Int J CARS (2007) 2:169–181
DOI 10.1007/s11548-007-0125-1

ORIGINAL ARTICLE

Statistical validation metric for accuracy assessment in medical image segmentation

Aleksandra Popovic · Matías de la Fuente · Martin Engelhardt · Klaus Radermacher

Received: 8 March 2007 / Accepted: 28 June 2007 / Published online: 28 July 2007

Abstract
Objective Validation of medical image segmentation algorithms is an open question, considering variance of individual pathologies and the related clinical requirements for accuracy. In this paper, we propose a validation metric capable to distinguish between an over and under-segmentation and account for different clinical applications.
Materials and methods In this paper, we propose a validation metric representing a tradeoff between sensitivity and specificity. The metric has an advantage of differentiating between an over or under-segmentation which is an important feature for validating large sets of segmentation results, as human inspection is exhausting and time consuming. Although it is oriented to the accuracy measurement it is also closely related to the robustness of a method.
Results Features of the metrics are analyzed alongside their medical impact. A set of numerical simulations is performed in order to compare the proposed metric with standardly used discrepancy measures. The metric is illustrated with a clinical case study, presenting accuracy assessment of an algorithm for calvarial tumor segmentation, validated on six patients.

Introduction

Evaluation of medical image segmentation algorithms is an important issue, not only if image processing systems are to be introduced into routine clinical practice, e.g., for image-guided therapy [1], but also during the algorithm design phase. Different evaluation frameworks and criteria have been proposed. Udupa et al. [2,3] define three components in assessment of efficacy of a segmentation method: *precision*, reproducibility or reliability of the segmentation method in presence of measurement variations inherent to the process, *accuracy*, degree to which segmentation result agrees with the real object, and *efficiency*, addressing computer and user time needed to perform an operation. Moreover Jannin et al. [1] proposed *robustness*, measure of performance in presence of disruptive factors, and *fault detection capability*. This paper focuses on the accuracy by introducing a metric for measurement of the discrepancy between an algorithmic segmentation result and the reference segmentation.

An important issue in medical image segmentation validation is a selection of a reference segmentation. A manual segmentation performed by an expert suffers from intra- and inter-rater variability and is not reliable for conclusive results. In the recent years, methods to combine segmentation results from various expert raters have been published and discussed [4–6], resulting in so called latent gold standards. Digital and physical phantoms present further evaluation prospects, although availability of such data-sets is limited to a small number of applications [7,8], due to the difficulties in the modeling of pathologies.

Although numerous validation metrics have been proposed and utilized, adapted from general computer vision techniques or medical tests evaluation, there is no general consensus in medical image processing society which metric(s) are to be used for the validation of segmentation results. Many authors in both computer vision [9] and medical image processing [2,3,10,11] emphasize the need for unification of validation procedures.

A. Popovic (✉) · M. de la Fuente · K. Radermacher
RWTH Aachen University, Chair of Medical Engineering,
Helmholtz-Institute for Biomedical Engineering,
Pauwelsstr. 20, 52062 Aachen, Germany
e-mail: popovic@hia.rwth-aachen.de

M. Engelhardt
Ruhr University Bochum, Clinic for Neurosurgery,
In der Schornau 23–25, 44892 Bochum, Germany

Even though medical image processing is a subset of image processing in general, validation in medicine is substantially different from general image processing since algorithm accuracy has to be considered alongside clinical relevancy and impact. For instance, while some computer vision metrics rate the same amount of over or under-segmentation equivalently, in some cases in medical image segmentation an over-segmentation could lead to the trauma of the sensitive healthy tissue with severe consequences for the patients life. Likewise, in other tumor segmentation applications, under-segmentation could result in a recurrence and a need for another intervention. Therefore, each medical image segmentation application has to be validated not only based on the accuracy of the method itself, but on the impact on clinical outcome and consequently patients' health. Udupa et al. [3] define an *application domain*, which is to be specified for each validation process.

Jannin et al. [12] define a *model of the validation objective*, describing both clinical context and medical objective of the validation study. In this context, we recognize a need for a segmentation validation metric that can account for clinical aspects of discrepancy.

According to Cardoso and Corte-Real [9], empirical validation metrics in image segmentation are classified into two types: goodness methods and discrepancy methods.

Goodness methods validate image-specific properties of the segmented object, e.g., color uniformity, shape, edge quality. However, this approach might not suitable in medical image processing, due to the subjectivity of selected property and the incapability to measure clinical relevance. As further discussed by Cardoso and Corte-Real [9], those properties are used to design algorithms and are not suitable for evaluation.

Discrepancy methods measure the amount of agreement between the gold standard and a segmentation result.

Validation of medical image segmentation is basically addressing three key issues.

- What is the discrepancy between the desirable and the achieved segmentation result?
- What is the clinical impact of this discrepancy?
- How robust is the algorithm in relation to patient anatomy variation and image properties fluctuations?

In order to handle these matters, a medically-oriented metric for comparison between an algorithmic segmentation and the reference has to be developed.

Commonly used validation metrics in medicine, sensitivity and specificity, have drawbacks not only for medical image processing but also for medical tests evaluation. As stated by Kraemer [13]: "Sensitivity and specificity are uncalibrated measures of test quality, measures with a variable zero-point and scale". Different prevalence adjusted discrepancy measures were introduced during the last century, of which Dice similarity coefficient (DSC) [14] is mentioned most frequently. Prastawa et al. [15] validated brain segmentation results using DSC together with surface distances, using segmentation validation tool VALMET [16]. Zou et al. [17,18] proposed a logit transformation of Dice similarity coefficient for better statistical inference. Further used are Jaccard score [19,20], kappa and χ^2 [5] statistics. Those measures are normally referred to as *statistical* metrics since they do not take spatial relations between image elements and/or edges into account. An alternative is to use *geometrical* measures that evaluate spatial distances between the objects. Those measures are either edge or volume oriented. Although widely used for validation in general image processing [21,22], especially for edge-oriented segmentation algorithms, geometrical measures are less often used for medical applications, mostly because they are not sufficiently clinically intuitive, as for instance sensitivity and specificity are.

In this paper a novel metrics for validation of segmentation results is proposed. The focus is on the validation of the segmentation processes that result in a binary decision, i.e., object or background.

Statistical validation metrics: an overview and limitations

Confusion matrix

Statistical validation metrics assume spatial independence between the image elements (to generalize: voxels).

Let $I(x, y, z) : \mathbb{R}^3 \rightarrow \mathbb{R}$ be a 3D medical image, and $S(I(x, y, z)) : \mathbb{R}^3 \rightarrow \Omega$, $\Omega = \{0, 1\}$, be a binary decision segmentation of the image $I(x, y, z)$. If the gold standard segmentation is G and a segmentation result R, each voxel can be classified as followed:

$$\begin{array}{ll} \text{true positive,} & G(x, y, z) = 1 \wedge R(x, y, z) = 1, \\ \text{false positive,} & G(x, y, z) = 0 \wedge R(x, y, z) = 1, \\ \text{true negative,} & G(x, y, z) = 0 \wedge R(x, y, z) = 0, \\ \text{false negative,} & G(x, y, z) = 1 \wedge R(x, y, z) = 0. \end{array}$$

Based on these, statistical parameters are defined: number of true positives (TP), number of false positives (FP), number of true negatives (TN), and number of false negatives (FN). Those values are typically presented in a 2×2 matrix, usually referred as the confusion matrix (Table 1) [23].

