metrics considers very small changes, which do not necessarily reflect an improvement, e.g. differences caused by chance. For this reason, we provide another correlation analysis at system level, in Section 3.3, that uses significance testing to decide whether one system has better performance than another.

Manual Ranking 1 (MRNK 1)			Manual Ranking 2 (MRNK2)		
metric		Pearson's correlation	metric		Pearson's correlation
Average distance	AVD	0.57	Rand Index	RI	0.56
Adjusted Rand Index	ARI	0.54	Variation of Information	VOI	0.56
Dice	DICE	0.54	Average distance	AVD	0.56
F-Measure	FMS	0.54	Accuracy	ACU	0.56
Interclass correlation	ICC	0.54	Global Consistency Error	GCE	0.55
Cohens KAP	KAP	0.54	Adjusted Rand Index	ARI	0.52
Probabilistic Distance	PBD	0.54	Dice	DICE	0.52
Rand Index	RI	0.54	F-Measure	FMS	0.52
Jaccard index	JAC	0.54	Interclass correlation	ICC	0.52
Accuracy	ACU	0.53	Cohens KAP	KAP	0.52
Variation of Information	VOI	0.53	Jaccard index	JAC	0.52
Global Consistency Error	GCE	0.53	Probabilistic Distance	PBD	0.51
Mutual Information	MI	0.47	Mutual Information	MI	0.46
Mahalanobis Distance	MHD	0.44	Mahalanobis Distance	MHD	0.41
Hausdorff distance	HD	0.43	Hausdorff distance	HD	0.40
Area under ROC curve	AUC	0.39	positive predictive value	PPR	0.38
True positive rate	TPR	0.39	Area under ROC curve	AUC	0.36
Volumetric Similarity	VS	0.27	True positive rate	TPR	0.36
positive predictive value	PPR	0.27	Volumetric Similarity	VS	0.30
Fallout	FPR	0.17	Fallout	FPR	0.26
True negative rate	TNR	0.17	True negative rate	TNR	0.26

Table 2: Pearson’s correlation between each of the metrics presented in Table 1 and the manual rankings MRK1 and MRK2 at segmentation level. The metrics are sorted according to decreasing correlation.

Metrics, in contrast to manual judgment, always attempt to assign two different ranks to two segmentations, regardless of how small the difference in the metric scores is. Therefore, we do not expect high correlation between metric rankings and manual rankings at segmentation level. The aim of this analysis is rather to find the metric(s) with the highest correlation, regardless of the absolute value of the correlation. In Section 3.3, we analyze the correlation between manual and metric ranking at system level, where a stronger correlation is expected.

3.3 Correlation between manual and metric ranking at system level

In this experiment the manual ranking of the individual segmentations done by the medical experts will be used to rank the systems that produced the segmentations. Instead of considering groups consisting of a medical case (volume) and the corresponding segmentations as done in Section 3.2, we consider all the segmentations produced by a particular system for the same organ together. This enables comparing the systems based on a statistical test. This is done as follows: For each organ separately, the metric scores of the systems are calculated as the average over all segmentations produced by the each system. Also the manual ranks were averaged over the same segmentations (in this case, the manual ranks are considered as scores). Now, for each system we have an average manual score and an average score for each metric. From these scores the systems are ranked depending on (i) the average scores, and (ii) significance test using the sign test to ensure that the difference between the average scores is significant. The sign test is performed on the score groups that have been averaged. In a first step, the systems are sorted according to their average scores ascending. Then the ranks are given as follows: starting with the first system (S_1), with the lowest score, it is given the rank 1. Then for each next system (S_i), if there is no significant difference to the previous system (S_{i-1}), according to a sign test, then it is assigned the same rank as (S_{i-1}), otherwise the next rank. Note that the same ranking is generated from the manual average ranks.

Table 3 shows the Pearson's correlation between each of the metrics in Table 1 and the manual ranking.

3.4 Automatic metric selection

In this section, we perform the formal metric selection method, proposed in [2], on the same segmentations that have been ranked manually by the two radiologists. Given a set of effectiveness metrics and a set of segmentations, this formal method aims to find the most suitable metric(s) for evaluating the segmentation. The selection method is primarily based on that metrics can be biased towards or against properties of the images being segmented, meaning that particular metrics over-penalize or over-reward segmentations given particular properties. According to this method, the bias of a particular metric to a particular property is inferred by automatically analyzing how the average scores of groups of segmentations differ in two cases, the first when the segmentations are grouped randomly and the second when they are grouped according to the property for which the bias is being measured. Once the biases of each metric to each property are inferred, the selection of the metrics is achieved based on the sum of biases, i.e. metrics are selected that have the least bias. For this set of segmentations, the following properties have been used: segment size, volume size, noise, deviation, shape signatures, sphericity, boundary smoothness, and recall. It is important to mention that the method provides the possibility to use property weighting to reflect the subjective preference that meet the specific goal of the segmentation (more details in [2]). However, this feature has not been used in this experiment, because further analysis of the comments of the ranking radiologists is required to determine the property weights.

