Chapter 13

VALIDATION

Concepts

validation of an analysis method estimates correctness of results from tests on a limited number of samples. For carrying out the validation, suitable samples need to be selected, comparison measures have to be chosen that reflect the quality of the result, and a norm is required against which the method is tested. These aspects are realized differently for a delineation task, a detection task, or a registration task.

› Overlap and outlier measures for delineation tasks: over- and under-segmentation, Dice and Jaccard coefficient, Hausdorff distance;

› The ROC curve; Success in detection: type I and type II errors, sensitivity and specificity, precision and recall rates;

› Measuring registration errors;

› Ground truth: manual delineation, hardware and software phantoms;

› Training and test data;

› Significance: *t*-test and Welsh test; › The *p*-value.

characteristics of a validation procedure

An analysis method needs to be tested with respect to the quality with which results are achieved. Unfortunately, a direct way to prove the quality does not exist, since the subjects of analysis are not accessible, except by indirect means. It results in a number of characteristics of a validation procedure:

- Validation is carried out in a statistical fashion, because the quality of an analysis procedure is tested on a number of sample cases. The outcome is assumed to be representative for all cases.
- Validation is relative, as it gives information about the current method with respect to some other way to generate the analysis results.
- Validation is indirect by comparing features that a correct analysis method should produce instead of comparing analysis methodologies.

documentation of the validation scenario

documentation of the validation scenario is an integral part of any validation. It gives a potential user of a method the chance to judge whether the validation is appropriate.

For image-guided surgery that requires segmentation, registration as well as object detection, a number of characteristics that should be addressed and documented:

- Accuracy measures the deviation of results from known ground truth.
- Precision and reproducibility measure the extent to which equal or similar input produces equal or similar results.
- Robustness characterizes the change of analysis quality if conditions deviate from assumptions made for analysis (e.g., when noise level increases or if object appearance deviates from prior assumptions).
- Efficiency describes the effort necessary to achieve an analysis result.
- Fault detection is the ability to detect potential false results during application of an analysis method.

validation scenario

The documentation of the validation scenario builds on the description of the analysis method.

Given the intended use of an analysis method and assumptions about the data, the description of a validation scenario should contain the following information:

- Description of the data on which the validation is to be carried out
- Description and justification of what is assumed to be the ground truth
- Criteria by which the quality is to be measured
- Definition and justification of what constitutes a successful validation.

Measures of Quality

- Quality depends on the kind of analysis that has been carried out on the data.
- If an object is delineated, it will be a correspondence measure between the delineated object and some reference segmentation.
- If the task was object detection, it will be a ratio between correct and incorrect decisions.
- If it has been a registration, it will be the deviation from the correct registration transformation.
- Since an analysis task may involve combinations of these methods, the kind of validation procedure will have to be selected based on the intended application.

Quality for a Delineation Task

- Measuring quality for a delineation task requires a measure of correspondence between delineated structure f and some known, true reference delineation g (see Fig. 13.1).
- For now, let us assume that ground truth data exists. Hence, for an image consisting of voxels or pixels v, a function g is known with

$$g(\mathbf{v}) = \begin{cases} 1, & \text{if } \mathbf{v} \text{ belongs to the delineated object} \\ 0, & \text{otherwise.} \end{cases}$$

Correspondence can be measured in different ways:

- **Volumetric measurements compute volume or area differences between f and g.**
- **Overlap measures compute the overlap between object and background elements in f and g.**
- **Distance measurements compute the deviation between the boundaries of f and g.**
- **Outlier measures compute the maximum deviation between f and g.**

