Measuring Inter-observer Agreement in Contour Delineation of Medical Imaging in a Dummy Run Using Fleiss’ Kappa*

G. Rücker1; T. Schimek-Jasch; U. Nestle2
1Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg, Germany; 2Department of Radiology, University Medical Center Freiburg, Freiburg, Germany

1. Introduction

Imaging techniques such as positron emission tomography (PET), computed tomography (CT) or magnetic resonance imaging (MRI) help making diagnosis and treatment decisions. In radiotherapy, it is vitally important to identify the tumor and to define, by contour delineation, the clinical target volume to treat [1, 2]. However, observer agreement can be studied in so-called dummy runs where a number of trained observers contour the same target volume such that observer agreement on this volume can be measured. For the special case of two observers, Zijdenbos et al. [3] proposed a version of Cohen’s kappa for measuring agreement in contour delineation called kappa index. This index was implemented into the software system ARTIView (AQUILAB, Lille, France) together with a number of other pairwise indices such as overlap ratio, volume ratio, common contoured volume, to mention only a few [4]. Combined with slice thickness, the concepts of kappa index and overlap ratio for two observers can be extended to 3D measures of observer agreement.

1.1 Example: The PET-Plan Study Dummy Run

The PET-Plan study is an ongoing prospective, multicenter, randomized controlled trial (RCT) on the optimization of radiotherapy by means of PET-based target volume definition in patients with locally advanced non small cell lung cancer. The primary outcome is progression-free survival [5–7]. As a part of the quality assurance, a multicenter contouring dummy run on protocol adherence in the contouring of the primary tumor (Gross Tumor Volume, GTV) and the anatomical contouring of mediastinal lymph node stations (Clinical Target Volume, CTV) was carried out. Target volume delineation of a given dummy patient was done by 11 study centers twice, once before and once after training. The
goals were to verify the compliance of the study centers with the contouring guidelines of the study protocol and to measure the inter-observer agreement between centers, preferably by one single measure before and one after training. For organizational reasons, the centers as well as the observers within the centers were not necessarily the same before and after training. All study centers used their own software for contouring and sent their data sets to the dummy run coordinators. The evaluation was done using the ARTIView software.

2. Objective

A limitation of the kappa index proposed by Zijdenbos et al. [3] is that it is restricted to measuring pairwise agreement between two observers. For the case of more than two, say \( n \) observers, agreement can be described by an \( n \times n \) matrix of \( \left( \begin{array}{c} n \\ 2 \end{array} \right) \) potentially different values, each of them measuring pairwise agreement between a pair of two observers. Instead of looking at a matrix such as this, we aimed at having a single measure of agreement for \( n \) observers. In this article, we propose a generalization of the kappa index to an arbitrary number \( n \geq 2 \) of observers, using the derivation method introduced by Zijdenbos et al. [3].

3. Methods

When contouring a target volume based on a number of images corresponding to slices of the volume, observers separate the observed region into two parts, the region inside and the region outside the contour. In other words, given one slice, they characterize pixels by allocating them to one of two categories, inside (category 1) or outside (category 0) the contoured region. The number of pixels in each category corresponds to the area of the corresponding region. If the target volume is a 3D structure, separated into a series of slices of equal thickness, the pixels of each slice can be interpreted as voxels. The number of voxels in each category (1 or 0) corresponds to the volume of the corresponding region. Contour delineation is thus interpreted as a classification procedure where each one of \( n \) observers classifies a large number of pixels/voxels into two categories while the spatial structure of the image is ignored.

3.1 Two Observers: Cohen’s Kappa

Cohen’s kappa is a measure of association that was proposed by Cohen in 1960 [8, 9]. It is often applied when the interest is in measuring agreement between two observers who categorize a number of subjects (e.g., patients) into one of a finite number of discrete categories (e.g., disease states) [10, 11]. The generic form of overall kappa is given by

\[
\kappa = \frac{\hat{P} - \hat{P}_e}{1 - \hat{P}_e}
\]

where \( \hat{P} \) is the observed proportion agreement, \( \hat{P}_e \) the extent of agreement that would be expected by mere chance, and thus \( \hat{P} - \hat{P}_e \) the observed degree of agreement between observers above chance, whereas \( 1 - \hat{P}_e \) is the maximal attainable degree of agreement.