D4.5 Result meta-analysis

Manual Ranking 1 (MRNK1)			Manual Ranking 2 (MRNK2)		
metric		Pearson's correlation	metric		Pearson's correlation
Volumetric Similarity	VS	0.81	Mahalanobis Distance	MHD	0.75
Jaccard index	JAC	0.81	Hausdorff distance	HD	0.66
Dice	DICE	0.81	Adjusted Rand Index	ARI	0.65
F-Measure	FMS	0.81	Dice	DICE	0.64
Interclass correlation	ICC	0.81	F-Measure	FMS	0.64
Cohens KAP	KAP	0.81	Interclass correlation	ICC	0.64
Adjusted Rand Index	ARI	0.80	Cohens KAP	KAP	0.64
Area under ROC curve	AUC	0.72	Jaccard index	JAC	0.62
True negative rate	TNR	0.72	Accuracy	ACU	0.56
Accuracy	ACU	0.71	Global Consistency Error	GCE	0.56
Global Consistency Error	GCE	0.71	Rand Index	RI	0.56
Rand Index	RI	0.71	Variation of Information	VOI	0.56
Variation of Information	VOI	0.71	Average distance	AVD	0.54
positive predictive value	PPR	0.64	positive predictive value	PPR	0.53
Mahalanobis Distance	MHD	0.47	Fallout	FPR	0.48
Probabilistic Distance	PBD	0.41	True positive rate	TPR	0.48
Average distance	AVD	0.39	Volumetric Similarity	VS	0.47
Hausdorff distance	HD	0.38	Probabilistic Distance	PBD	0.36
Fallout	FPR	0.23	Area under ROC curve	AUC	0.34
True positive rate	TPR	0.23	True negative rate	TNR	0.34
Mutual Information	MI	0.19	Mutual Information	MI	0.14

Table 3: Pearson's correlation between each of the metrics presented in Table 1 and the manual rankings MRK1 and MRK2 at system level. Sorted according to decreasing correlation.

After performing the method for each metric and each property, the metrics were then sorted ascending according to their sum of bias (Table 4), and given ranks based on the sum of bias, which indicate their suitability for evaluating this set of segmentations, where the metric(s) with the least bias are the most suitable.

Now, these ranks of suitability of the metrics are compared with the ranks inferred from their correlation with the manual ranking, obtained from the experiment in Section 3.3. In other words, we compare the automatic metric selection, based on this method, with the metric selection that would be done based on the manual ranking in order to test the efficiency of the selection method. The results in Table 4 show moderate correlation between the automatic metric selection and the selection depending on the manual ranking. Note that the moderate correlation could be increased by using weights for the properties that reflect the goal of each ranker. Such weights have not been used in this experiment. Normally, such weights can be defined based on an analysis of the comments of the radiologist that explain their judgments. To understand the usefulness of using such weights, consider the fact that the correlation between the manual rankings of the two radiologists is moderate (0.62), which implies that we should not expect a correlation stronger than moderate without using settings, e.g. weights that reflect the individual goals of the segmentation.

Metric		Automatic		Manual ranking 1		Manual Ranking 2	
		Sum of biases	Automatic rank of suitability	correlation with MRK1	Rank according MRK1	Correlation with MRK2	Rank according MRK2
Volumetric Similarity	VS	30,59	2	0.81	1	0.47	10
Jaccard index	JAC	30,59	2	0.81	1	0.62	5
Cohens KAP	KAP	30,29	1	0.81	1	0.64	4
Dice	DICE	30,59	2	0.81	1	0.64	4
F-Measure	FMS	30,59	2	0.81	1	0.64	4
Interclass correlation	ICC	32,43	3	0.81	1	0.64	4
Adjusted Rand Index	ARI	30,29	1	0.80	2	0.65	3
Area under ROC curve	AUC	32,77	4	0.72	3	0.34	12
True negative rate (Specificity)	TNR	54,09	13	0.72	3	0.34	12
Accuracy	ACU	42,40	10	0.71	4	0.56	6
Rand Index	RI	50,34	10	0.71	4	0.56	6
Global Consistency Error	GCE	50,34	11	0.71	4	0.56	6
Variation of Information	VOI	53,19	12	0.71	4	0.56	6
positive predictive value	PPR	58,10	13	0.64	5	0.53	8
Mahalanobis Distance	MHD	36,41	8	0.47	6	0.75	1
Probabilistic Distance	PBD	34,82	7	0.41	7	0.36	11
Average distance	AVD	40,19	9	0.39	8	0.54	7
Hausdorff distance	HD	34,41	6	0.38	9	0.66	2
True positive rate (Sensitivity)	TPR	32,77	4	0.23	10	0.48	9
Fallout	FPR	58,10	13	0.23	10	0.48	9
Mutual Information	MI	61,68	14	0.19	11	0.14	13
Pearson's correlation with automatic ranking				0.57		0.42	