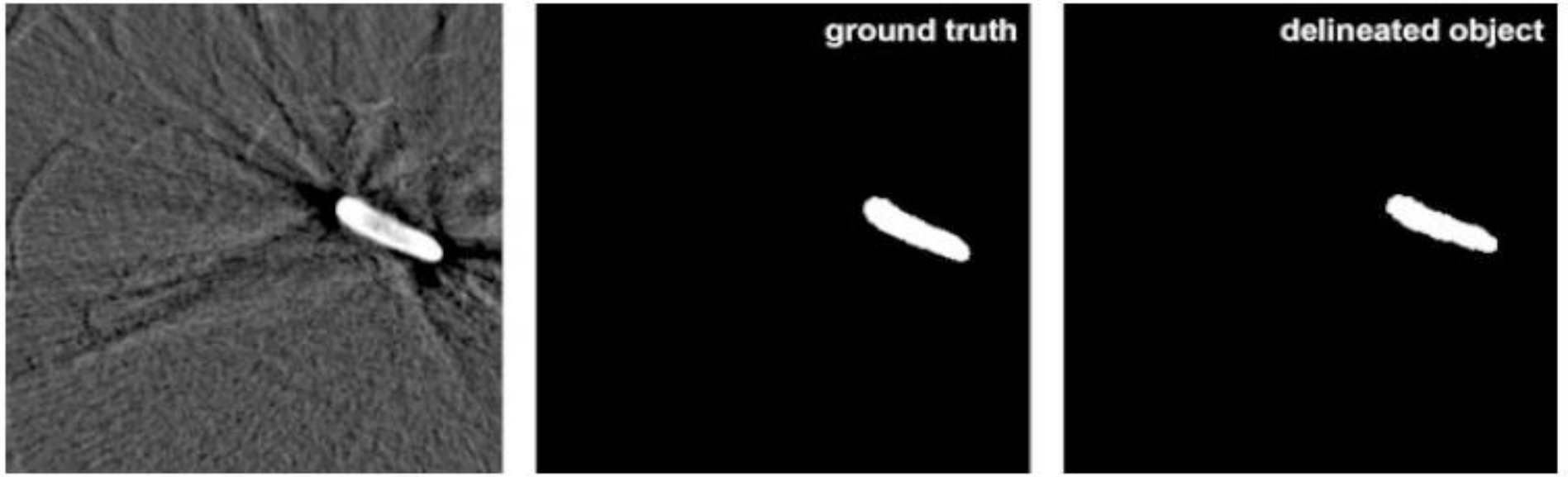


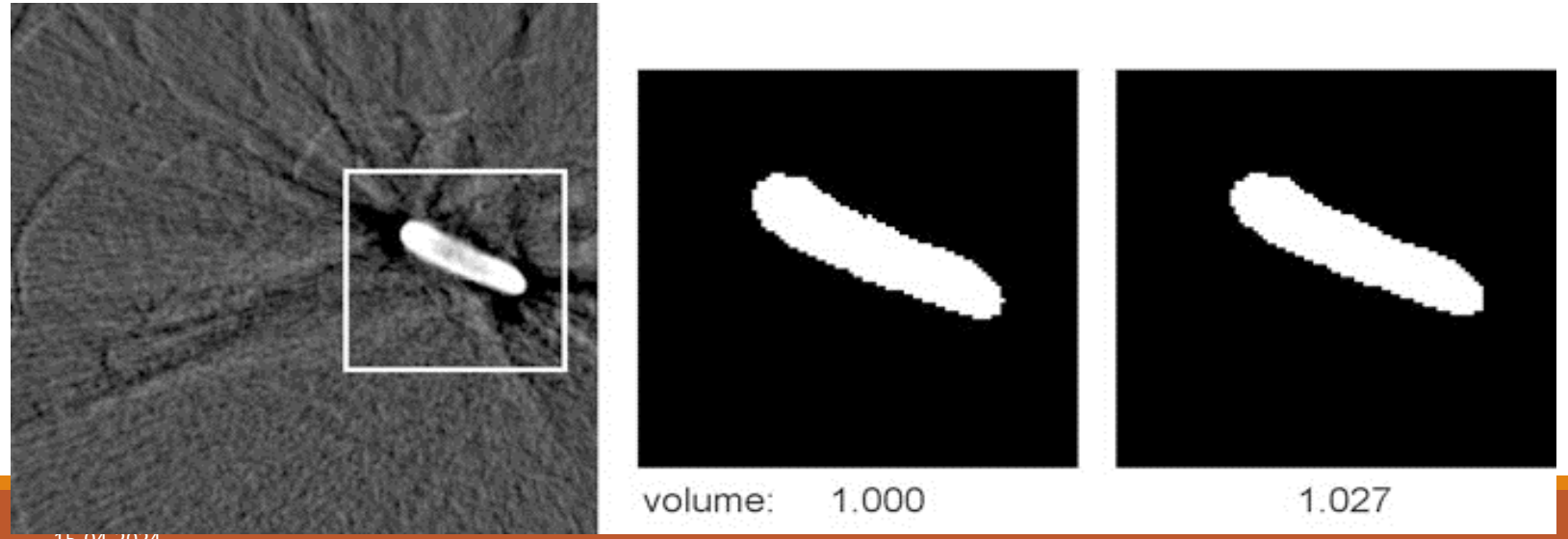
Fig. 13.1 Validation of a delineation requires a measure of comparison between some reference (the often so-called ground truth) and the delineated object

Volumetric measurements

- A volumetric measurement only compares a property derived from the delineation.
- An overlap measure computes an element-wise correspondence.
- Distance measurements do the same with corresponding boundary points. They may return substantially different results than an overlap measure, since a large change of boundary detail not necessarily results in an equally large change of volume.
- Finally, outlier measurements capture singular deviations between f and g. Selecting a type of quality measure depends on the application.
- **Volumetric measurements are easy to compute. They simply count the number of elements |F| and |G| with F = {v|f(v) = 1} and G = {v|g(v) = 1} weighted by the area or volume covered by each scene element (pixel or voxel) v**

Overlap measures

- Overlap measures compute the amount of over-segmentation, under-segmentation, or a combination of the two (see Fig. 13.3).
- **Over-segmentation is simply the number of elements v, for which g(v) = 0 and f(v) = 1, while under-segmentation is the opposite, i.e., g(v) = 1 and f(v) = 0.**
- In order to make these numbers comparable for objects of different size, the amounts are usually given as percent over- or under-segmentation.