In our setting, the number of categories is two (0 and 1), and voxels are categorized. Instead of the number of subjects in a category, the volume of the selected/not selected region is measured.

Zijdenbos et al., considering only the case of two observers, applied Cohen’s kappa to this setting [3]. The outlying region (voxels contoured by neither observer) is thought to be of infinite size. They derived from Cohen’s kappa the formula which is now implemented as kappa index (KI) in the ARTIView software [4].

We propose a generalization of this method to more than two observers by replacing Cohen’s kappa with Fleiss’ kappa. Fleiss’ kappa (denoted by \( \kappa_n \)) is a measure of association that generalizes Cohen’s kappa for \( n \geq 2 \) indistinguishable observers [12–15].

3.2 More than Two Observers

First, consider a single slice. Let \( n \) be the number of observers and \( V_i \) the total number of voxels (that is, the volume of the slice) that is contoured by exactly \( i \) observers (\( i = 1, \ldots, n \)). For example, \( V_0 \) is the size of the volume that is delineated by all \( n \) observers, \( V_{n-1} \) the sum of all volumes delineated by all but one observer, and so on; \( V_2 \) is the sum of all volumes delineated by only two observers, and \( V_1 \) is the sum of all volumes delineated by only one of the observers. It is convenient to mark each of the regions corresponding to \( V_1, \ldots, V_n \) with different colors, as done by the ARTIView software, e.g., \( V_1 \) is coloured red. We can also consider \( V_0 \) which is the volume outside all contours. It is important to note that \( V_0 \) is not well-defined since it depends on the size of the image/the slice. Figure 1 illustrates this for the case of three observers.

3.2.1 Data Structure

The data is given as a (usually very large) table. Each row corresponds to a pixel in a 2D slice and is interpreted as a voxel in the 3D union of all slices. The first two columns are identifiers for the voxel and the slice. The next \( n \) columns correspond to \( n \) observers and contain a number indicating whether the observer categorized this voxel as being inside (1) or outside (0) the contoured region. Table 1 shows a detail from this data structure for \( n = 11 \) observers. Each row corresponds to a voxel of the given slice, here slice No. 84. Columns ‘1’ to ‘11’ correspond to observers 1 to 11. The number of observers contouring the voxel...
(between 0 and 11) is seen as the sum of the row entries in the rightmost column.

This data set is condensed to the final data set shown in Table 2. For each of 19 slices (rows), it contains the distribution of the number of observers per voxel. The bottom row shows the sum, which is interpreted as the distribution of the number of observers (i = 1, ..., n) per voxel in the total 3D volume, i.e., Vn is the number of voxels contoured by exactly i observers.

We note that it is not necessary to aggregate information like in Table 2, since we do not further distinguish between slices. However, Table 2 provides some insight into the slice-based structure of the total volume.

### 3.3 An Overall Kappa for an Arbitrary Number of Observers

Generalizing the approach used by Zijdenbos et al. [3] to an arbitrary number of observers and applying Fleiss’ kappa as a generalization of Cohen’s kappa to this setting, we show that overall kappa κn can be written

$$\kappa_n = \frac{1}{n-1} \sum_{i=1}^{n} \frac{i(i-1)V_i}{\sum_{i=1}^{n} iV_i}.$$

Note that if Vn (the region where all observers agree) approaches infinity, κn approaches 1, which makes sense (to see this, divide both numerator and denominator by Vn). The next subsection 3.3 gives an outline of the proof. Full details are found in Appendix A. An R function for calculating κn from a table structured as Table 1 is found in Appendix B.

#### Outline of the Proof of the Formula for κn

To derive a formula for κn, we start from Fleiss’ definition of κ for n observers, two categories (here: a voxel is allocated inside/outside the delineated region) and a number of ‘subjects’ to be categorized, where the ‘subjects’ here are voxels. Their number corresponds to the total volume

$$\sum_{i=1}^{n} V_i$$

‘inside’ the contoured region. Let V0 be the volume ‘outside’ of all contours, so that \(\sum_{i=0}^{n-1} V_i\) is the total volume. As mentioned above, V0 is not well-defined. For the first part of our considerations, V0 is held fixed. As all observers will agree to a large extent, particularly about the large volume ‘outside’, we then look what happens if V0 tends to infinity.