Table 1 2 × 2 confusion matrix

		Gold standard	
		+	–
Result	+	TP	FP
	–	FN	TN

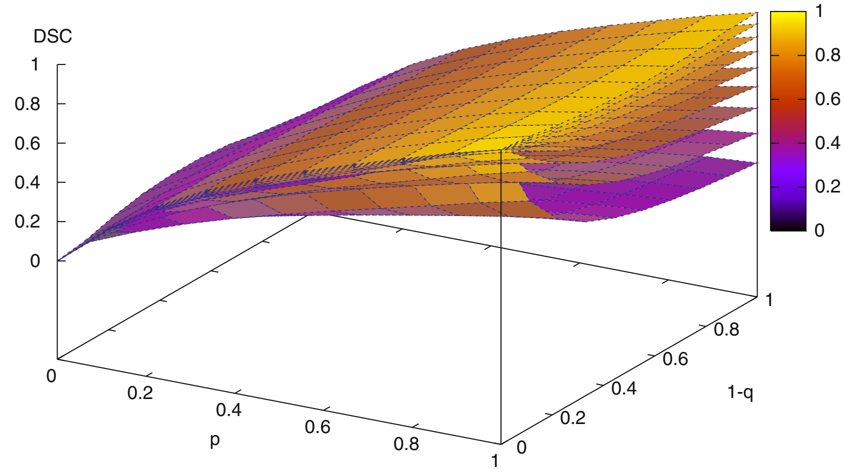

Fig. 1 Isometric (π) surfaces showing DSC in the ROC-space. Topmost surface is for $\pi = 0.9$, lowermost for $\pi = 0.1$

Left diagonal elements in the matrix represent correctly classified voxels (hits) while right diagonal represents falsely classified voxels (misses). The confusion matrix has three degrees of freedom.

Sensitivity, specificity, dice similarity measure, receiving operating characteristics (ROC) analysis

From the confusion matrix, different statistical parameters can be derived, e.g.:

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$q = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

$$\pi = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{3}$$

$$\theta = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{4}$$

where p, q, π, and θ are sensitivity, specificity, prevalence, and level of test, respectively. Sensitivity represents probability of the positive test for the positive instances while specificity represents probability of the negative test for negative instances. Since sensitivity depends only on measurements of diseased subjects while specificity only on healthy subjects, neither one depends on the disease prevalence, which makes them popular for biomedical diagnosis accuracy assessment [24].

DSC is defined as the amount of the intersection between a segmented object and the gold standard, or in terms of statistical parameters [18]:

$$\mathrm{DSC} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{5}$$

DSC is sometimes referred to as precision-recall balanced F-measure or positive specific agreement [25]. It is related with previously defined sensitivity, specificity, and prevalence (see [17]):

$$\mathrm{DSC} = \frac{2\pi p}{\pi(1 + p) + (1 - \pi)(1 - q)}. \tag{6}$$

From the Eq. 6 and Fig. 1, it is obvious that DSC is dependent on the disease prevalence which could be a limiting factor for the application in the medical image segmentation validation. Although geometrically intuitive, it lacks the information about the type of segmentation error, namely whether over or under-segmentation occurred, as shown in Fig. 2.

DSC is a special case of kappa statistics if the number of true negatives is significantly larger than the other three statistical parameters [18,26].

The ROC curve is a frequently used representation for biomedical diagnosis [24], decision making in machine learning and data mining problems [27]. It represents a relationship between the rate of true positives (p), on the y-axis, and the rate of false positives ($1-q$), on the x-axis. In ROC analysis, the area under the curve (AUC) defines to which amount the classifier under investigation is better than a random classifier (AUC being 0.5). AUC has some important features. It represents how well are positive and negative cases separated and is invariant to the prior probabilities [28]. The AUC can also be interpreted as the average sensitivity for all values of specificities and vice versa [24].

The ROC analysis can be used if the segmentation algorithm has a continuous response to parameter changes in the ROC space, e.g., edge detection algorithms [5], or image classifiers [29]. Grova et al. [30] propose a modification to ROC analysis for the evaluation of EEG localization methods. They separately validate those responses that are close to the simulated spike generator, i.e., close to the ground truth, and those far away from the source, i.e., false positives. With this method, the bias introduced by the low prevalence is reduced. However, more complex segmentation algorithms are dependent on the larger parameter sets, yielding a field of points in the ROC space, rather than a curve. A common

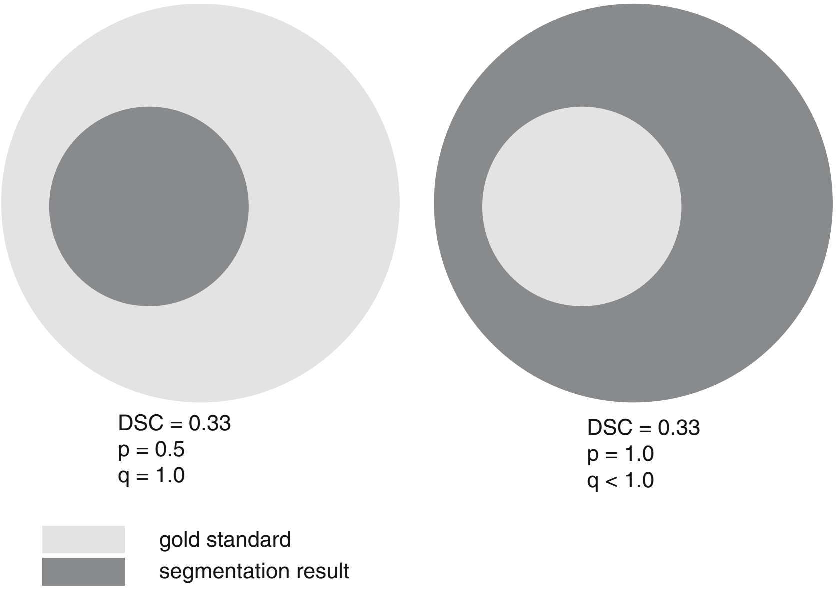

Fig. 2 An example of symmetry of DSC in relation to over or under-segmentation

approach to overcome this difficulty is to capture the most “north-western” points in a convex hull. This method could be misleading since the variance, i.e., robustness, is not being measured. Furthermore, computation complexity of a convex hull is $O(N \log N)$, N being number of points. An alternative measure is AUC of a point ($\text{AUC}_{p,q}$) defined as the area under the trapezoid bounded by points [0,0], $[p, 1-q]$, [1,1]. $\text{AUC}_{p,q}$ is the mean between p and q.

Requirements for statistical validation metrics in medical image segmentation

Accuracy of an algorithm in medical image segmentation has to be analyzed in a conjunction with the impact on the clinical outcome and the patients health. An often application of segmentation algorithms is a radiological diagnosis and image-guided procedures. Therefore, it is of crucial importance, not only to measure a discrepancy between an algorithmic segmentation result and the reference segmentation, but also to qualitatively determine the nature of the discrepancy. Medical researchers have been using cost (in terms of outcome) corrected tests for long time, but, due to the nature of medical diagnosis, they are highly dependent on population size, prevalence, and level of test. Applying diagnostic test on medical image segmentation and observing image elements as ‘patients’ and the image as test population is a speculative approach, due to the specificities of the problem (intra-element decision dependance, typically low prevalence, etc.). The following requirements for accuracy assessment in medical image segmentation are defined:

1. *Prevalence independent in the ROC-space*. The metric should perform a trade-off between sensitivity and specificity independently from the prevalence.
2. *Asymmetric*. The accuracy metrics should distinguish between over and under-segmentations. As discussed previously, for most of the clinical applications, one of these features is more important and has to be identified properly. An intuitive range of the metrics, in order to achieve both asymmetry of over and under-segmentation as well as universality of method is $[-1, 1]$, where -1 is full under-segmentation ($p = 0$, $q = 1$) and +1 full over-segmentation ($p = 1$, $q = 0$).
3. *Clinically intuitive*. The metric has to refer to the metrics traditionally used for medical test evaluation.