The volume difference is the simplest FOM for comparing a delineation with the ground truth. It does, however, not account for shape differences

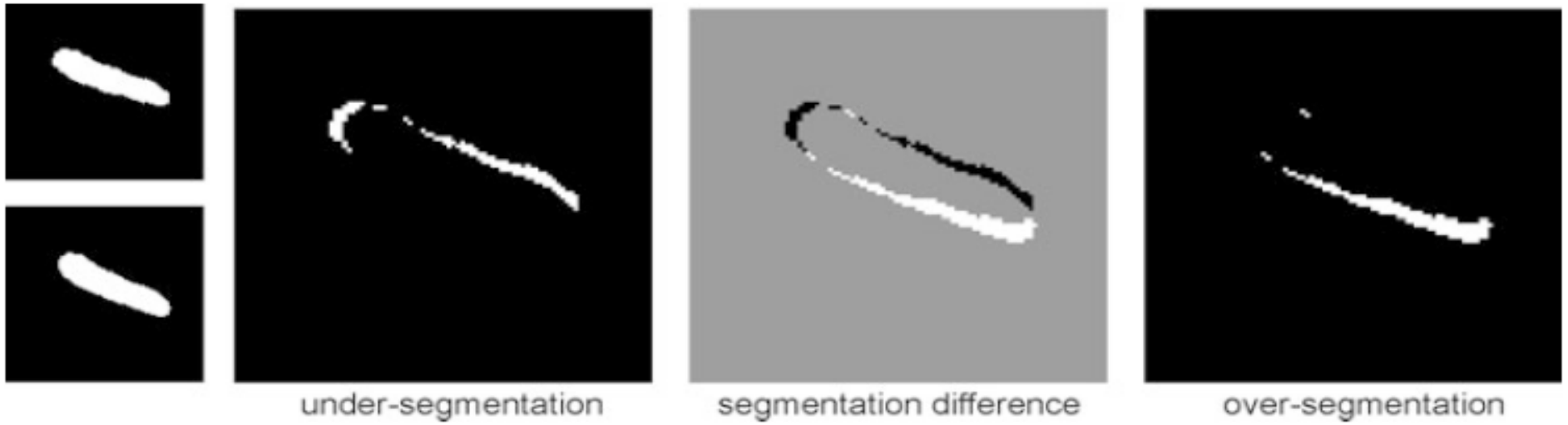


Fig. 13.3 Over- and under-segmentation as well as a combination of the two represent differences in shape between ground truth and the segmentation to be tested

- For compactness of presentation, results from the two measures may be combined. Two
- often used quality measures are the Dice coefficient and the Jaccard coefficient . They are borrowed from statistics where they are used to rate the similarity of data sets.

Dice coefficient and Jaccard coefficient

- The Dice coefficient (Dice 1945) is defined as where $F \cap G$ as the set of all elements v with f(v) = 1 and g(v) = 1.

$$d = \frac{2|F \cap G|}{|F| + |G|},$$

- The coefficient is 1, if the correspondence is perfect, and smaller than 1 otherwise.
- The Dice coefficient was found to agree well with perceived variability of results of a segmentation in Zou et al. (2004).

Jaccard index

The Jaccard index (Jaccard 1912) is given by

$$j = \frac{|F \cap G|}{|F \cup G|},$$

where F ∪ G is the set of all elements v with f(v) = 1 or g(v) = 1.

Again, this coefficient is 1, if the correspondence is perfect, and decreases otherwise.

The Jaccard index is also known as Tanimoto coefficient on sets.

Outliers cannot be measured by the criteria listed above although they may sometimes be critical. An example would be a task where organ boundaries are to be delineated as part of access planning in minimally invasive surgery. In such case, the maximum deviation of the delineated boundary from the true boundary is an important quality. It can be measured by the *Hausdorff distance* between the two data sets F and G (used, e.g., in Chalana and Kim (1997), Gerig et al. (2001), see Fig. 13.4). The Hausdorff distance h is the maximum of all minimal distances d between points in F and points in G:

$$h = \max\left(\inf_{f \in F} d(f, G), \inf_{g \in G} d(g, F)\right) \tag{13.7}$$

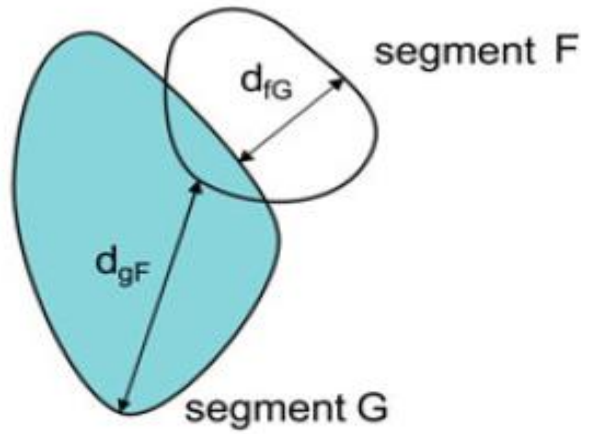


Fig. 13.4 The Hausdorff distance is the maximum of all shortest distances between points of one set to the other set

Sample problems

Problem 1: Suppose you have two binary images representing ground truth and predicted segmentation masks for a specific object. The ground truth mask has 120 pixels labeled as the object, and the predicted mask has 110 pixels labeled as the object. The intersection between the ground truth and predicted masks (pixels labeled as the object in both) is 100 pixels. Calculate the Dice and Jaccard coefficients.

Solution: Dice coefficient is given by the formula:

$$Dice = 2 \times Intersection \; / \; Ground\ Truth + Predicted$$

Given:

- Intersection = 100 pixels, Ground Truth = 120 pixels, Predicted = 110 pixels

Substituting the values:

$$Dice = 2 \times 100 \; / \; 120 + 110 = 200/230 \approx 0.8696$$

Solution: Jaccard coefficient is given by the formula:

Jaccard=Intersection/ Union

Given:

- Intersection = 100 pixels (same as before), Ground Truth = 120 pixels (same as before), Predicted = 110 pixels (same as before)

Union=Ground Truth+Predicted−Intersection=120+110−100=130 Substituting the values:

Jaccard=100/ 130≈0.7692

Problem: Compute the Hausdorff distance between two sets of points in a 2D plane.