Table 4: the results of performing the automatic metric selection [2]. The column ‘sum of bias’ provides the bias of each metric over all properties used, and is the base for ranking the metrics according their suitability, column automatic rank. In the next columns, the correlation and resulting metric suitability rank according to the expert rankings, MRK1 and MRK2

3.5 Discussion of the manual ranking analysis

The following conclusions can be inferred from the results of the analysis using the manual rankings (results presented in Table 2 and Table 3)

- Table 3 shows correlations at system level that are significantly stronger than the correlations of rankings at segmentation level (Table 2). Actually, this is intuitive because the errors (differences from the manual ranking) in the ranking at segmentation level are higher than in rankings at system level. This stems from the fact that ranking single segmentations using metrics is sensitive to small differences, in contrast to manual rankings, where small differences are ignored. Using significance testing in ranking at system level efficiently solves the problem, since the ranking becomes similar to the manual ranking: only systems that have significant performance difference are assigned different rankings, otherwise the same rank. The results of this experiment shows the necessity of using significance tests for ranking.
- The four metrics selected for evaluating segmentation in the VISCERAL project, namely the Dice coefficient (DICE), the interclass correlation (ICC), the average Hausdorff distance

(AVD), and the adjusted Rand index (ARI) are in general (except for the AVG in Ranking 1) ranked at the top, which means they have strong correlation with expert ranking. These four metrics have been selected from the 21 metrics based on a correlation analysis on brain tumor segmentations from the BRATS challenge [8], using the automatic metric selection method proposed in [2].

- One observation is interesting for a further analysis, namely the differences in how the metrics are placed in Table 3 for MRK1 and MRK2. For example, the volumetric similarity (VS) is placed at the top for MRK1, but at the bottom in MRK2. This is also the case for many other metrics. This can be explained by the weak correlation between the two rankers – the correlation between the two manual rankings is only 0.62 (Section 3.1.2). However, these differences should be related to the criteria considered in the manual ranking by each of the rankers, i.e. the subjective rating of the different qualities of the segmentations. A possible further analysis is linking this issue to the comments provided by the rankers for each segmentation.

4 Fuzzy segmentation and fuzzy metrics

In this section, we analyze the impact of using fuzzy metrics on the ranking results. We analyze this issue from several sides trying to answer the following questions: (i) how is the behavior of the different metrics as a result of including fuzzy segmentations? (ii) What is the impact of using binary ground truth to validate fuzzy segmentation using fuzzy metrics? (iii) How are the results different when using fuzzy ground truth? (iv) How are they different when using a threshold at some value?

4.1 Fuzzy segmentations

Fuzzy segmentations are common in the medical volume segmentation. As ground truth (GT), such segmentations can be as a result of averaging annotations done by different annotators. Another case is the fusion of automatic segmentations to produce a silver corpus. In these cases, voxels are not assigned to structure as a binary relation, but rather as a probability of membership to the structure.

Fuzzy segmentation is also common as automatic segmentations produced by the algorithms being ranked. Depending on the approach used for segmentation, the membership of the voxels can be defined as probabilities resulting in fuzzy segmentations. Also dependent on the segmentation task, there could be cases where fuzzy memberships are required to represent boundaries where an exact separation between structures is not possible.

The aim of this analysis is to infer how sensitive metrics are against fuzzification of images, in other words, how a particular metric responds to smoothing a particular volume segmentation. This analysis is motivated by the following. On the one hand, metrics with high fuzzification sensitivity are required to distinguish the accuracy of the systems. This is required if there is fuzzy ground truth available and the segmentations being evaluated are fuzzy as well. On the other hand, if only a binary ground truth is available, and the segmentations being evaluated are fuzzy or mixed, then the question is about the negative impact of using fuzzy ground truth on evaluation results. This holds also for the opposite case.

In the Anatomy2 Benchmark, only one of the participating algorithms produces fuzzy segmentations. This algorithm is denoted as Algorithm A throughout this section. However, only binary ground truth segmentations have been used in the evaluation of Anatomy1 and Anatomy2. Here, it has been observed that there are differences in the system rankings due to using the threshold option in the evaluation. These differences are not negligible.

The segmentations involved in this analysis are:

- **Binary ground truth (BGT):** This is the official binary ground truth, used for validating the challenge.

- **Synthetic fuzzy ground truth (FGT):** Since there are only binary ground truth segmentations, the fuzzy ground truth was generated synthetically: from each of the ground truth segmentations, a fuzzy variant was produced by smoothing the corresponding ground truth using a mean filter.
- **Binary silver ground truth (BSGT):** the silver corpus was generated by fusing all the automatic segmentations
- **Fuzzy silver ground truth (BSGT):** in another variant, a fuzzy silver corpus is generated by fusing all the automatic segmentations.
- **Binary automatic segmentations (BAS):** these are the automatic segmentations produced by most of the participating algorithms.
- **Fuzzy automatic segmentation (FAS):** these are the fuzzy segmentations produced by one of the participating algorithms, denoted by Algorithm A throughout the section.