All voxels are categorized depending on how many observers selected them by contouring, as described above. We arrange entries in a table (Table 2) that facilitates the calculation of Fleiss’ kappa for these data; this arrangement was inspired by an analogous table [16]. Each row in the table is identified by a number i (i = 0, ..., n) (first column), corresponding to the number of observers selecting the voxels falling in this row. The second column contains the number of voxels found in this category which is the same as the volume Vn(i = 0, ..., n). This serves as the weight of the row.

The third and fourth column contain characteristics of the row: ‘inside’ is the number of observers who contoured the voxels in this row (which is again i), ‘outside’ is the number of observers who did not cover this region (which is n – i). Both numbers add up to n. The last column contains Pi which is the extent to which observers agree on voxels in row i. Pi is given by

$$P_i = \frac{1}{n(n-1)} (i^2 + (n-i)^2 - n).$$

### Table 1 Original data structure based on voxels (detail from one slice)

<table>
<thead>
<tr>
<th>Voxel</th>
<th>Slice</th>
<th>Observer number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1   2   3   4   5   6   7   8   9   10   11 Sum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...</td>
</tr>
<tr>
<td>5489</td>
<td>84</td>
<td>0     1     0     0     0     0     0     0     1     0     1     1</td>
</tr>
<tr>
<td>5490</td>
<td>84</td>
<td>0     1     0     0     0     0     0     0     1     0     0     2</td>
</tr>
<tr>
<td>5491</td>
<td>84</td>
<td>0     1     1     0     1     0     1     0     0     0     3</td>
</tr>
<tr>
<td>5492</td>
<td>84</td>
<td>0     1     1     0     1     0     0     0     1     0     4</td>
</tr>
<tr>
<td>5493</td>
<td>84</td>
<td>0     1     1     0     1     0     1     0     0     1     5</td>
</tr>
<tr>
<td>5494</td>
<td>84</td>
<td>0     1     1     0     1     0     1     0     0     1     5</td>
</tr>
<tr>
<td>5495</td>
<td>84</td>
<td>0     1     1     0     1     0     1     0     0     1     5</td>
</tr>
<tr>
<td>5496</td>
<td>84</td>
<td>0     1     1     0     1     0     1     0     0     1     5</td>
</tr>
<tr>
<td>5497</td>
<td>84</td>
<td>0     1     1     0     1     0     1     0     0     1     6</td>
</tr>
<tr>
<td>5498</td>
<td>84</td>
<td>0     1     1     0     1     1     1     0     0     1     7</td>
</tr>
<tr>
<td>5499</td>
<td>84</td>
<td>0     1     1     0     1     1     1     0     1     0     8</td>
</tr>
<tr>
<td>5500</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5501</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5502</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5503</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5504</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5505</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5506</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5507</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
<tr>
<td>5508</td>
<td>84</td>
<td>1     1     1     1     1     1     1     1     1     11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

© Schattauer 2012 Methods Inf Med 6/2012
The average of the $P_i$, weighted by the volume $V_i$ and still depending on $V_0$ is

$$\bar{P}(V_0) = \frac{1}{n(n-1)\sum_{i=0}^{n} i V_i} \sum_{i=0}^{n} i V_i (i^2 + (n-i)^2 - n)$$

This goes into the formula for overall kappa. The other part, $P_0$, is calculated as $P_0 = p_i^2 + p_0^2$ using the overall proportions $p_i$ and $p_0$ of assignments to the categories 'inside' and 'outside' in the last line of Table 3 and given by

$$p_0 = \frac{1}{n} \sum_{i=0}^{n} (n-i)V_i$$

$$\bar{P}_e(V_0) = p_i^2 + p_0^2 = \frac{1}{n^2 \left( \sum_{i=0}^{n} V_i \right)^2} \left[ \left( \sum_{i=0}^{n} i V_i \right)^2 + \left( \sum_{i=0}^{n} (n-i)V_i \right)^2 \right]$$

Using this notation, we have

$$\kappa_n(V_0) = \frac{\bar{P}(V_0) - \bar{P}_e(V_0)}{1 - \bar{P}_e(V_0)},$$

still depending on $V_0$. The idea going back to Zijdenbos et al. [3] is now to make the 'outside' region $V_0$ infinitely large and to investigate whether $\kappa_n(V_0)$ tends to a limit. This is the case as seen in Appendix A.

