Inter-observer variability of manual contour delineation of structures in CT

L. Joskowicz\textsuperscript{1} PhD, D. Cohen\textsuperscript{1} MSc, N. Caplan\textsuperscript{2} MD MS, J. Sosna\textsuperscript{2}

\textsuperscript{1} The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel.
\textsuperscript{2} Department of Radiology, Hadassah Hebrew University Medical Center, Jerusalem, Israel.

Appendix

The following definitions and metrics are used for the evaluation. Fig. 1A illustrates the main metrics and concepts.

A1. Scan volume

Let $I$ be a volumetric scan consisting of image slices represented by a 3D gray-level matrix $V(i, j, k)$ of voxels $v_{ijk}$ of size $W \times W \times H$ where $i, j$ is the corresponding pixel location of the voxel $v_{ijk}$ in image slice $k$. Let $S \subseteq V$ be a set of voxels defined by $S = \bigcup_{k=1}^{S} [S]_k$ where $[S]_k$ is a set of voxels on each image slice $k$. Let $|S|$ be the number of voxels in a set $S$ of voxels and let $\text{volume}(S)$ be the volume of a $S$, i.e., the number of voxels in $S$ times the voxel volume, $\text{volume}(S) = |S| \times W \times W \times H$.

A2. Volumes comparison

We define two measures to compare the volumes of two voxel sets $S_1, S_2$. The volume difference of two volumes is:

$$\text{vol}\_\text{diff} ([S_1]_k, [S_2]_k) = \text{volume}([S_1]_k) - \text{volume}([S_2]_k) \text{ for slice } k$$

$$\text{vol}\_\text{diff} (S_1, S_2) = \text{volume}(S_1) - \text{volume}(S_2) = \sum_{k=1}^{n} (\text{volume}([S_1]_k) - \text{volume}([S_2]_k))$$

The volume overlap similarity, also called the Dice index is:

$$\text{dice}([S_1]_k, [S_2]_k) = \frac{2 \times \text{volume}([S_1]_k \cap [S_2]_k))}{\text{volume}([S_1]_k) + \text{volume}([S_2]_k)}$$
Fig. A1: Illustration of the notation: (a) volumetric scan of a structure $S$ represented as a 3D gray-level matrix $V(i,j,k)$ of voxels $v_{ijk}$. Voxel $v_{ijk}$ corresponds to pixel $(i,j)$ in image slice $k$; (b) image slice with structure $[S]_k$ (blue), delineation contour $[d]_k$ consisting of a set of voxels (dark orange), and its corresponding delineation volume set $v_s([d])_k$ (light and dark orange); (c) volume overlap and volume overlap variability of two delineations $d_1$ and $d_2$ and their corresponding delineation volume sets $v_s(d_1)$ and $v_s(d_2)$ (the slice index $k$ is omitted for simplicity); (d) Possible, Consensus and Variability sets and the mean delineation of a delineation set $D_4 = \{d_1, d_2, d_3, d_4\}$ consisting of four delineations.
\[
dice(S_1, S_2) = \frac{2 \times \text{volume}(S_1 \cap S_2)}{\text{volume}(S_1) + \text{volume}(S_2)}
\]

The volume overlap difference is \(1 - \text{the Dice index:}
\]

\[
\text{vol\_overlap\_diff}(S_1, S_2) = 1 - \dice(S_1, S_2)
\]

When the delineations are error-free, the volume overlap difference between their volume sets is the volume overlap variability between the delineations:

\[
\text{vol\_overlap\_variability}(S_1, S_2) = \text{vol\_overlap\_diff}(S_1, S_2) \text{ when } S_1, S_2 \text{ are correct}
\]

A3. Group-wise delineations comparison

Let \(O\) be a structure of interest in \(I\) and let \(d\) be a delineation of \(O\) in volume \(V\) created manually by an observer or generated automatically by an algorithm. A delineation \(d\) is the set of voxels on the structure boundary defined by the set of all lists of voxels \([d]_k\) on the structure boundary on each image slice \(k\), \(d = \bigcup_{k=1}^{n} [d]_k\). The volume set \(\text{vs}(d)\) of a delineation is the set of voxels on or inside the delineation boundary defined by the set of voxels \(\text{vs}([d]_k)\) on or inside the structure boundary on each image slice \(k\), \(\text{vs}(d) = \bigcup_{k=1}^{n} \text{vs}([d]_k)\).