A new accuracy metric: *C*-Factor

A novel image segmentation validation metric called C-Factor (from *Coverage*) is proposed. If d is a discrepancy measure, defined as:

$$d = \frac{2 \cdot p \cdot (1-q)}{p + (1-q)} + \frac{2 \cdot (1-p) \cdot q}{(1-p) + q} \tag{7}$$

the C-Factor is defined as:

$$C = \begin{cases} d, & p \geq q \wedge p > 1-q \\ -d, & p < q \wedge p > 1-q \\ \text{undefined}, & p \leq 1-q. \end{cases} \tag{8}$$

For the case if $p \leq 1-q$, an algorithm is performing less accurate than the random classifier, and the C-Factor is undefined. In further analysis, we take into considerations only those algorithms that perform above the random diagnosis line in the ROC-space.

A similar criterion, called Gini-ROC[1] (d = Gini-ROC) was proposed by Flach [31] as a decision tree splitting

[1] Note: This is not the Gini-coefficient defined as $2 \cdot \text{AUC} - 1$.

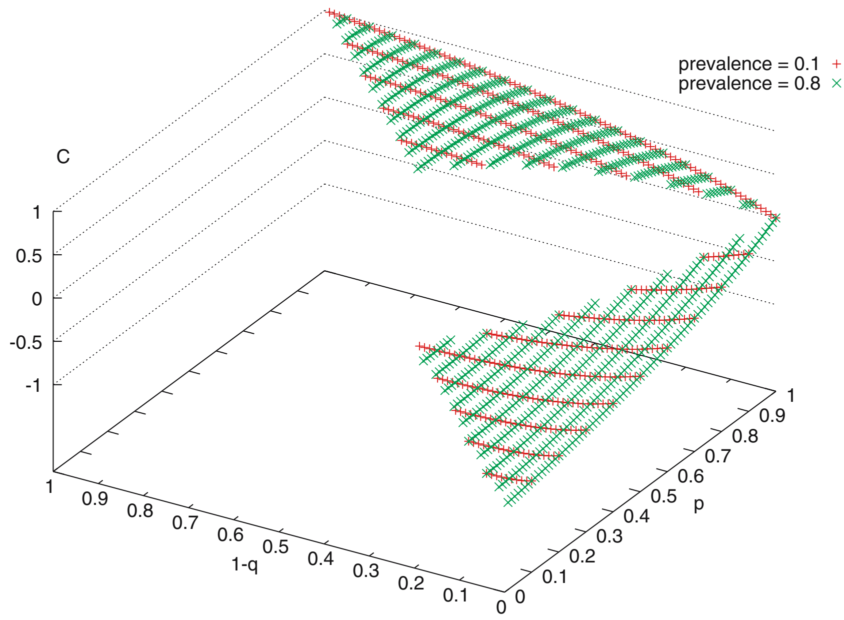

Fig. 3 Full simulation of entire segmentation space (all possible results) for two fixed object sizes (prevalence = 0.1 and 0.8). Both surfaces are analytically the same, although they are not defined in all points (some p–q combinations are not possible for a fixed π)

criteria in machine learning applications, which is an extension to the Gini index, used as a difference measure in various applications, e.g., [32,33]. Flach [31] proved that Gini-ROC is a prevalence insensitive variation of the χ^2 statistics. In the Appendix, the origin of the C-Factor is given.

Dependence on prevalence and level of test in the ROC-space

It is obvious from Eq. 8, that the C-Factor is insensitive to prevalence variations in ROC-space. This is achieved through a tradeoff between sensitivity and specificity. Figure 3 shows a 3D-ROC space depicting dependance of the C-Factor from sensitivity and specificity. Is is important to notice (Fig. 4) that the proposed metric is dependent on level of test, sign-inverse symmetrical around zero.

Asymmetry

For the same amount of over and under-segmentation C-Factor has the same absolute value but different sign. From the Eq. 8, following could be concluded:

$$\forall\{p_1, q_1\} \wedge \{p_2, q_2\} | p_1 = q_2 \wedge p_2 = q_1 \wedge p_1 \neq q_1 \\ \Rightarrow C(p_1, q_1) = -C(p_2, q_2). \tag{9}$$

Further features of C-Factor are:

$$\forall\{p, q\} | p > q \quad \Rightarrow C(p, q) > 0, \tag{10a}$$

$$\forall\{p, q\} | p < q \quad \Rightarrow C(p, q) < 0. \tag{10b}$$

Let us consider the diagnostic meaning of the sensitivity being larger then the specificity and vice versa. If N is a total number of image elements ($N = \text{TP} + \text{FP} + \text{TN} + \text{FN}$), then Eqs. 1 and 2 could be redefined as follows:

$$p = \frac{\text{TP}}{N\pi}, \tag{11a}$$

$$q = \frac{\text{TN}}{N - (\text{TP} + \text{FN})} = \frac{\text{TN}}{N(1-\pi)}. \tag{11b}$$

Hence:

$$p > q \Rightarrow \frac{\text{TP}}{N\pi} > \frac{\text{TN}}{N(1-\pi)} \\ \Rightarrow (1-\pi) \cdot \text{TP} > \pi \cdot \text{TN} \Rightarrow \text{TP} > \frac{\pi}{1-\pi} \cdot \text{TN}. \tag{12}$$

Similarly:

$$p > q \Rightarrow \text{FP} < \frac{1-\pi}{\pi} \cdot \text{FN}. \tag{13}$$

The prevalence dependent factor $\frac{\pi}{1-\pi}$ also called *skew ratio* [34,31] shows whether positives or negatives are more important in the ROC-space. Consequently, in the diagnostic sense, if the objective of the segmentation and the succeeding surgical procedure is the removal of diseased tissue and if some margin is allowed for removal of healthy tissue [35], a developer of the segmentation algorithm has to focus to the cases where C-Factor is positive and close to zero. However, if the segmentation is performed to assess some spatial feature of objects (e.g., brain volume), then the attention has to be given to both positive and negative space. Although DSC could be sufficient in this case, since it is exactly showing the spatial overlap, validating results using C-Factor can provide information on the tendency of algorithms and inherent parameter sets to under or over-segmentation, which is important information for a potential algorithm redesign.

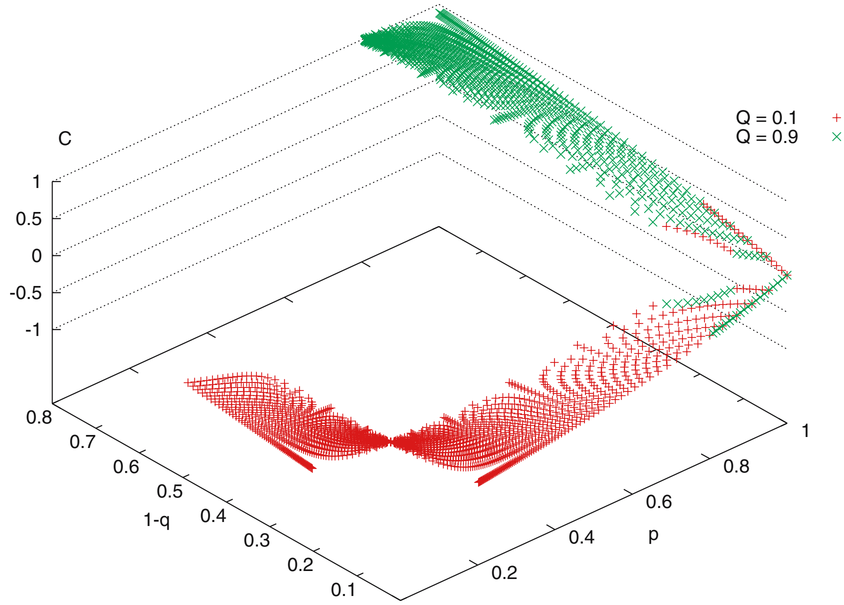

Fig. 4 Full simulation of entire segmentation space (all possible results) for two fixed segmentation results sizes (level of test = 0.1 and 0.9). For the small level of test (10%) the tendency is to <0 and for large (90%) >0. Note that this is a simulation of all possible cases with prevalence changing as well