4.2 Fuzzy metrics

From the metrics presented in Table 1, metrics that have fuzzy implementation are only those that are based on the confusion matrix, namely DICE, JAC, TPR, TNR, FPR, PPR, ACU, FMS, VS, RI, ARI, MI, VOI, PBD, KAP, and AUC. More about the fuzzy implementation of the metrics is available in [1]. All other metrics, e.g. spatial based distances, do not have fuzzy implementation. As a workaround for those metrics, fuzzy images are cut at 0.5 threshold, before they are compared as binary images. In this section, only metrics with fuzzy definition are considered.

4.3 Analysis

4.3.1 Metric sensitivity against fuzzification

The aim of this experiment is to infer how invariant metrics are against fuzzification of images, that is, how a particular metric responds to smoothing a particular volume segmentation. To this end we compared each binary volume in the silver corpus (BSGT) with its corresponding volume from the fuzzy silver corpus (FSGT) using each of the 16 metrics for which fuzzy implementations exist. This results in 16 similarities/distances per comparison (segmentation pair), which are then averaged over all pairs to get 16 average scores, presented in Figure 4. The assumption is that metrics that measure less average discrepancy between the binary volumes and their fuzzy variants are more invariant against fuzzification.

Results in Figure 4 show that metrics are differently invariant against fuzzification, that is, they have different capabilities in discovering changes due to fuzzification. Metrics that include the true negatives (TN) in their definitions (e.g. ARI, ACU, TNR) are in general less sensitive to fuzzification, in contrast to other metrics not considering the TN, like DICE, KAP, and JAC. Also one can observe that the discrepancy metrics FPR, PBD, and VOI are also invariant against fuzzification because they provide very small distances ($\ll 0.01$ voxel) between binary images and their corresponding smoothed images.

4.3.2 Ranking systems using binary/fuzzy ground truth

The aim of this experiment is to infer how system rankings, using metrics, change when using fuzzy instead of binary ground truth in two cases: (i) when the segmentations being evaluated are binary, (ii) when they are fuzzy. The following types of segmentations have been used in this experiment: The binary ground truth, officially used for validating the Anatomy2 Benchmark, denoted by (BGT). The synthetic fuzzy ground truth generated from each of the binary ground truth by smoothing each volume using a mean filter (FGT). The binary automatic segmentations, i.e. the output from most of the participating algorithms (BAS). The fuzzy automatic segmentation, the segmentations produced by one single participating algorithm, denoted by Algorithm A (FAS).

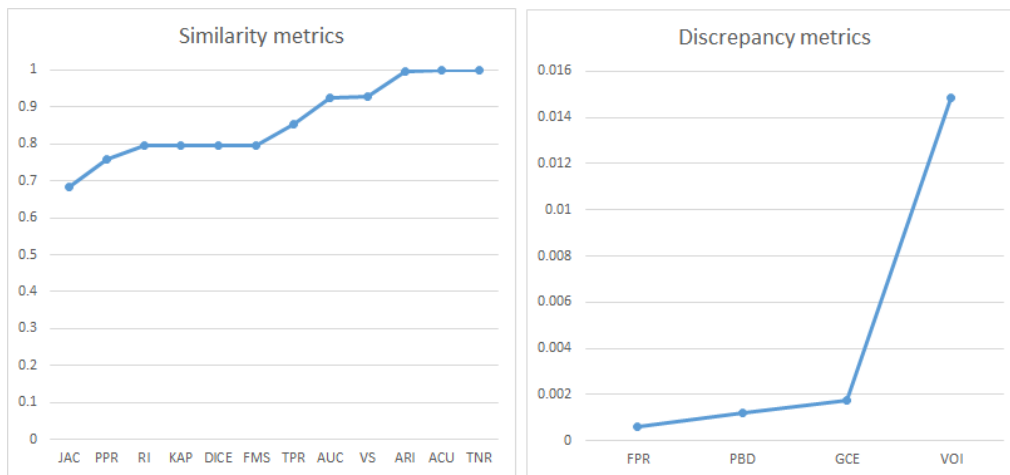


Figure 4: the average similarity between binary volumes and their corresponding fuzzy variant.

In Figure 5, Figure 6, and Figure 7 are the results of the same analysis performed for three selected metrics, namely Dice coefficient (DICE), Interclass Correlation (ICC), and Adjusted Rand Index (ARI) respectively. The three metrics are selected to represent three different metric categories according to Table 1. Note that no metrics are selected for the spatial distance based metrics, because there is still no fuzzy implementation for such metrics in the evaluation software.

There are seven systems to be ranked (Systems A to L in the figures below) depending on the quality of the segmentations produced for each of seven organs (left kidney, right kidney, liver, left lung, right lung, left psoas major muscle, and right psoas major muscle), which means the systems are ranked for each organ separately. The participating algorithms B to L produce only binary volumes whereas Algorithm A produces only fuzzy segmentations.