### 3.4 Variance Estimation for $\kappa_n$

For obtaining a confidence interval for $\kappa_n$, the variance must be estimated. However, a unit-of-analysis problem occurs with the common asymptotic variance estimator of Fleiss' kappa [14]. In the usual application, the variance depends on the number of observers and subjects to categorize: the more observers take part in the study and the more subjects are judged, the more precise the estimation becomes. In the application given here, however, the number of subjects is replaced with the number of voxels categorized which is very large, but completely arbitrary, depending on the number and size of images (slices) originating from only one patient. The usual formula yields [17]

$$\text{Var}(\kappa_n) = \frac{2}{n(n-1)\sum_{i=1}^{n} V_i}$$

which does not make sense because of the arbitrary number of voxels in the denominator. An alternative proposal is to make the variance depend only on the number of observers by setting each $V_i$ to 1, which yields

$$\text{Var}(\kappa_n) = \frac{2}{n^2(n-1)}.$$

Alternatively, bootstrap confidence intervals were calculated using two different approaches. In the first approach, samples with replacement were drawn from the set of observers. In the second approach, observers were fixed and samples were drawn from the set of all voxels. For both approaches, the mean was estimated from the resulting distribution of kappa values in a large number of bootstrap samples (10000), drawn with replacement. The 95% confidence intervals were determined by taking the observed 2.5% percentile of the distribution as the lower limit and the 97.5% percentile as the upper limit of the confidence interval.
4. Results

4.1 Examples of $\kappa_n$

For the special case $n = 2$ that was already treated by Zijdenbos et al., $\kappa_2$ reduces to

$$\kappa_2 = \frac{2V_2}{V_1 + 2V_2}$$

which agrees with the formula given before [3]. As $V_2$ denotes the volume delineated by both observers and $V_1$ the sum of the two volumes delineated by either of them, $\kappa_2$ is the share of the overlapping volume in the average volume of both observers.

For $n = 3$ (as in Fig.1) we obtain

$$\kappa_3 = \frac{V_2 + 3V_3}{V_1 + 2V_2 + 3V_3}.$$  

Table 3 Table for derivation of overall kappa

<table>
<thead>
<tr>
<th>$i$</th>
<th>Volume inside</th>
<th>outside</th>
<th>$P_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$V_0$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>$V_1$</td>
<td>$n-1$</td>
<td>$\frac{n-2}{n}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$i$</td>
<td>$V_i$</td>
<td>$n-i$</td>
<td>$\frac{1}{n(n-1)}(i^2 + (n-i)^2 - n)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$n-2$</td>
<td>$V_{n-2}$</td>
<td>$n-2$</td>
<td>$2$</td>
</tr>
<tr>
<td>$n-1$</td>
<td>$V_{n-1}$</td>
<td>$n-1$</td>
<td>1</td>
</tr>
<tr>
<td>$n$</td>
<td>$V_n$</td>
<td>$n$</td>
<td>0</td>
</tr>
</tbody>
</table>

Total $\sum_{i=0}^{n} V_i$, $\sum_{i=0}^{n} iV_i$, $\sum_{i=0}^{n} (n-i)V_i$.

Proportion in category $P_1$, $P_0$.

4.2 Results of the PET-Plan Dummy Run

From 11 study centers participating in the first dummy run, 8 also took part in the second dummy run. In addition to these, three new centers participated in the second dummy run. Observers within centers were not generally the same before and after training. We use the gross tumor volume (GTV) as an example of an outcome. Data for GTV were available from 11 observers in the first run (19 slices, 17,225 voxels) and 10 observers in the second run (18 slices, 14,430 voxels); data from one of the centers were missing in the second run. Overall kappa was slightly enhanced from $\kappa_{11} = 0.59$ (95% CI 0.51–0.67) before training to $\kappa_{10} = 0.69$ (0.59–0.78) after training. The average pairwise kappa increased from 0.57 to 0.72.

The bootstrap 95% confidence intervals were for the first approach (drawing from the set of observers) 0.62 (0.52–0.73) before training and 0.72 (0.66–0.79) after training. For the second approach (drawing from the voxels) we got 0.588 (0.583–0.593) before training and 0.69 (0.68–0.73) after training.