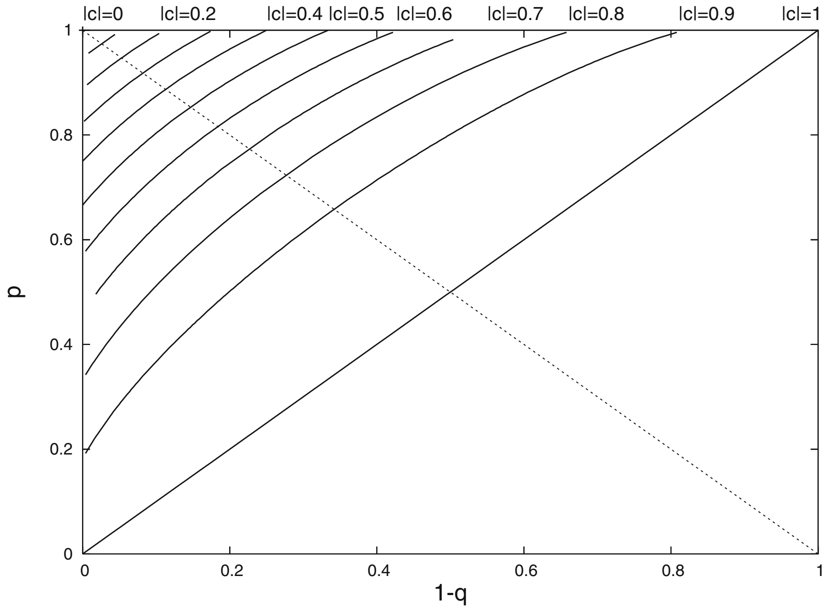

Fig. 5 Ten isometric curves of the C-Factor absolute value in the ROC-space (range [0:1], step 0.1). In the area below the dashed line, C-Factor is negative. Area above each curve is approx. $\frac{C^2}{8}$ for $C \ll 1$

Absolute value

The absolute value of the C-Factor can most effectively be analyzed in the ROC-space (Fig. 5). A C-Factor isometric curve ($C = c$) has the following intersections with axes in the ROC-space (Eq. 8):

$$p = 1 \Rightarrow q = 2 \cdot \frac{1-c}{2-c} \Rightarrow 1 - q = \frac{c}{2-c}, \tag{14}$$

$$q = 1 \Rightarrow p = 2 \cdot \frac{1-c}{2-c}. \tag{15}$$

Further important geometrical property of the C-Factor is its relation to the AUC. For cases in which C-Factor is small, the isometric curve could be approximated with a line. Hence, AUC is:

$$\mathrm{AUC}_C = 1 - \Delta, \tag{16}$$

where Δ is the area of the triangle in the left corner of ROC space, with vertexes in [0, 1], $[0, 2 \cdot \frac{1-c}{2-c}]$, and $[\frac{c}{2-c}, 1]$, according to Eq. 14. Hence:

$$\Delta = \frac{1}{2} \cdot \left(1 - 2 \cdot \frac{1-c}{2-c}\right) \cdot \frac{c}{2-c} \tag{17}$$

$$= \frac{1}{2} \cdot \frac{c^2}{(2-c)^2}. \tag{18}$$

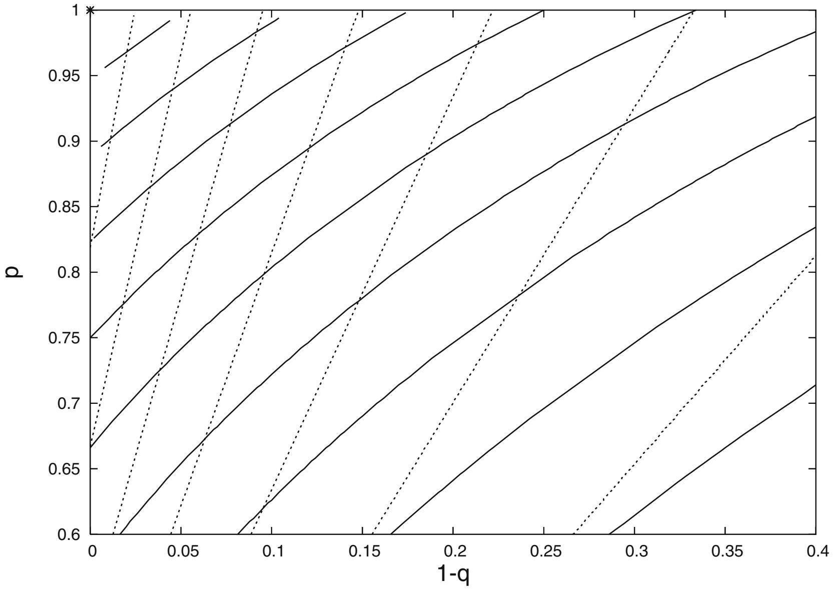

Fig. 6 Isometric lines of the C-Factor absolute value and DSC (*dashed lines*, for a fixed prevalence $\pi = 0.1$). Note that in the area closest to the up-left corner of the ROC-space (perfect match), DSC is overrating specificity against sensitivity. As the prevalence grows, the slope of DSC isometric lines declines and the lines shift right. Relevant p, q range [0.6,1.0] is depicted, only

For $c \ll 1 \Rightarrow (2-c)^2 \approx 4$, we can conclude:

$$\forall C, C \ll 1 \Rightarrow \mathrm{AUC}_C \approx 1 - \frac{C^2}{8}. \tag{19}$$

Since the C-Factor is defined for a single point in the ROC space, Eq. 19 needs more attention. If a point in the ROC space has a $C = c, c \ll 1$, it rests on a virtual ROC curve (Fig. 5) with an AUC $\approx 1 - \frac{c^2}{8}$. This quadratic relation makes C-Factor more sensitive to performance changes in the area close to the up-left corner.

Figure 6 represents a comparison of the DSC and C-Factor isometric curves. For the simplicity of the image, the prevalence is set to 0.1. It is Straight forward to demonstrate that for growing prevalence, the slope declines and the lines shift forward on the x-axis (Fig. 1). In the case near left-up corner, DSC is overrating the specificity over the sensitivity. Further analysis of relations between different figures of merit is given in the next section.

Comparison of statistical validation metrics

Definitions

To compare two different segmentation validation measures, $f(R, G)$ and $g(R, G)$, R and G being a segmentation result and the ground truth, respectively, three criteria are defined (similar definitions could be found in [34] and [36]). Λ is defined as the domain of R and G, and N as a total number of possible instances of R in Λ.

Definition 1 A segmentation result R_1 is *better* than a segmentation result R_2, for a given validation metric $f(R, G)$ if f is rating R_1 over R_2. The symbol for this relation is $\succ$: $f(R_1, G) \succ f(R_2, G)$.

Definition 2 (bias) The segmentation validation metrics $f(R, G)$ is biased if a pair $\{R_1, G\}$ and $\{R_2, G\}$, such that $f(R_1, G) = f(R_2, G)$ and $R_1 \neq R_2$ exists.