The ranking is performed in three different cases: (i) The ground truth is binary (BGT) and the segmentations are as is (fuzzy for algorithm A and binary otherwise). This case is denoted by “binary GT” in the results. (ii) The ground truth is fuzzy (FGT) and the segmentations are as is. This case is denoted by “fuzzy GT” in the results. (iii) Fuzzy segmentations of Algorithm A are cut at 0.5 threshold to get binary representations. The other segmentations and the ground truth are as is, thus all images, involved in this case are binary, are binary. This case is denoted by “threshold at 0.5” in the results.

To indicate how average scores are deviated between the algorithms, as well as between the three cases, we added standard deviation columns and a standard deviation row, as shown in Figure 5, Figure 6, and Figure 7.

The first observation is regarding Algorithm A, which produces fuzzy segmentations as a single algorithm. Here, Algorithm A has the best ranking when the corresponding segmentations are evaluated using 0.5 threshold or against a fuzzy ground truth, but it is has a considerable disadvantage when using the binary ground truth. Thus it is strongly recommended to use a threshold option when the segmentations/ground truth are mixed in terms of binary and fuzzy modes.

The second observation is that the sensitivity in the resulting rankings is dependent on the deviations between the average scores of the systems, the less the deviation, the more the rankings change between the three cases. That is, if the algorithms are similar in their performance, then using a binary instead of a fuzzy ground truth, or the opposite, has a considerable impact on the system ranking. For example the average scores of the systems have the most deviation with kidney and liver, so the rankings of the systems is exactly the same in the three cases. On the contrary, system average scores have low deviations with lungs and psoas major muscles, therefore the rankings of the systems considerably change between the three cases. We recommend therefore to take the score deviations into account when there are mixed segmentations/ground truth in terms of fuzzy and binary.

D4.5 Result meta-analysis

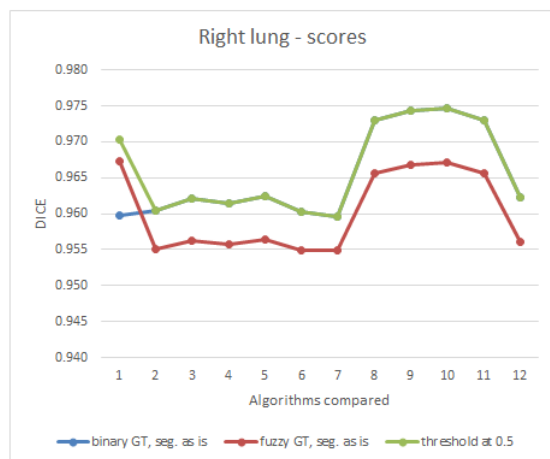
Ranking using the DICE measure in different combinations of binary and fuzzy images

algorithms compared	left kidney		right kidney		liver		left lung		right lung		left psoas major muscle		right psoas major muscle		Volume count
	29663		29662		58		1326		1302		32249		32248		
	binary GT	Standard deviation (std.)	binary GT	Standard deviation (std.)	binary GT	Standard deviation (std.)	binary GT	Standard deviation (std.)	binary GT	Standard deviation (std.)	binary GT	Standard deviation (std.)	binary GT	Standard deviation (std.)	
	Fuzzy GT	threshold at 0.5	Fuzzy GT	threshold at 0.5	Fuzzy GT	threshold at 0.5	Fuzzy GT	threshold at 0.5	Fuzzy GT	threshold at 0.5	Fuzzy GT	threshold at 0.5	Fuzzy GT	threshold at 0.5	
A	0.906	0.008	0.838	0.012	0.907	0.012	0.959	0.005	0.960	0.004	0.808	0.021	0.786	0.026	55
B	0.760	0.002	0.623	0.002	0.929	0.002	0.958	0.003	0.960	0.003	0.833	0.004	0.823	0.004	55
C	0.873	0.004	0.871	0.004	0.934	0.002	0.959	0.003	0.962	0.003	0.813	0.004	0.770	0.004	55
D	0.867	0.003	0.867	0.004	0.931	0.002	0.959	0.003	0.961	0.003	0.833	0.004	0.823	0.004	55
E	0.820	0.003	0.870	0.004	0.930	0.002	0.960	0.003	0.962	0.003	0.827	0.004	0.828	0.005	55
F	0.870	0.003	0.904	0.003	0.931	0.002	0.958	0.003	0.960	0.003	0.827	0.004	0.818	0.004	55
G	0.778	0.002	0.748	0.002	0.831	0.001	0.952	0.002	0.960	0.002	0.777	0.003	0.747	0.003	55
H	0.784	0.002	0.787	0.002	0.860	0.001	0.971	0.004	0.973	0.003	0.806	0.003	0.787	0.003	69
I	0.746	0.001	0.790	0.002	0.866	0.001	0.972	0.004	0.974	0.004	0.784	0.003	0.776	0.003	69
J	0.784	0.002	0.785	0.002	0.860	0.001	0.971	0.004	0.975	0.004	0.806	0.003	0.787	0.003	69
K	0.781	0.002	0.744	0.002	0.846	0.001	0.966	0.004	0.973	0.003	0.803	0.003	0.777	0.003	69
L	0.682	0.002	0.649	0.001	0.821	0.001	0.941	0.003	0.962	0.003	0.765	0.002	0.738	0.002	69
std.	0.062	0.065	0.085	0.086	0.042	0.044	0.008	0.009	0.006	0.006	0.021	0.025	0.028	0.033	