5. Discussion

We have introduced overall kappa $\kappa_n$, a measure of observer agreement in contour delineation that generalizes the kappa index proposed by Zijdenbos et al. [3] to more than two indistinguishable observers. This was achieved by replacing Cohen’s kappa with Fleiss’ kappa in the derivation following these authors. For overall kappa, it is not required to have observers fixed. To compare overall kappa between different groups or settings (e.g., before and after training sessions) it is also not necessary that their number is equal. The formula for overall kappa is very simple. Overall kappa can be calculated if information for each voxel, each slice and all observers are known, or if at least the $n$ volumes contoured by exactly $i$ ($i = 1, \ldots, n$) observers are available. At present, the ARTIview software displays for each slice the areas of different extent of agreement in different colors, but does not output them as numbers [4].

A similar measure seems to have been applied in development of consensus guidelines for the definition of the clinical target volume for radiation therapy for prostate cancer. However, details of its derivation are not reported [1].

The PET-Plan study is a therapeutic RCT where contour delineation plays the role of a means to define the clinical target volume which should ideally lead to similar results in all study centers for a given case. The dummy run was part of the quality assessment in the multicenter trial, while the teaching session followed by a repetition of the dummy run was intended to improve contouring consistency among the centers. We were primarily interested in observer agreement.

5.1 Limitations

Our approach has limitations. First, though $\kappa_n$ is always between 0 and 1 (in contrast to kappa in general, which can be negative), there is no straightforward interpretation of overall kappa as a proportion or a measure of probability, except for the case $n = 2$. Secondly, estimation of the variance is difficult, as there is uncertainty as to what unit of analysis has to be used. We propose a conservative variance estimate that does not depend on the arbitrary number of voxels. Alternatively, bootstrap confidence intervals were calculated. However, the results depend on which unit of analysis is chosen. When drawing from the set of ob-
servers, there is concern that bootstrapping systematically overestimates kappa and underestimates its standard error, as sampling with replacement tends to increase inter-observer agreement. When drawing from the set of voxels, kappa is estimated unbiased, but with very narrow confidence intervals not reflecting that there are only few observers and only one patient. Moreover, for the run after training the kappa values simulated in this way surprisingly showed a bimodal distribution, which makes confidence intervals difficult to interpret.

In our application, there was only one dummy patient, which limits the value of \( \kappa \) to this patient. In this sense, \( \kappa \) is a patient-specific measure of observer agreement which could be further evaluated in a future study to analyze it for a greater number of patients. If agreement between observers is to be assessed based on images from multiple patients, the patient-specific kappa values may be averaged over all respective patients. We note that this also concerns all pairwise coefficients calculated so far by the contouring software ARTI-View.

In the case at hand we asked for inter-observer agreement, and the same image was assessed twice by (partly) different observers. If the same image is assessed more than once by exactly the same observers, it is possible, and may be of additional interest, to assess intra-observer reliability by calculating observer-associated kappa values using our method. This provides apparently no answer to the question whether inter-observer agreement has changed over time. Finding a ‘paired’ method seems difficult since inter-observer agreement is not a function of the individual observer, but of the whole group of observers. Even if observer agreement is satisfying, validity may be low. If, for instance, all observers agree but simultaneously fail to hit the true target because of correlated errors, validity is poor. In this case, the assumption of conditional independence of the observers would be violated, a known problem in latent class analysis [18, 19]. Under the assumption that observers are independent conditional on the true classification, the STAPLE algorithm (Simultaneous truth and performance level estimation) is a powerful tool in estimating validity [20]. It is an application of the EM (expectation-maximization) algorithm allowing maximum likelihood estimation in the presence of incomplete information [21]. However, its estimates may be biased if the correlation between observers is misspecified.

6. Conclusion

If the main focus is on observer agreement, overall kappa is of use as a simple measure of observer agreement for a set of two or more indistinguishable observers. In the presence of a gold standard, the focus should be on validity. If there is no gold standard and the focus is on validity, the STAPLE algorithm [20] may be used.

Acknowledgment

The authors gratefully thank Romain Viard and Baptiste Germain (AQUILAB, Lille, France) who kindly provided the original data source from the software ARTIView. Further, we are indebted to Werner Vach for critical comments and discussion and to Caroline Mavergames for polishing English language. We also thank the PET-Plan study centers for cooperation in the dummy run.

References