Definition 3 (consistency) Two segmentation validation metrics $f(R, G)$ and $g(R, G)$ are consistent if a pair $\{R_1, G\}$ and $\{R_2, G\}$, such that $f(R_1, G) \succ f(R_2, G)$ and $g(R_1, G) \prec g(R_2, G)$ does not exist.

Definition 4 (discriminancy) Segmentation validation metric $f(R, G)$ is more discriminating than $g(R, G)$ if a pair $\{R_1, G\}$ and $\{R_2, G\}$, such that $f(R_1, G) \neq f(R_2, G)$ and $g(R_1, G) = g(R_2, G)$ exists and a pair $\{R_1, G\}$ and $\{R_2, G\}$ such that $f(R_1, G) = f(R_2, G)$ and $g(R_1, G) \neq g(R_2, G)$ does not exist.

A formal analysis of segmentation validation metrics in terms of consistency and discriminancy is possible. However, as will be shown below, validation metrics rarely have total consistency, therefore we define probabilistic measures.

Definition 5 (degree of bias) If a segmentation validation metrics $f(R, G)$ has N_b pairs $\{R_i, G\}$ and $\{R_j, G\}$, such that $f(R_i, G) = f(R_j, G)$ and $R_i \neq R_j$, degree of bias is defined as:

$$\mathbf{B} = \frac{N_b}{N}. \tag{20}$$

Definition 6 (degree of consistency) If for two segmentation validation metrics $f(R, G)$ and $g(R, G)$ there exist N_c pairs $\{R_i, G\}$ and $\{R_j, G\}$, such that $f(R_i, G) \succ f(R_j, G)$ and $g(R_i, G) \prec g(R_j, G)$, degree of consistency is defined as:

$$\mathbf{C} = \frac{N - N_c}{N}. \tag{21}$$

Definition 7 (degree of discriminancy) If for two validation metrics $f(R, G)$ and $g(R, G)$ there exist N_{fg} pairs $\{R_i, G\}$ and $\{R_j, G\}$, such that $f(R_i, G) \neq f(R_j, G)$ and $g(R_i, G) = g(R_j, G)$, and N_{gf} pairs $\{R_i, G\}$ and $\{R_j, G\}$, such that $g(R_i, G) \neq g(R_j, G)$ and $f(R_i, G) = f(R_j, G)$, degree of discriminancy is defined as:

$$\mathbf{D}_{fg} = \frac{N_{fg}}{N_{gf}}. \tag{22}$$

Value of degree of consistency is in the range [0, 1]. Degree of discriminancy is in the range $[0, \infty]$. Degrees of consistency are symmetrical while the degrees of discriminancy are reciprocal ($\mathbf{D}_{fg} = \mathbf{D}_{gf}^{-1}$).

Numerical simulations

Numerical simulations have been performed to test the relations from Definitions 5 to 7. Three validation metrics have been selected for the simulations: DSC, C-Factor, and $\mathrm{AUC}_{p,q}$. $\mathrm{AUC}_{p,q}$ is chosen since it represents a trade-off between sensitivity and specificity, and is, in this manner, similar to the C-Factor.

Dividing right sides of the Eqs. 1 and 2 by N, where N is a total number of image elements (voxels), it is obvious that both sensitivity and specificity are dependent on the *rate* of statistical parameters from the confusion matrix:

$$p = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}}, \tag{23}$$

$$q = \frac{\mathrm{tn}}{\mathrm{tn} + \mathrm{fp}}, \tag{24}$$

where tp = TP/N, fp = FP/N, tn = TN/N, fn = FN/N, are rates of true positives, false positives, true negatives, and false negatives, respectively, such that:

$$\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn} = 1 \tag{25}$$

$$0 \leq \{\mathrm{tp}, \mathrm{fp}, \mathrm{tn}, \mathrm{fn}\} \leq 1. \tag{26}$$

From the Eqs. 7 and 23, it can be concluded that the C-Factor depends also from rate of statistical parameters only. Therefore, different combinations from the confusion matrix can be simulated by selecting all possible the rates of statistical parameters between zero and one. For the experiments described here, all possible combinations of the {tp, fp, tn, fn} have been generated with a sampling rate $= 10^{-3}$. Two simulated segmentation results are considered different if at least two statistical parameters from the confusion matrix are different. Only values above the random line in the ROC space have been taken into consideration, since the C-Factor is defined above the random line only.

Table 2 Results of numerical simulations for degree of bias

C[a]	0.308×10^{-3}
AUC	13.24×10^{-3}
DSC	4.21×10^{-3}

[a] Biased C-Factor values are > 0.85

Table 3 Results of numerical simulations of degree of consistency of three validation metrics

	$\lvert C\rvert$	AUC	DSC
$\lvert C\rvert$	\	0.935	0.847
AUC	0.935	\	0.933
DSC	0.847	0.933	\

Table 4 Results of numerical simulations of degree of disriminancy of three validation metrics

	$\lvert C\rvert$	AUC	DSC
$\lvert C\rvert$	\	∞	∞
AUC	0	\	0
DSC	0	∞	\

Results of simulations are given in Tables 2, 3, and 4. Notice that the absolute value of C-Factor has been analyzed for discriminancy and consistency, to avoid application specific decisions for $\{(C_1, C_2) | C_1 = -C_2\}$.

The most biased metric is AUC. It is invariant to mirroring over axes in the ROC space. As discussed above, DSC is insensitive to type of the error occurred (Fig. 2). Numerical experiments have shown that the C-Factor is biased in a smaller number of cases ($B(C) \propto B(\mathrm{DSC}) \cdot 10^{-1}$ and $B(C) \propto B(\mathrm{AUC}) \cdot 10^{-2}$). Further analysis has shown that biased values do not occur for $C < 0.85$. Although bias is an undesired feature, it is less critical if it occurs for segmentation results that could be discarded as bad based on their high C-Factor value.

The numerical experiments have yielded consistencies above 0.8. DSC and C-Factor are inconsistent in the cases if $p \approx 1$ and $q < 1$, when DSC is overvaluing leaks over the object borders. This effect could be analyzed for a case where $p = 1$ and $q < 1$. From Eq. 6:

$$\mathrm{DSC}(p = 1) = \frac{2\pi}{2\pi + (1 - \pi)(1 - q)}, \tag{27}$$

with DSC $\in [\frac{2\pi}{\pi+1}, 1]$. Therefore, DSC can overestimate over-segmentations for the larger values of the prevalence. Recommendations by [37] and [18] state that a 'very good' segmentation result is obtained for DSC > 0.7. However, this might be misleading since for $\pi \geq 0.54$ lower bound of DSC, for $q = 0$, is approx. 0.7 (Eq. 27). Therefore, if DSC is to be used as a quality measure, a region of interest for the validation has to be carefully selected to avoid misinterpretations.

Based on the similar reasoning as for the bias, C-Factor is more discriminating then AUC$_{p,q}$ and DSC. Table 3 shows that C-Factor is more discriminating, Definition 4, than other two metrics.

Huang and Ling [36] claim that a metric $f(R, G)$ is *better* then $g(R, G)$ if $\mathbf{C} > 0.5$ and $\mathbf{D}_{fg} > 1$.

According to the criteria by Huang and Ling [36] that a metric $f(R, G)$ is *better* then a metric $g(R, G)$ if $\mathbf{C} > 0.5$ and $D_{fg} > 1$, a conclusion could be made that C-Factor is *"better"* than DSC or AUC$_{p,q}$. However, this criteria might not be sufficient and needs stronger statistical analysis, which is beyond the area of this paper.