Let \(D_n = \{d_1, \ldots, d_n\}\) be a set of \(n\) delineations of the same structure \(O\) in a volumetric scan \(V\) defined by the set of delineations on each image slice \(k\), \([D_n]_k = \{[d_1]_k, \ldots, [d_n]_k\}\). We define \(\text{Consensus}(D_n)\) as the set of voxels that are included in all delineations, \(\text{Possible}(D_n)\) as the set of voxels that were included in at least one delineation, and \(\text{Variability}(D_n)\) as the set of voxels for which there is a disagreement between delineations:

\[
\text{Consensus}(D_n) = \bigcap_{i=1}^{n} \text{vs}(d_i); \quad \text{Possible}(D_n) = \bigcup_{i=1}^{n} \text{vs}(d_i)
\]

\[
\text{Variability}(D_n) = \bigcup_{i=1}^{n} \text{vs}(d_i) - \bigcap_{i=1}^{n} \text{vs}(d_i)
\]

and
\begin{align*}
    Consensus([D_n]_k) &= \bigcap_{i=1}^{n} vs([d_i]_k); \quad Possible([D_n]_k) = \bigcup_{i=1}^{n} vs([d_i]_k) \\
    Variability([D_n]_k) &= \bigcup_{i=1}^{n} vs([d_i]_k) - \bigcap_{i=1}^{n} vs([d_i]_k)
\end{align*}

By definition, the volume sets $Consensus(D_n)$, $Possible(D_n)$ and $Variability(D_n)$ depend on the number of delineations $n$: higher values of $n$ indicate a more comprehensive, and hence a more accurate characterization of the actual delineation variability. Note that $Consensus(D_n)$ monotonically decreases and $Possible(D_n)$ and $Variability(D_n)$ monotonically increase as $n$ increases. The measures are the worst-case upper bounds on the delineation variability of a structure in a scan. More precisely, for all $k \leq n$:

1. $volume(Consensus(D_k)) \geq volume(Consensus(D_n))$ since every new delineation may remove voxels which were included in the existing delineations. $Consensus(D_k)$ monotonically increases with $n$.

2. $volume(Possible(D_k)) \leq volume(Possible(D_n))$ since every new delineation may add voxels that were not included by the existing delineations. $Possible(D_k)$ monotonically increases with $n$.

3. $volume(Variability(D_k)) \leq volume(Variability(D_n))$ since $Variability$ is defined by the difference between $Possible$ and $Consensus$. $Variability(D_k)$ monotonically increases with $n$.

**A4. Group-wise delineations comparison metrics**

Let $mean(D_n)$ to be the mean delineation of the structure of interest $O$ in volume $V$ defined by the mean of the delineations $mean([D_n]_k)$ of $O$ in slice $k$, $mean(D_n) = \bigcup_{k=1}^{S} mean([D_n]_k)$. The mean contour delineations $mean([D_n]_k)$ are computed with the STAPLE method [10].

For each case (image scan and structure delineation), we define the following measures.

1. **Manual tracing.** The minimum delineation variability of a set of delineations $D_n$ is defined on the slice $k$ for which the volume of the delineations variability divided by the volume of the mean delineation is smallest:
minimum_delineation_variability \( (D_n) = \min_k \left\{ \frac{\text{volume}(\text{Variability}([D_n]_k))}{\text{volume}(\text{vs}(\text{mean}([D_n]_k)))} \right\} \)

2. Pairs of observers. The mean delineation variability between two delineations \( d_i, d_j \) is defined as the mean of the volume of the pairwise delineation variability divided by the volume of the mean delineation for all slices \( k \):

\[
\text{pairwise_mean_variability} \ (d_i, d_j) = \text{mean}_k \left( \frac{\text{volume}(\text{Variability}([d_i]_k, [d_j]_k)))}{\text{volume}(\text{vs}(\text{mean}([d_i]_k, [d_j]_k))))} \right)
\]