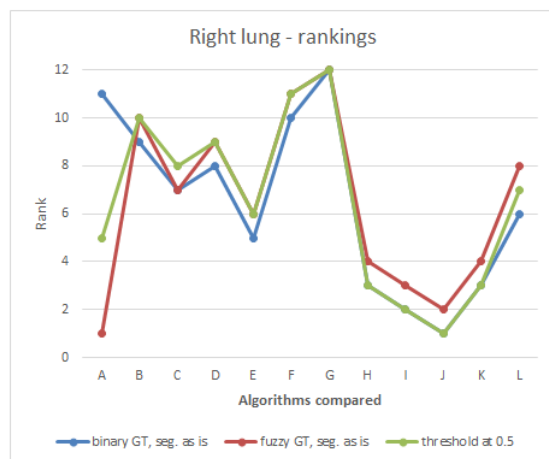
(A)

	Ranking			Ranking			Ranking			Ranking			Ranking			Ranking												
A	1	1	1	0.000	5	5	5	0.000	6	6	1	2.357	7	1	4	2.449	11	1	5	4.110	6	1	1	2.357	7	5	1	2.494
B	10	10	10	0.000	12	12	12	0.000	5	5	6	0.471	9	8	9	0.471	9	10	10	0.471	2	3	3	0.471	2	2	3	0.471
C	2	2	2	0.000	2	2	2	0.000	1	1	2	0.471	8	9	8	0.471	7	7	8	0.471	5	6	6	0.471	10	10	10	0.000
D	4	4	4	0.000	4	4	4	0.000	2	2	3	0.471	6	7	7	0.471	8	9	9	0.471	1	2	2	0.471	3	3	4	0.471
E	5	5	5	0.000	3	3	3	0.000	4	4	5	0.471	5	6	6	0.471	5	6	6	0.471	4	5	5	0.471	1	1	2	0.471
F	3	3	3	0.000	1	1	1	0.000	3	3	4	0.471	10	10	10	0.000	10	11	11	0.471	3	4	4	0.471	4	4	5	0.471
G	9	9	9	0.000	9	9	9	0.000	11	11	11	0.000	11	11	11	0.000	12	12	12	0.000	11	11	11	0.000	11	11	11	0.000
H	6	6	6	0.000	7	7	7	0.000	8	8	8	0.000	2	3	2	0.471	3	4	3	0.471	7	7	7	0.000	5	6	6	0.471
I	11	11	11	0.000	6	6	6	0.000	7	7	7	0.000	1	2	1	0.471	10	10	10	0.000	9	9	9	0.000	9	9	9	0.000
J	6	6	6	0.000	8	8	8	0.000	8	8	8	0.000	2	3	2	0.471	1	2	1	0.471	7	7	7	0.000	5	6	6	0.471
K	8	8	8	0.000	10	10	10	0.000	10	10	10	0.000	4	5	5	0.471	3	4	3	0.471	9	9	9	0.000	8	8	8	0.000
L	12	12	12	0.000	11	11	11	0.000	12	12	12	0.000	12	12	12	0.000	6	8	7	0.816	12	12	12	0.000	12	12	12	0.000

(B)



(C)



(D)

Figure 5: (A) Validating segmentations using the DICE in three different combinations of binary/fuzzy segmentations. The standard deviations of the scores are to show the quality variance between the algorithms, and the score variance between the combinations (B) The resulting system ranking. (C) Score details of the right lung as a selected case. (D) The resulting system ranking for the right lung.

D4.5 Result meta-analysis

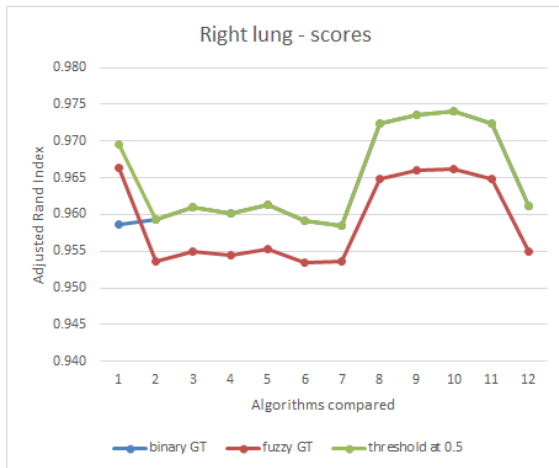
Ranking using the Adjusted Rand Index (ARI) measure in different combinations of binary and fuzzy images