Sets of Points:

- Set A: {(1, 2), (3, 4), (5, 6), (7, 8)}
- Set B: {(2, 2), (4, 4), (6, 6), (8, 8)}

Solution: The Hausdorff distance between two sets A and B is defined as the maximum of the minimum distances from each point in set A to the nearest point in set B, and vice versa.

To compute the Hausdorff distance, we need to find:

1. For each point in set A, calculate its distance to the nearest point in set B.
2. For each point in set B, calculate its distance to the nearest point in set A.
3. Take the maximum of the minimum distances obtained from steps 1 and 2.

1.Calculate distances from each point in set A to the nearest point in set B:

1. Distance from (1, 2) to nearest point in set B: √((1-2)^2 + (2-2)^2) = 1
2. Distance from (3, 4) to nearest point in set B: √((3-4)^2 + (4-4)^2) = 1
3. Distance from (5, 6) to nearest point in set B: √((5-4)^2 + (6-4)^2) = √5
4. Distance from (7, 8) to nearest point in set B: √((7-8)^2 + (8-8)^2) = 1

2.Calculate distances from each point in set B to the nearest point in set A:

1. Distance from (2, 2) to nearest point in set A: √((2-1)^2 + (2-2)^2) = 1
2. Distance from (4, 4) to nearest point in set A: √((4-3)^2 + (4-4)^2) = 1
3. Distance from (6, 6) to nearest point in set A: √((6-5)^2 + (6-6)^2) = 1
4. Distance from (8, 8) to nearest point in set A: √((8-7)^2 + (8-8)^2) = 1

3.Take the maximum of the minimum distances obtained:

1. Maximum of {1, 1, √5, 1} = √5

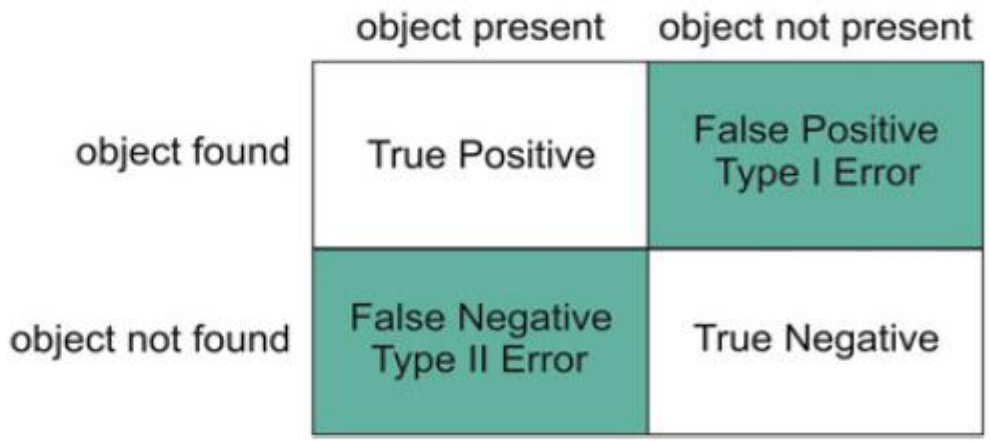
So, the Hausdorff distance between set A and set B is √5.

Quality for a Detection Task

Given two sets T and F of objects in the ground truth data, where T consists of objects to be detected and F consists of all other objects, and two types p and n of decisions of the detector, where p is a positive decision that the object is detected and n is the negative decision, four different cases arise (see Fig. 13.6):

- True positive detections TP are those belonging to T, which have been rightly detected by a positive decision p.
- True negative detections TN are those belonging to F, which are rightly resulted in a negative decision n.
- False positive results FP are those that do belong to F but resulted in a positive decision p.
- False negative results FN are those belonging to T but were classified as n.

Fig. 13.6 A detection task may result in two different kinds of error



Listing detection results organized as in Fig. 13.6 produces the **confusion matrix** for a two-class classification problem. This is also a common means to convey classification results in multi-label detection tasks, as it quickly reveals important characteristics of the tested detector.

A good detection method would produce as many TP and TN as possible. However, since false positive results (e.g., a tumor is detected while no tumor is present) and false negative results (e.g., a tumor is overlooked) may have wildly different consequences, the two types of error are measured separately. The former is called a *type-I error* and the latter is a *type-II error*. Since the absolute numbers of FP or FN detections do not carry much information, they are normalized with the number of cases tested. This is expressed by the *sensitivity* and *specificity* of a method. Sensitivity Sv is defined as

$$\mathrm{Sv} = \frac{\mathrm{TP}}{T} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{13.9}$$

It represents the rate of positive detections with respect to all elements of T. Sensitivity tells how likely it is to miss a detection by the analysis method (in the information retrieval community it is also called *recall rate Rc*). The specificity Sp is defined as

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{13.10}$$

It represents the rate of negative decisions with respect to all elements of N. Specificity tells how likely it is that the detection method produces a false alarm. In the information retrieval community this is replaced by a different measure, called *precision rate* Pr of the retrieval

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{13.11}$$

It reflects a different view at the detection task as it measures the amount of "noise" (the false positives) which is generated by the detection algorithm.