A case study: segmentation of calvarial tumors

To illustrate a practical use of the C-Factor, segmentation of calvarial tumors in computed tomography (CT) images for computer and robot assisted surgery is selected [38,39]. Surgical management of calvarial tumors implies total removal of the tumor [40]. Since calvarial tumors are not deep seeded in the brain, under-segmentation is highly critical, leading to post operative recurrence. Slight over-segmentation is less problematic, if there is no sensitive structures in vicinity of the tumor. Thus, we focus on segmentation results with the C-Factor larger than zero. Images are segmented with a knowledge-based level set segmentation algorithm and varying parameter set. More details on the algorithm could be found in [35,41].

Six CT data-sets have been used in this study, with post-operatively histologically validated diagnoses (Table 5). Resolution for patient A was 0.43 mm × 0.43 mm × 1 mm and for patients B–F: 0.44 mm × 0.44 mm × 2 mm. A latent gold standard, computed with the STAPLE algorithm [6] from manual expert segmentations has been used as the reference segmentation for validation. Table 5 lists the number of expert manual segmentation available for validation and histological diagnoses. Table 6 shows the variability of expert segmentations.

Table 5 Patient diagnoses

Patient	Diagnosis	Nr experts	Volume (cm^3)
A	Meningioma	4	7.3
B	Metastasis	1	59
C	Meningioma	4	77
D	Meningioma	3	18
E	Meningioma	2	7.5
F	Metastasis	3	24

Table 6 Variability of expert segmentations

Patient	DSC (Min, Mean, Max)
A	0.74, 0.77, 0.82
B	–
C	0.78, 0.81, 0.83
D	0.79, 0.84, 0.86
E	0.78
F	0.88, 0.9, 0.92

The motivation for introducing a novel figure of merit for validation of segmentation of calvarial tumors is two-fold. Results of level set based segmentation algorithm used are sensitive to parameter changes. The validation procedure allows us to observe behavior of the algorithm for given parameters without a visual inspection of all results. Due to the inter-subject variability, parameters could not be trained on previous patients or digital phantoms. Manual try-and-error procedure could be a tedious task, especially for an unexperienced user. For this purpose, an automatic parameter optimization procedure has been developed [42]. It requires an automatic segmentation result quality assessment. C-Factor could be useful for this task, due to it's discriminant and asymmetric nature.

Segmentation results

Leave-one-out cross validation has been performed (five patients for knowledge training and one patient for validation). Different segmentation results for each patient dataset have been generated from a predefined parameter space (see [42]). Those results have been used for the comparison between DSC and C-Factor. Figure 7 represents a boxplot

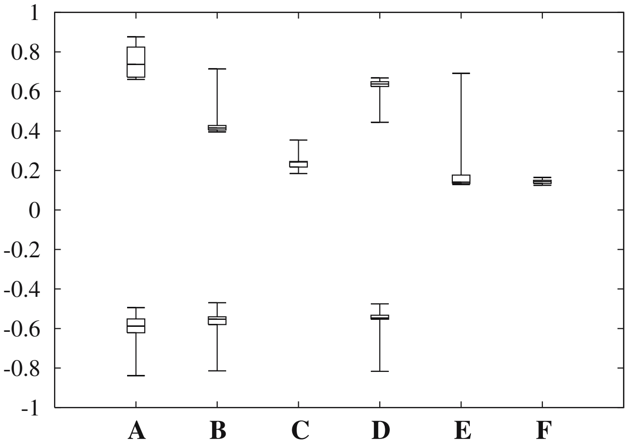

Fig. 7 Minimum, 1st quartile, median, 3rd quartile, maximum of C-Factor values for all six patients. Note: Two cases (C-Factor < 0, C-Factor > 0) are differentiated; therefore the box plot is divided into two parts

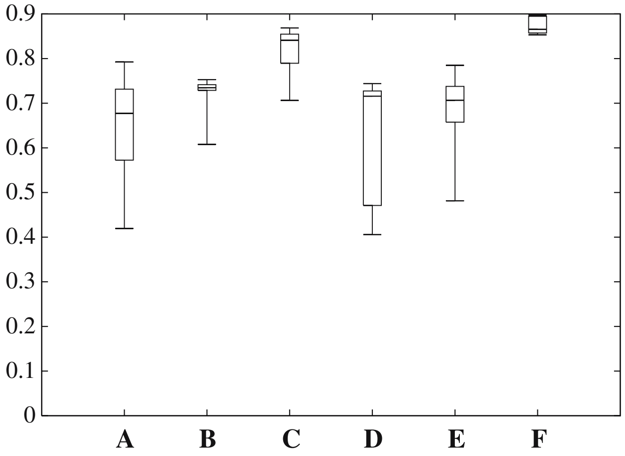

Fig. 8 Minimum, 1st quartile, median, 3rd quartile, maximum of DSC values for all six patients

of results obtained for C-Factor. Results obtained with DSC are shown in Fig. 8. Figures 9 and 10 show example slices of segmentation results for patients A and D.

Discussion of the clinical study

As discussed above, in contrast to DSC, C-Factor offers an insight into the behavior of the algorithm and the nature of errors occurred.

For patient A, based on the DSC observation, the conclusion could be drawn that the algorithmic parameter set with the best overlap (0.79) is optimal for the given patient. However, this segmentation result exhibits a negative C-Factor (–0.49). Optimal segmentation according to C-Factor, observing positive values, is the one with C-Factor 0.66 (DSC = 0.68). Example slices for this case are given in Fig. 7. It is obvious that, despite better overlap, the segmentation result with the best DSC value has more false negatives than the one with best C-Factor. Similar conclusions can be drawn for the patient D (Fig. 10).

From Fig. 7, it could be reasoned out that for patient A, the algorithm reaches the better overlap (smaller absolute value of C-Factor) for negative C-Factors. This could be taken into consideration for further algorithm development, e.g., attention should be given to the stopping criteria for level set propagation, to forbid surface leaks causing large absolute values of the C-Factor. Similar reasoning could be used for other patients.

Discussion

The accuracy assessment of segmentation results is not only an important prerequisite for the quality management (clinical approval) but also during the algorithm design and re-design stages. An appropriate parameter selection for a specific task or an individual patient is a challenging task due to variance in anatomy, pathology, image properties and quality. Recent endeavors were made to automatize the parameter selection process [43,44]. For such design problems with a large number of segmentation results, that have to

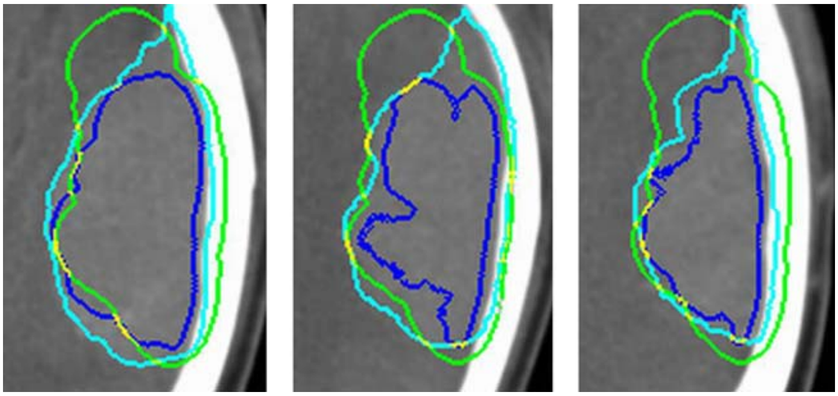

Fig. 9 Segmentation results at three different slices for the patient A; *cyan*: gold standard derived from manual segmentations; *blue*: the best segmentation as evaluated by DSC; *green*: the best segmentation as evaluated by C-Factor

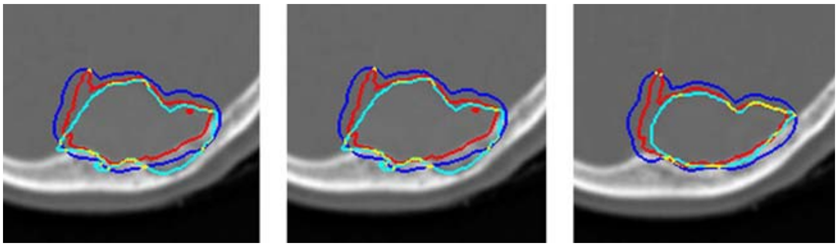

Fig. 10 Segmentation results at three different slices for the patient D; *cyan*: gold standard derived from manual segmentations; *red*: the best segmentation as evaluated by DSC; *blue*: the best segmentation as evaluated by C-Factor

be validated and compared, a metric being able to unambiguously measure the accuracy and the type of discrepancy is important.