algorithms compared	left kidney 29663		right kidney 29662				liver 58				left lung 1326				right lung 1302				left psoas major muscle 32249				right psoas major muscle 32248				Volume count		
	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (s.d.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (s.d.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (s.d.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (s.d.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (s.d.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (s.d.)					
	A	0.906	0.918	0.925	0.008	0.837	0.849	0.866	0.012	0.905	0.913	0.933	0.012	0.958	0.965	0.969	0.005	0.959	0.966	0.969	0.005	0.808	0.826	0.858	0.021	0.786		0.804	0.847
B	0.760	0.755	0.760	0.002	0.622	0.618	0.622	0.002	0.927	0.924	0.927	0.002	0.957	0.951	0.957	0.003	0.959	0.954	0.959	0.003	0.833	0.825	0.833	0.004	0.823	0.814	0.823	0.004	55
C	0.873	0.865	0.873	0.004	0.870	0.862	0.870	0.004	0.933	0.928	0.933	0.002	0.957	0.951	0.957	0.003	0.961	0.955	0.961	0.003	0.812	0.805	0.812	0.004	0.769	0.761	0.769	0.004	55
D	0.867	0.860	0.867	0.003	0.867	0.859	0.867	0.004	0.930	0.926	0.930	0.002	0.958	0.952	0.958	0.003	0.960	0.954	0.960	0.003	0.833	0.825	0.833	0.004	0.823	0.814	0.823	0.004	55
E	0.820	0.813	0.820	0.003	0.870	0.862	0.870	0.004	0.929	0.925	0.929	0.002	0.959	0.953	0.959	0.003	0.961	0.955	0.961	0.003	0.827	0.818	0.827	0.004	0.827	0.818	0.827	0.005	55
F	0.869	0.863	0.869	0.003	0.904	0.897	0.904	0.003	0.929	0.925	0.929	0.002	0.957	0.951	0.957	0.003	0.959	0.954	0.959	0.003	0.827	0.819	0.827	0.004	0.818	0.809	0.818	0.004	55
G	0.778	0.773	0.778	0.002	0.748	0.743	0.748	0.002	0.829	0.826	0.829	0.001	0.951	0.946	0.951	0.002	0.958	0.954	0.958	0.002	0.777	0.772	0.777	0.003	0.747	0.741	0.747	0.003	55
H	0.784	0.780	0.784	0.002	0.787	0.783	0.787	0.002	0.858	0.855	0.858	0.001	0.970	0.962	0.970	0.004	0.972	0.965	0.972	0.004	0.805	0.799	0.805	0.003	0.787	0.780	0.787	0.003	69
I	0.746	0.743	0.746	0.001	0.790	0.786	0.790	0.002	0.863	0.861	0.863	0.001	0.971	0.963	0.971	0.004	0.974	0.966	0.974	0.004	0.784	0.778	0.784	0.003	0.776	0.770	0.776	0.003	69
J	0.784	0.780	0.784	0.002	0.784	0.780	0.784	0.002	0.858	0.855	0.858	0.001	0.970	0.962	0.970	0.004	0.974	0.966	0.974	0.004	0.805	0.799	0.805	0.003	0.787	0.780	0.787	0.003	69
K	0.781	0.777	0.781	0.002	0.744	0.740	0.744	0.002	0.844	0.841	0.844	0.001	0.966	0.958	0.966	0.004	0.972	0.965	0.972	0.004	0.803	0.797	0.803	0.003	0.777	0.771	0.777	0.003	69
L	0.682	0.678	0.682	0.002	0.649	0.646	0.649	0.001	0.818	0.815	0.818	0.001	0.940	0.934	0.940	0.003	0.961	0.955	0.961	0.003	0.765	0.760	0.765	0.002	0.737	0.732	0.737	0.002	69
std.	0.062	0.063	0.065		0.085	0.084	0.086		0.042	0.043	0.044		0.009	0.008	0.009		0.006	0.006	0.006		0.021	0.021	0.025		0.029	0.028	0.033		

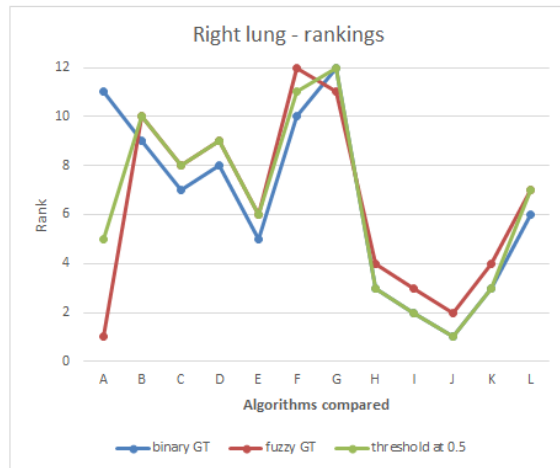
(A)