Sensitivity and specificity as well as precision and recall are separate measures of the type-I and the type-II error. If the quality of a result shall be described by a single scalar, they can be combined. An often used measure for this is the F_β-score. It is defined as

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{Pr} \cdot \mathrm{Rc}}{\beta^2 \cdot \mathrm{Pr} + \mathrm{Rc}} = \frac{(1 + \beta^2) \cdot \mathrm{TP}}{(1 + \beta^2) \cdot \mathrm{TP} + \beta^2 \cdot \mathrm{FN} + \mathrm{FP}}. \tag{13.12}$$

receiver operator characteristic (ROC)

it measures the performance of a parametrizable detector, it is also used as a quality measure for determining its performance. It is called receiver operator characteristic (ROC) and measures the ratio of sensitivity versus specificity for each parameter setting a:

$$\mathrm{ROC}(\alpha) = (\mathrm{Sv}(\alpha) \quad 1 - \mathrm{Sp}(\alpha)).$$

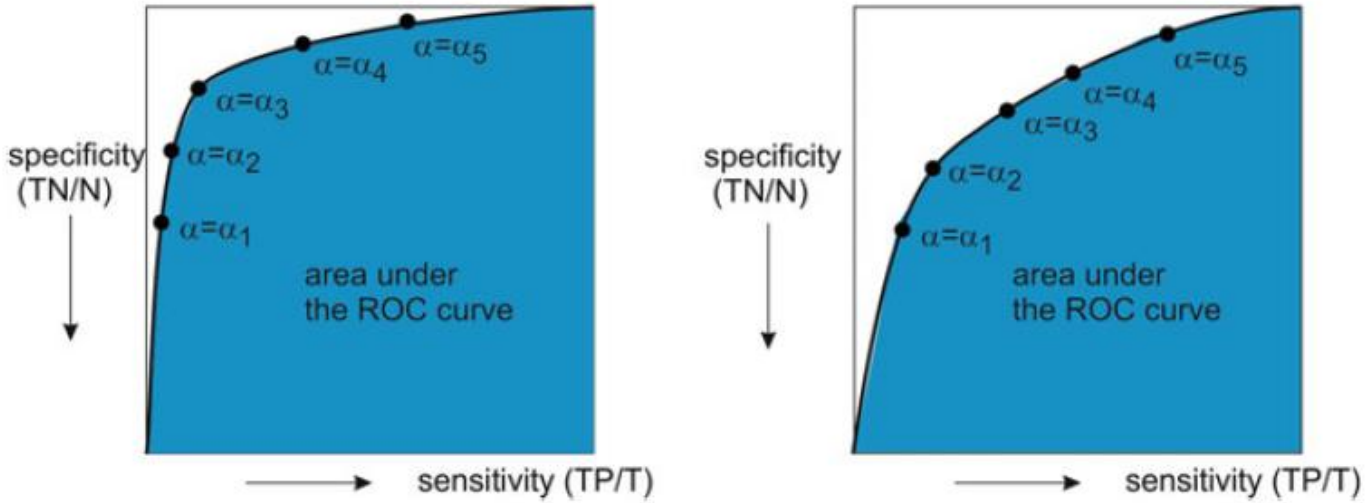


Fig. 13.7 The ROC curve rates sensitivity versus specificity for different parameter settings of a detection method. The method described by the curve on the right is inferior to the method described by the curve on the *left*, since specificity decreases faster with increased sensitivity

This ROC curve represents the performance of the detector independent of the parameterization a. Hence, it is a way to compare two different, parametrizable detectors. An ideal ROC curve would produce sensitivity and specificity of 100% irrespective of the value of a. Since the diagram is normalized with Sv and 1 − Sp ranging from 0 to 1, the area under this ROC curve would be 1.0. The worst result would be an ROC curve, where an increase of Sv would cause an equal increase of 1 − Sv. This would mean that increasing the number of correct answers would increase the number of false answers in the same fashion. In this case, the ROC curve would be a diagonal and the area under the curve would be 0.5. Computing the area under the ROC curve allows comparing two different, parameterizable detectors.

Quality for a Registration Task

- Quality of a registration task is measured by direct or indirect measurement of the difference between true and computed registration transformation.
- Direct comparison requires the true transformation parameters to be known (which is sometimes possible when phantom data is used for providing the ground truth).
- Average deviation of transformation parameters plus detection of outliers is then the appropriate means to describe how well the registration method succeeded. If the true transformation is unknown, an indirect way to compute a quality measure is to exchange moving and still image . The transformation should be the inverse and any deviations are taken as inaccuracies in the computation of the registration transformation.
- An indirect way, which is often used for computing validity of a registration transformation, is to compute deviations of point locations after registration. If a number of point pairs are known that represent semantically equivalent locations in the two images, applying the registration transformation to these point locations in one image should map them exactly on locations of their counterparts in the other image. The **point pairs are called fiducial markers**. Of course, the point pair correspondence must not have been used for computing the registration transformation.

- Using fiducial markers has the advantage that it evaluates a property that is usually the reason for computing the registration transformation in the first place.