The definitions for the mean pairwise Consensus and Possible delineations variability, pairwise_mean_consensus and pairwise_mean_possible are obtained by replacing \( \text{Variability}([d_i]_k, [d_j]_k) \) by \( \text{Consensus}([d_i]_k, [d_j]_k) \) and \( \text{Possible}([d_i]_k, [d_j]_k) \) respectively. The definitions for minimum and maximum pairwise delineation variability, pairwise_min_variability and pairwise_max_variability are obtained by replacing the \( \text{mean} \) operator by the \( \text{min} \) and \( \text{max} \) operators, respectively.

3. Case and structure type. For each case, we compute the volume overlap between the volume of each delineation \( d_k \) and the volume of the mean delineation \( \text{mean}(D_n) \) with the Dice similarity coefficient:

\[
\text{delineation_similarity} \ (d_k, D_n) = \text{dice}(\text{vs}(d_k), \text{vs}(\text{mean}(D_n)))
\]

For all cases of the same type -- level of difficulty and structure type -- we compute the minimum, maximum and mean delineation similarity from the individual delineation similarity \( \text{delineation -- similarity} \ (d_k, D_n) \).

4. Expertise of observers. For each case and for each observer, we compute the volume overlap between the volume of the observer delineation \( d_k \) and the volume of the mean delineation \( \text{mean}(D_n) \) with the Dice similarity coefficient as in item 5 above. We also compute the mean, maximum, and minimum delineation similarity by type of structure and by observer expertise. We compare the pairwise differences with standard Student t-test. Significance is determined for \( p < 0.05 \).

5. Groups of observers. The mean delineation variability of a given group of \( l \) delineations \( D_l \subseteq D_n, 1 \leq l \leq n \), is a generalization of the pairwise mean delineation variability. It is the mean of
the volume of the delineation variability divided by the volume of the mean delineation for all slices \( k \):

\[
\text{group\_mean\_variability} (D_l) = \text{mean}_k \left( \frac{\text{volume}(\text{Variability}([D_l]_k))}{\text{volume}(\text{vs(mean}([D_l]_k)))} \right)
\]

For all groups of \( m \) delineations out of \( n \), the sets \( D_l \subseteq D_n, |D_l| = m, 1 \leq l \leq \binom{n}{m} \):

\[
\text{all\_groups\_mean\_variability} (D_n, m) = \text{mean}_l \left( \text{group\_mean\_variability} (D_l) \right)
\]

The definitions for \textit{Consensus}, \textit{Possible}, minimum and maximum one group and all groups variability are defined by replacing the corresponding terms for \( \text{Variability}([d_i]_k, [d_j]_k) \) by \( \text{Consensus}([d_i]_k, [d_j]_k) \) and \( \text{Possible}([d_i]_k, [d_j]_k) \) and by replacing the \textit{mean} operator in the two equations above with the \textit{min} and \textit{max} operators, respectively.

6. \textit{Non-agreement between observers}. The disagreement between \( m \) observers for a set of delineations \( D_n \) is defined as the total number of voxels in the variability volume set that were chosen to be on or inside the structure by \( m \) observers plus those not chosen to be on the structure by \( n - m \) observers divided by the total number of voxels in the variability volume set. Formally for a given \( m, 1 \leq m \leq \left\lfloor \frac{n}{2} \right\rfloor \):

\[
\text{discrepancy}(D_n, m) = \frac{\sum_{v \in \text{Variability}(D_n)}(\text{rank}(v, m) + \text{rank}(v, n - m))}{|\text{Variability}(D_n)|}
\]

where \( \text{rank}(v, l) \) is a function returns 1 when voxel \( v \) in the variability set of all delineations \( \text{Variability}(D_n) \) has been selected by exactly \( m \) observers. Note the symmetry between \( \text{rank}(v, m) \) and \( \text{rank}(v, n - m) \): the voxel was selected by exactly \( m \) observers, and the voxel was not selected by exactly \( n - m \) observers. Thus, exactly \( m \) observers disagree.