	Ranking			Ranking			Ranking			Ranking			Ranking			Ranking								
A	1	1	1	0.000	5	5	5	0.000	6	6	6	2.357	7	7	7	4.110	11	11	11	2.357	7	7	7	2.494
B	10	10	10	0.000	12	12	12	0.000	5	5	5	0.471	9	9	9	0.471	9	10	10	0.471	2	3	3	0.471
C	2	2	2	0.000	2	2	2	0.000	1	1	1	0.471	8	8	8	0.471	7	8	8	0.471	5	6	6	0.000
D	4	4	4	0.000	4	4	4	0.000	2	2	2	0.471	6	7	7	0.471	8	9	9	0.471	1	2	2	0.471
E	5	5	5	0.000	3	3	3	0.000	4	4	4	0.471	5	6	6	0.471	5	6	6	0.471	4	5	5	0.471
F	3	3	3	0.000	1	1	1	0.000	3	3	3	0.471	10	10	10	0.000	10	12	11	0.816	3	4	4	0.471
G	9	9	9	0.000	9	9	9	0.000	11	11	11	0.000	12	11	11	0.471	11	11	11	0.000	11	11	11	0.000
H	6	6	6	0.000	7	7	7	0.000	8	8	8	0.000	2	3	3	0.471	3	4	3	0.471	7	7	7	0.000
I	11	11	11	0.000	6	6	6	0.000	7	7	7	0.000	1	2	1	0.471	2	3	2	0.471	10	10	10	0.000
J	6	6	6	0.000	8	8	8	0.000	8	8	8	0.000	2	3	2	0.471	1	2	1	0.471	7	7	7	0.000
K	8	8	8	0.000	10	10	10	0.000	10	10	10	0.000	4	5	5	0.471	3	4	3	0.471	9	9	9	0.000
L	12	12	12	0.000	11	11	11	0.000	12	12	12	0.000	12	12	12	0.000	6	7	7	0.471	12	12	12	0.000

(B)



(C)



(D)

Figure 7: Validating segmentations using the Adjusted Rand Index (ARI) in three different combinations of binary/fuzzy segmentations. The standard deviations of the scores are to show the quality variance between the algorithms, and the score variance between the combinations (B) The resulting system ranking. (C) Score details of the right lung as a selected case. (D) The resulting system ranking for the right lung.

5 Conclusion

We provide analysis on 21 evaluation metrics for medical volume segmentation that have been implemented in the evaluation tool EvaluateSegmentation. We show that the correlation among these metrics gives information about the nature, sensitivities, and bias of these metrics. It is important to take these properties into account in selecting evaluation metrics. In an analysis using manual rankings provided by two radiologists, compared to the rankings produced by the 21 evaluation metrics, we show that the correlation between metric rankings and manual rankings is significantly stronger when using significance tests, since small performance differences are mostly ignored by manual rankers. The automatic metric selection method [2] is performed on the same segmentations that have been ranked manually, to test the efficiency of the method. The results show a moderate correlation of the manual ranking. We show in an analysis on synthetic fuzzy segmentations, generated using smoothing functions, that using binary ground truth to evaluate fuzzy segmentations or the opposite (fuzzy ground truth to evaluate binary segmentation) has a considerable impact on the system ranking, given the systems are similar in their performance. Therefore it is strongly recommended to always use a threshold option, if the segmentations/ground truth are mixed in terms of fuzzy and binary modes. Furthermore, we show that different metrics are differently invariant against fuzzification, i.e. differently sensitive to the combinations of fuzzy/binary volumes.

6 References

- [1] Abdel Aziz Taha and Allan Hanbury. An efficient tool for calculating Medical volume Segmentation Metrics. BMC Medical Imaging. Dec. 2014 (submitted).
- [2] Abdel Aziz Taha, Allan Hanbury, and Oscar Jimenez del Toro. A formal method for selecting evaluation metrics for image segmentation. In 2014 IEEE International Conference on Image Processing (ICIP) (ICIP 2014), pages 932–936, Paris, France, Oct 2014.
- [3] Langs, G., Mueller, H., Menze, B.H., Hanbury, A.: VISERAL: Towards large data in medical imaging - challenges and directions. In: MCBR-CDS MICCAI Workshop, vol. 7723. Nice, France, pp. 92-98 (2013)
- [4] Fatourech, M., Ward, R.K., Mason, S.G., Huggins, J., Schloegl, A., Birch, G.E.: Comparison of evaluation metrics in classification applications with imbalanced datasets. In: ICMLA, pp. 777-782 (2009)
- [5] Fenster, A., Chiu, B.: Evaluation of segmentation algorithms for medical imaging. In: Conf Proc IEEE Eng Med Biol Soc., vol. 7, pp. 7186-7189 (2005)
- [6] Udupa, J.K., LeBlanc, V.R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L.M., Hirsch, B.E., Woodburn, J.: A framework for evaluating image segmentation algorithms. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society 30 (2006)
- [7] Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 9999, 2837-2854 (2010)
- [8] Menze, B., Jakab, A., Bauer, S., Reyes, M., Prastawa, M., Leemput, K.V. (eds.): MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation BRATS2012. MICCAI,(2012).
- [9] Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.A.: Comparing images using the Hausdor distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15, 850{863 (1993)
- [10] Zou, K.H., Wells, W.M., Kikinis, R., Warfield, S.K.: Three validation metrics for automated probabilistic image segmentation of brain tumours. Statistics in Medicine 23 (2004)