- One has to be aware, however, that the quality of this mapping is evaluated only at the locations of the fiducial markers (see Fig. 13.10). This may become a problem, if markers are placed in locations that are very different from those where a high accuracy is required (e.g., when only using skin surface markers to test a registration of brain images), or if the registration transformation has many degrees of freedom and only few marker positions are used (e.g., when evaluating a non-rigid registration transformation with few markers).

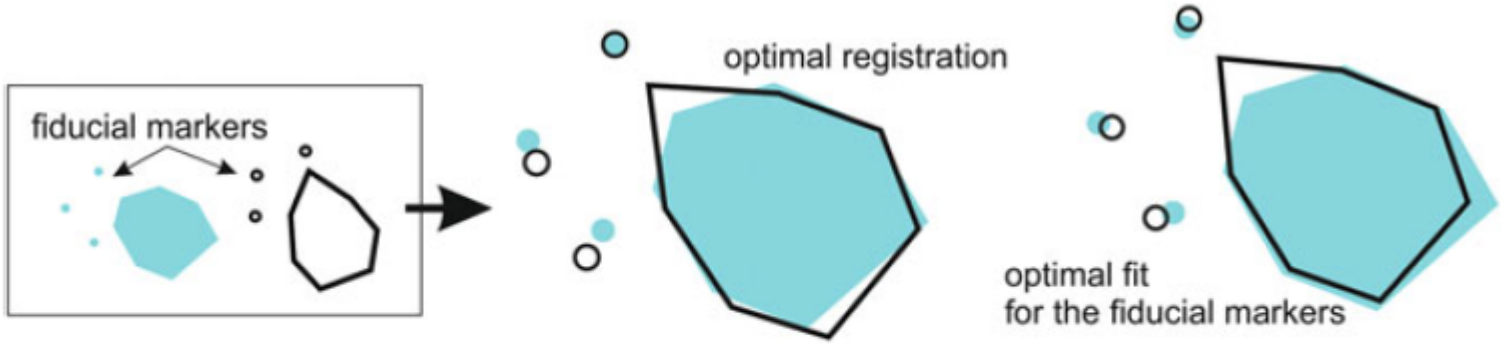


Fig. 13.10 Using fiducial markers to validate a registration may yield unexpected results if the markers are far away from the objects to be registered, if they have localization errors, or if they are too few given the number of degrees of freedom for a registration. The optimal fit for fiducial markers does then not automatically mean that the registration itself is optimal

The Ground Truth

- All measures to estimate the quality of an analysis procedure require comparison of the analysis result with the true information.
- The truth is difficult to come by, since the reason for producing images in the first place was to gather information about the human body that cannot be accessed otherwise.
- Ground truth is the true analysis result of data to which the analysis method is applied. The data can be real or artificial.
- For automated analysis, the observers will be replaced by an automatic scoring system.
- Under the assumption that data experts produce unbiased variations of an unknown ground truth (an
- assumption which also gave rise to the **STAPLE procedure**, the method is able to rate a new method correctly even if the exact ground truth is unknown.

Ground Truth from Real Data

- Ground truth from real data results from applying the currently established best method to it. This is often difficult to determine, however. If an established method exists at all. Another problem is that the implementation of the established state-of-the-art method is not always available.
- If a currently best method does not exist, analysis by a human expert is an option to produce ground truth data. It requires some effort on the developer's side and a lot more effort on the data expert's side.
- Intra-observer variability can be estimated that describes variation of judgment by the same individual. A high variability may indicate that the knowledge on which decisions are based is either not very clear, not sufficient, or cannot be applied easily.
- Different experts may have different opinions about the content in the data. Hence, inter-observer variability should be measured as well by asking several experts to analyze the data.

STAPLE

- STAPLE (simultaneous truth and performance level estimation) (Warfield et al. 2004) is an elegant way to solve this problem for delineation and object detection tasks.
- The ground truth is assumed to be a set of labeled voxels and the input is a labeling from different experts (or segmentation algorithms) whose reliability is unknown.
- STAPLE defines the problem probabilistically and treats expert segmentations as samples that deviate in an unknown fashion from the ground truth.
- It reconstructs the most likely ground truth data given the samples. This, in turn, defines the reliability of each sample.
- Hence, STAPLE is an application of a maximum likelihood (ML) algorithm to the sample results.

The STAPLE method for a two-label segmentation (e.g., foreground vs. background) presumes that a number of R segmentations exist that assign a binary label to each of N scene elements. It estimates the true segmentation $\mathbf{t} = (t_1, \ldots, t_N)$ from the R segmentations $\mathbf{d}_j$, $j = 1, R$ and $\mathbf{d}_j = (d_{j,1}, \ldots, d_{j,N})$. It further estimates sensitivity $\mathbf{p} = (p_1, \ldots, p_R)$ and specificity $\mathbf{q} = (q_1, \ldots, q_R)$ for each of the R segmentations. Sensitivity and specificity of STAPLE relates to the number of classified voxels or pixels. If one of the R segmentations is new and shall be tested against the other segmentations, sensitivity and specificity of the new method can be used to rate it against the other methods. Other quality measures such as the Dice coefficient can be computed based on $\mathbf{t}$.

Ground Truth from Phantoms

- Phantoms will exhibit varying degrees of realism. The more realistic the phantom is, the less accessible is the information represented by the phantom. Selecting the appropriate phantom is always a compromise.
- Based on the degree of realism, **phantoms fall into four groups:**
- • Cadaver phantoms (human or animal).
- • Artificial hardware phantoms constructed of material that is known to produce a similar image signal than real data.
- • Software phantoms representing the imaged measurement distribution.
- • Software phantoms representing the reconstructed image.