Formal analysis of proposed figure-of-merit (fifth section) and commonly used statistical measures have demonstrated that the absolute value of C-Factor is more effective concerning degree of bias and capability to differentiate between different segmentation results. The biased cases, i.e., same C-Factor and different segmentation results, are observed for less relevant range in the (p, q) space (the values of C-Factor over 0.85), whereas for DSC, they are distributed over the entire parameter space. These features have been observed in the medical case study 6. C-Factor offered more insight into the behavior of the segmentation algorithm used in the study. The segmentation tendency inherent to the algorithm, parameters, and tumor type, i.e., different patient cases, could be observed.

We have demonstrated that the proposed measure meets all the criteria defined in section "Requirements for statistical validation metrics in medical image segmentation". The major weakness of the C-Factor is that it is prevalence independent in the ROC-space only, i.e., C-Factor inherited uncalibrated nature of sensitivity and specificity. However, in cases of (automatic) parameter selection or comparison of different algorithms on the same images, this is not an issue, given the constant prevalence over the cases. Further solution for this problem could be sought in a careful selection of the region of interest for validation, e.g., only in disputed regions close to the gold standard.

Registration of two images originating from different medical imaging modalities is an optimization problem aligning same features (e.g., labeling of structures) across two images [45]. Therefore, measure of discrepancy between the labels could be used for the validation, but also as one of the optimization terms. This requires a generalization of Eq. 8 to a multi-class measure. A possibility is to compute the mean value of sensitivity and specificity across all labels and use it in the standard formula for C-Factor.

Conclusion

The accuracy assessment of medical image segmentation algorithms is the key issue in clinical validation but also an every-day problem in the design and testing of segmentation algorithms. We have proposed a novel metric measuring the accuracy of a segmentation result in comparison to the reference segmentation. We also have demonstrated that the proposed metric satisfies all the prerequisites defined in the section "Requirements for statistical validation metrics in medical image segmentation". Furthermore, new metrics could be effectively used for different clinical application spaces, dependent on which classifier characteristic is more desirable. Although not as simple to interpret as spatial overlap, it is free of ambiguousness inherent to DSC. Nevertheless, it could be interpreted as a trade-off between sensitivity and specificity, being well-established measures in both medical and image processing societies. We demonstrate a simple relation between AUC and the proposed metric.

The generalization of the proposed metric to probabilistic segmentation results validation and multi-class decision making is a challenging prospective for the future work.

Acknowledgment This project is funded in part by the German Research Foundation (DFG), Framework 'Surgical Navigation and Robotics' (DFG RA548/2-1).

Appendix: Origin of the C-Factor

The C-Factor has origin in a binary decision tree splitting criterion (Gini-split) [46]. Gini-split minimizes impurities in training subset after the tree splitting [47]. If an image segmentation is interpreted as splitting of the original image into two classes (or child nodes in decision tree terminology), namely: positive and negative instances, we can assume notation from [31]:

$$g = 1 - \left(\frac{\mathrm{TP} \cdot \mathrm{FP}}{(\mathrm{TP} + \mathrm{FP})^2} + \frac{\mathrm{TN} \cdot \mathrm{FN}}{(\mathrm{TN} + \mathrm{FN})^2} \right), \tag{28}$$

where g is the Gini-split.

The first term in Eq. 28 ($\frac{\mathrm{TP}\cdot\mathrm{FP}}{(\mathrm{TP}+\mathrm{FP})^2}$) represents a trade-off between false and true positives (i.e., impurity of the positive child) while the second term ($\frac{\mathrm{TN}\cdot FN}{(\mathrm{TN}+\mathrm{FN})^2}$) represents a trade-off between false and true negatives (i.e., impurity of the negative child). Notice that the skew insensitive Gini-split [31] is used.

From Eqs. 11a and 11b we can compute:

$$\mathrm{FN} = N \cdot (1 - p)\pi, \tag{29}$$

$$\mathrm{FP} = N \cdot (1 - q)(1 - \pi). \tag{30}$$

Substituting Eq. 29 in Eq. 4:

$$\theta = p\pi + (1 - q)(1 - \pi). \tag{31}$$

Finally, substituting Eqs. 11a, 11b, 29, 30, and 31 in Eq. 28:

$$g = 1 - \left(\frac{2 \cdot p \cdot (1 - q)}{p + (1 - q)} + \frac{2 \cdot (1 - p) \cdot q}{(1 - p) + q} \right). \tag{32}$$

According to the second requirement defined in section "Requirements for statistical validation metrics in medical image segmentation", term from Eq. 32 is subtracted from one and multiplied with the sign of $p - q$.