Cadaver

In survey on image segmentation methods in medical imaging, classified phantoms of the first **two kinds as physical phantoms and the latter two as computational phantoms.** The main difference is that a physical phantom is imaged by the same imaging device than the patient data, while for the latter influences from imaging have to be simulated.

- A cadaver or an animal specimen has similar properties than real patient data. Since image acquisition is equal to that of generating patient data, the following attributes of the imaging procedure are represented by the phantom:

- • Material properties • Measurement properties

  • Influences from image reconstruction • Shape properties of the imaged object.

- In order to make the phantom data useful for validation analysis, results have to be generated in the phantom. That means, • for a detection task that number and locations of objects to be detected have to be specified.

  • for a registration task that fiducial markers have to be implanted that are visible in the images, or that the phantom is fixed to a reference frame from which transformation parameters can be deduced.

- Some problems may arise when using such a phantom for validation:
- • Attributes of the image phantom may still vary in an unknown fashion from that of the real data.
- • Anatomical variability (normal as well as pathological) is not captured by the phantom.
- **Artificial hardware phantoms** capture all but the material properties. Material properties are simulated by selecting material that produces a similar signal than the tissues to be imaged (see Fig. 13.12).
- Since the artificial structure is built on purpose, aspects for validation (extent of an object, specific landmarks, delineation of object boundaries) are usually known and accessible.
- Simple hardware phantoms just capture local properties to be measured by the imaging system

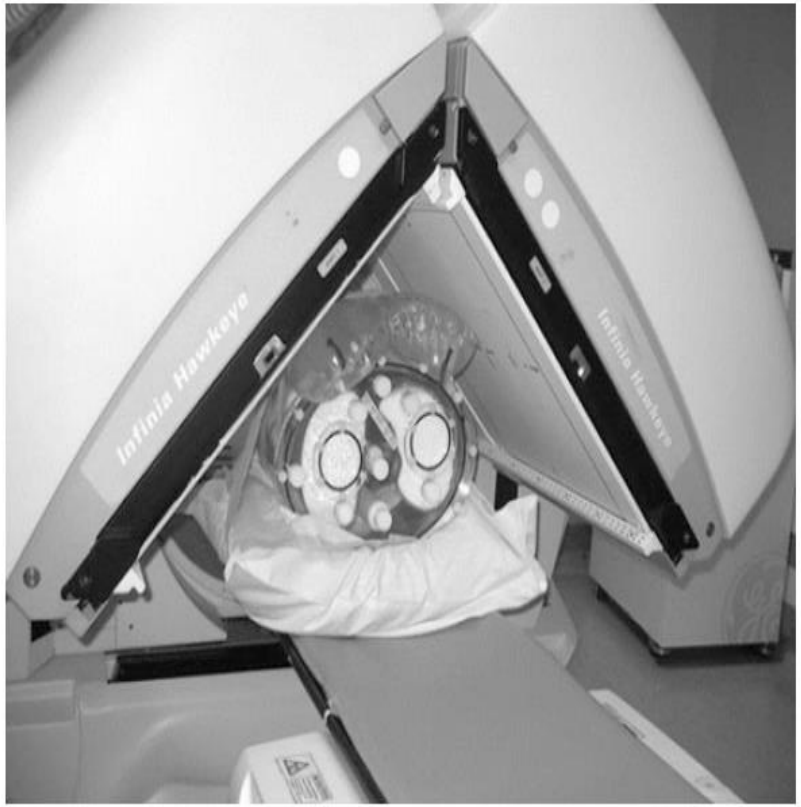


Fig. 13.12 Example of an artificial phantom. The picture shows a hardware phantom for th simulation of SPECT images of the heart. The phantom consists of several bottles o photon-emitting fluids that have a similar characteristic than the organs that are simulated. Th shape of the bottles resembles (at the spatial resolution of SPECT imaging) the shape of the organ (with kind permission of Anna Celler, Department of Radiology, University of British Columbia Vancouver)

Software phantoms

- Software phantoms representing the measured image signal are the next level of abstraction.
- The phantom consists of a 3d distribution of the signal to be measured by the image acquisition system (e.g., X-ray absorption for an X-ray CT phantom). The 3d signal distribution may be generated from interpreted image material [e.g., the MRI BrainWeb phantom].
- A number of software phantoms of this kind exist:
- • The BrainWeb phantom, www.bic.mni.mcgill.ca/brainweb/,
- The Field II ultrasound simulation program , to be found at http://server.electro.dtu.dk/personal/jaj/field/,
- The group of Segars et al., http://www.bme.unc.edu/*wsegars/index.html and http://dmip1.rad.jhmi.edu/xcat/.
- The advantage of a software phantom compared to a hardware phantom is that it easily allows inclusion of known anatomical variation by creating several different shape phantoms. Also, the original properties of the objects to be represented are easily accessible.

Representativeness of Data

- besides trusting ground truth and the appropriateness of the quality measures, representativeness is another issue for validation.
- **Using representative data means that all data properties potentially influencing the performance of the analysis method are reflected in the test data.**
- A couple of strategies strengthen the argument of representativeness:
- • Separation between test and training data
- • Identification of sources of variation
- • Identification of outliers
- • Investigation of robustness with respect to parameter variation.