References

1. Jannin P, Fitzpatrick JM, Hawkes DJ, Pennec X, Shahidi R, Vannier MW (2002) Validation of medical image processing in image guided therapy. IEEE Trans Medical Imaging 21(12):1445–1449
2. Udupa JK, LeBlanc VR, Schmidt H, Imielinska C, Saha PK, Grevera GJ, Zhuge Y, Molholt P, Jin Y, Currie LM (2002) A methodology for evaluating image segmentation algorithms. In: Proceedings of SPIE: medical imaging, pp. 266–277
3. Udupa JK, LeBlanc VR, Zhuge Y, Imielinska C, Schmidt H, Currie LM, Hirsch BE, Woodburn J (2006) A framework for evaluating image segmentation algorithms. Comput Medical Imaging Graph 30(2):75–87
4. Warfield SK, Zou KH, Wells WM (2002) Validation of image segmentation and expert quality with an expectation—maximization algorithm. In: Dohi T, Kikinis R (eds) Proceedings of MICCAI 2002, the fifth international conference. Springer, Heidelberg, pp 298–306
5. Yitzhaky Y, Peli E (2003) A method for objective edge detection evaluation and detector parameter selection. IEEE Trans Pattern Anal Mach Intell 25(10):1–7
6. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Medical Imaging 23(7):903–921
7. Collins DL, Zijdenbos AP, Kollokian V, Sled JG, Kabani NJ, Holmes CJ, Evans AC (1998) Design and construction of a realistic digital brain phantom. IEEE Trans Medical Imaging 17(3):463–468
8. Zubal IG, Harrell CR, Smith EO, Smith AL, Krischlunas P (1995) Two dedicated software voxel-based antropomorphic (Torso and Head) phantoms. In: Dimbylow PJ (ed) Proceeding of the international conference at the national radilogical protection board, pp 105–111
9. Cardoso JS, Corte-Real L (2005) Toward a generic evaluation of image segmentation. IEEE Trans Image Proces 14(11):1773–1782
10. Yoo TS, Ackerman MJ, Vannier M (2000) Toward a common validation methodology for segmentation and registration algorithms. In: Delp S, DiGioia A, Jaramaz B (eds) Proceedings of MICCAI 2000, the 3rd international conference, vol 1935 of Lecture Notes in Computer Science. Springer, Heidelberg, pp 422–431
11. Duncan JC, Ayache N (2000) Medical image analysis: Progress over two decades and the challenges ahead. IEEE Trans Pattern Anal Mach Intell 22(1):85–106
12. Jannin P, Grova C, Maurer CR Jr. (2006) Model for defining and reporting reference based validation protocols in medical image processing. Int J Comput Assisted Radiol Surg 1(2):63–73
13. Kraemer HC (1992) Evaluating medical tests. SAGE
14. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302
15. Prastawa M, Bullitt E, Ho S, Gerig G (2003) Robust estimation for brain tumor segmentation. In: Ellis RE, Peters TM (eds) Proceedings of MICCAI 2003, the sixth international conference, vol 2879 of Lecture Notes in Computer Science. Springer, Heidelberg, pp 530–537
16. Gerig G, Jomier M, Chakos M (2001) VALMET: a new validation tool for assesing and improving 3D object segmentation. In: Niessen WJ, Viergever MA (eds) MICCAI 2001, the fourth international conference, vol 2208 of Lecture Notes in Computer Science, pp 516–528
17. Zou KH, Wells WM, Kikinis R, Warfield SK (2003) Three validation metrics for automated probabilistic image segmentation of brain tumours. Statist Med 23(8):1259–1282
18. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, Wells WM, Jolesz FA, Kikinis R (2004) Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 11(2):178–189
19. Jaccard P (1912) The distribution of flora in the alpine zone. New Phytol 11:37–50
20. Shan ZY, Ji Q, Gajjar A, Reddick WE (2005) A knowledge-guided active contour method of segmentation of cerebella on mr images of pediatric patients with medulloblastoma. J Magn Reson Imaging 21:1–11
21. Roman-Roldan R, Gomez-Lopera JF, Atae-Allah C, Martinez-Aroza J, Luque-Escamilla PL (2001) A measure of quality for evaluating methods of segmentation and edge detection. Pattern Recogn 34:969–980
22. Goumeidane AB, Khamadja M, Belaroussi B, Benoit-Cattin H, Odet C (2003) New discrepancy measures for segmentation evaluation. In: International conference on image processing (ICIP), vol 2. IEEE, pp 411–414
23. Kohavi R, Provost F (1998) Glossary of terms. Editorial for the special issue on applications of machine learning and the knowledge discovery process. J Mach Learn 30(2/3) (in press)
24. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inf 38(5):404–415
25. Fleiss JL (1975) Measuring agreement between two judges on the presence or absence of a trait. Biometrics 31:651–659
26. Hripcsak G, Rothschild AS (2005) Agreement, the f-measure, and reliability in information retrieval. J Am Med Inf Assoc 12(3):296–297
27. Fuernkranz J, Flach PA (2005) Roc ‘n’ rule learning—towards a better understanding of covering algorithms. Mach Learn 58(1):39–77
28. Bradly A (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159
29. Horsch K, Giger ML, Venta LA, Vyborny CJ (2001) Automatic segmentation of breast lesions on ultrasound. Med Phys 28(8):1652–1659
30. Grova C, Daunizeau J, Lina J-M, Bnar CG, Benali H, Gotman J (2006) Evaluation of EEG localization methods using realistic simulations of interictal spikes. Neuroimage 29(3):734–753
31. Flach PA (2003) The geometry of ROC space: understanding machine learning metrics through roc isometrics. In: Proc 20th international conference on machine learning (ICML’03). AAAI Press, pp 194–201
32. Kuan Xu (2000) Inference for generalized Gini indices using the iterated-bootstrap method. J Bus Econ Statist 18(2):223–227
33. Castillo-Salgado C, Schneider C, Loyola E, Mujica O, Roca A, Yerg T (2001) Measuring health inequalities: Gini coefficient and concentration index. Epidemiol Bull Pan Am Health Organization 22(1) (in press)
34. Vilalta R, Oblinger D (2000) A quantification of distance bias between evaluation metrics in classification. In: ICML ’00: proceedings of the seventeenth international conference on machine learning. San Francisco, Morgan Kaufmann, pp 1087–1094
35. Popovic A, Engelhardt M, Radermacher K (2006) Knowledge-based segmentation of calvarial tumors in computed tomography images. In: Bildverarbeitung für Medizin, BVM 2006, Informatik-Aktuell. Springer, Heidelberg, pp 151–155
36. Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17(3) (in press)
37. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC (1994) Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 13(4):716–724

38. Popovic A, Engelhardt M, Wu T, Portheine F, Schmieder K, Radermacher K (2003) CRANIO—computer assisted planning for navigation and robot-assisted surgery on the skull. In: Lemke HU, Vannier MW, Inamura K, Farman AG, Doi K, Reiber JHC (eds), Proceedings of the 17th international congress and exhibition (CARS), vol 1256 of International Congress Series. Elsevier, pp 1269–1276
39. Bast P, Popovic A, Wu T, Heger S, Engelhardt M, Lauer W, Radermacher K, Schmieder K (2006) Robot- and computer-assisted craniotomy: resection planning, implant modelling and robot safety. Int J Med Robot Comput Assisted Surg 2(2):168–178
40. Engelhardt M, Bast P, Jeblink N, Lauer W, Popovic A, Eufinger H, Scholz M, Christmann A, Harders A, Radermacher K, Schmieder K (2006) Analysis of surgical management of calvarial tumours and first results of a newly designed robotic trepanation system. Minim Invasive Neurosurg 49(2):98–103
41. Popovic A, Engelhardt M, Wu T, Radermacher K (2006) Modeling of intensity priors for knowledge-based level set algorithm in calvarial tumors segmentation. In: Larsen R, Nielsen M, Sporring J (eds) Proceedings of 9th international conference on medical image computation and computer assisted intervention (MICCAI 2006), vol 4191 of Lecture Notes in Computer Science. Springer, Heidelberg, pp 864–871
42. Popovic A, Engelhardt M, Wu T, Radermacher K (2006) Towards automatic parameter optimization for medical image segmentation algorithms. In: Proceedings of the 11th international fall workshop, vision modeling, and visualization—VMV 2006
43. Maddah M, Zou KH, Wells WM, Kikinis R, Warfield SK (2004) Automatic optimization of segmentation algorithms through simultaneous truth and performance level estimation (STAPLE). In: Barillot C, Haynor DR, Hellier P (eds) Proceedings of MICCAI 2004, seventh international conference, vol 3216 of Lecture Notes in Computer Science. Springer, Heidelberg, pp 274–282
44. Abdul-Karim M-A, Roysam B, Dowell-Mesfin NM, Jeromin A, Yuksel M, Kalyanaraman S (2005) Automatic selection of parameters for vessel/neurite segmentation algorithms. IEEE Trans Image Proces 14(9):1338–1350
45. Crum WR, Camara O, Rueckert D, Bhatia KK, Jenkinson M, Hill DLG (2005) Generalized overlap measures for assessment of pairwise and groupwise image registration and segmentation. In: Duncan J, Gerig G (eds) Proceedings of MICCAI 2005, the 8th international conference, vol 3749 of Lecture Notes in Computer Science. Springer, Berlin, pp 99–106
46. Breiman L (1996) Technical note: some properties of splitting criteria. Mach Learn 24(1):41–47
47. Berzal F, Cubero J-C, Cuenca F, Martin-Bautista MJ (2003) On the quest for easy-to-understand splitting rules. Data Knowl Eng 44(1):31–48