Separation Between Training and Test Data

- the set of data that can be used for training and testing is very small.
- Separating it into two even smaller subsets would further reduce the significance of results.
- A workaround is to use the **leaving-one-out technique** (also called jackknife technique, for an example).
- A data set of N elements, for which the ground truth is known, is separated in a training set of N − 1 elements and a test set consisting of just one element.
- Parameter estimation is done on the training set and the quality is then measured on the single test element. This is done for all N subsets of N − 1-element training sets and 1-element test sets. The overall quality is then computed by combining the N different test results. The leaving-one-out technique is a special case of the leaving-n-out technique, where n instead of just one sample is left out.

Identification of Sources of Variation and Outlier Detection

- Representativeness means that typical sources of variation in the data are covered by the ground truth data. In order to test the behavior in extreme cases, outliers should be identified.
- If data consists of patient data that has been analyzed by the current best method or by a human expert, sources of variation and outliers can be found with the help of an expert.
- Another source of misunderstanding regarding variation and outliers is the difference of the task for the expert and for a computer method.
- Given N data sets, cross-validation computes predictions of these properties from all subsets of N − 1 elements and applies this prediction to the remaining data set.
- The N − 1 elements would be a sufficiently high number to generate an estimate of a likelihood function.

Robustness with Respect to Parameter Variation

- Finding parameters and testing the method with respect to changes of parameter values can be costly because every possible combination of parameters needs to be tested.
- It may become unfeasible, if the number of parameters is large. Hence, it is worthwhile trying to detect parameters that are most likely independent of each other.
- Doing this by simply testing would be unfeasible again. However, if the analysis methodology has been developed with exact documentation about information that is exploited for the analysis and about ways how this is used for carrying out the analysis, it is usually possible to hypothesize which parameters may be independent.

Significance of Results

- Significance of a test result can be computed by estimating the probability that it arose by chance.
- **This percentage is the p-value that is used to describe the significance of the outcome of an experiment**. The statement “the results are significant with $p < 0.01$” simply means that the probability that the results arose by chance is less than 1%.
- Intuitively, significance depends on the number of samples and on the amount of similarity or dissimilarity between the two populations (ground truth vs. test or method A vs. method B).
- **Significance can be tested by the Student’s t-test (also called t-test).**
- It computes the probability whether quality measurements qnew from a new method could be the result of some natural fluctuation of measurements qold that were computed with an earlier method

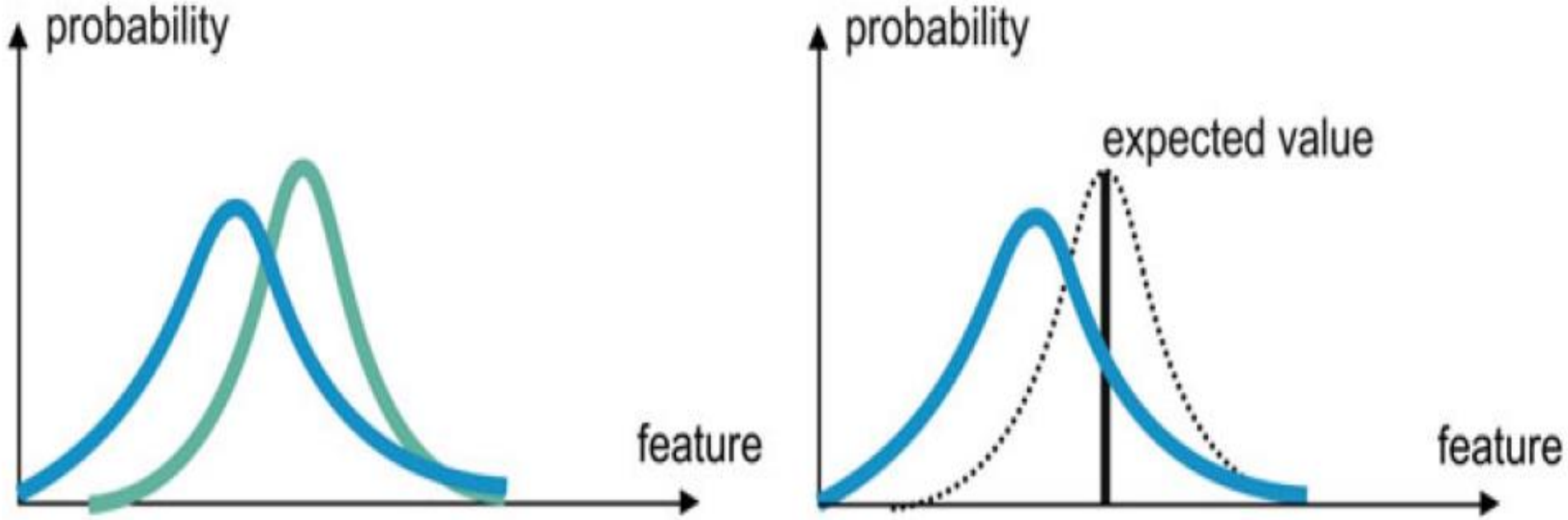


Fig. 13.18 The purpose of the student *t*-test is to determine how likely two distributions of observations belong to the same observation. It is often tested by estimating how likely the expected value of one of the two distributions belongs to the other distribution