How many radiologists does it take to reliably delineate a tumor? A large-scale clinical study of observer variability

D. Cohen[1], L. Joskowicz[1], N. Caplan[2], J. Sosna[2]

[1] The Rachel and Selim Benin School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel.

[2] Department of Radiology, Hadassah Hebrew University Medical Center, Jerusalem, Israel.

Background: The evaluation of segmentation algorithms requires ground-truth manual delineations of expert radiologists. In most cases, one, and sometimes two experts generate these delineations. These delineations are then used to quantitatively evaluate the segmentation algorithms with various measures, e.g. the Dice coefficient, and the mean surface distance. However, it is well known that different radiologists generate different delineations, as has been reported in both the clinical and technical literature. The delineations variations depend on many factors, e.g., the structure of interest, the resolution, contrast, and quality of the scan, the radiologist experience, and the radiologist available time, patience and dedication, among others. To properly assess the algorithms, it is thus essential to quantify the observer variability. While quantifying observer variability is recognized as a key issue by radiologists and technologists, very few large-scale studies have been conducted to actually quantify it.

Method: We conducted a large manual delineation study at the Hadassah University Medical Center to obtain ground truth segmentation variability data and to quantify the radiologists delineation variability. We retrospectively selected 16 CT studies, 5 from liver tumors, 5 from lung tumors, 6 left kidneys from our Center with dimensions 512x512x350-466 voxels and resolutions 0.76-0.98x0.76-0.98x1-3.3mm^3. Care was taken to choose diverse cases in terms of clinical pathology, structure aspect, image intensity differences, texture, structure size, and shape complexity. Eleven radiologists, including residents, mid-career, and expert radiologists to generate a total of 2,829 axial slices from the 16 CT scans. The delineations were made with the ITK-SNAP v3.4 software package after one hour of training. An expert radiologist coordinated the study and reviewed the delineations of all clinicians. She made corrections to about 10.5% of the delineations when she did not agree with them. Fig. 1 shows representative delineations of three structures.

The data analysis of the resulting delineations quantified the structure delineation variability as the difference between the union (possible) and the intersection (consensus) of the voxels inside the delineations of k observers divided by the mean structure volume. Results were compiled by: 1) pairs of annotators; 2) groups of annotators; 3) case type and difficulty; 4) annotators expertise; 5) annotators disagreement, and; 6) surface distance difference.

Results: The kidney, liver tumors, and lung tumors contour delineation variability is 7%, 14% and 16% for 3 observers and 13%, 26%, and 32% for 10 observers, respectively. Fig. 2 shows the how the liver tumors volume variability increases as a function of the number of observers. The variability convergence rate volume reaches to 52%-56% for 3 annotators, up to 86%-89% for 7 annotators. Overall, 40% of the variability is due to one annotator, and 60% to 2. At least 6 annotators needed to get a variability value. No statistical difference between annotators expertise was found.

Conclusions: The preliminary analysis of our results indicates that: 1) the observer variability spans a wide interval and varies significantly depending on the structure type and the case difficulty; 2) two or even three observers usually do not suffice to properly quantify observer variability; 3) the observer variability converges slowly as the number of observers increases; 4) there is minor delineation variability between the annotators based on their level of expertise; 5) there are significant differences between the convergence rates for different types of structures.

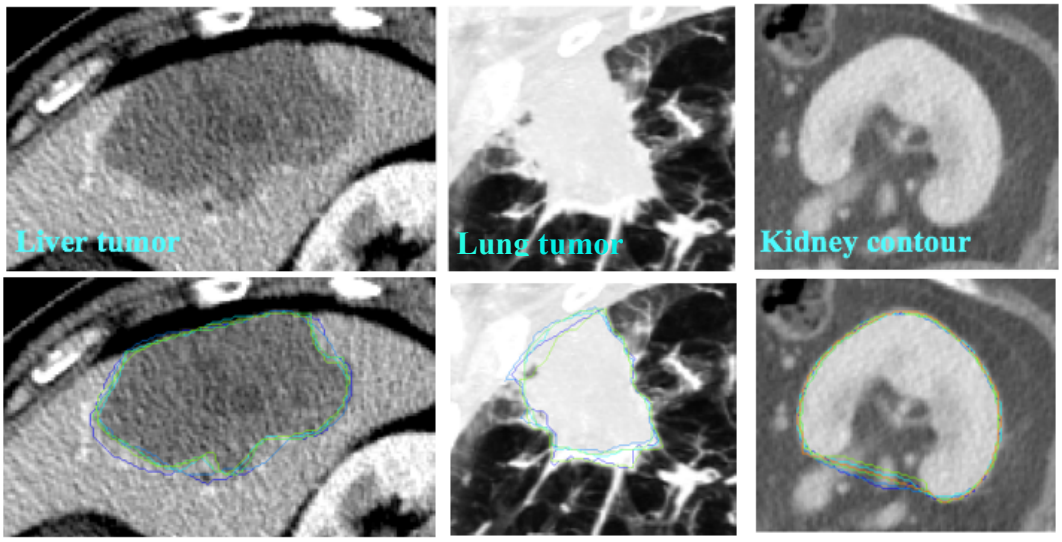

Fig. 1: Illustration of manual delineations on three anatomical structures: liver tumor, lung tumor, and kidney contour. Shown are representative axial slices with superimposed manual delineations in which each color shows the delineation of one radiologist.

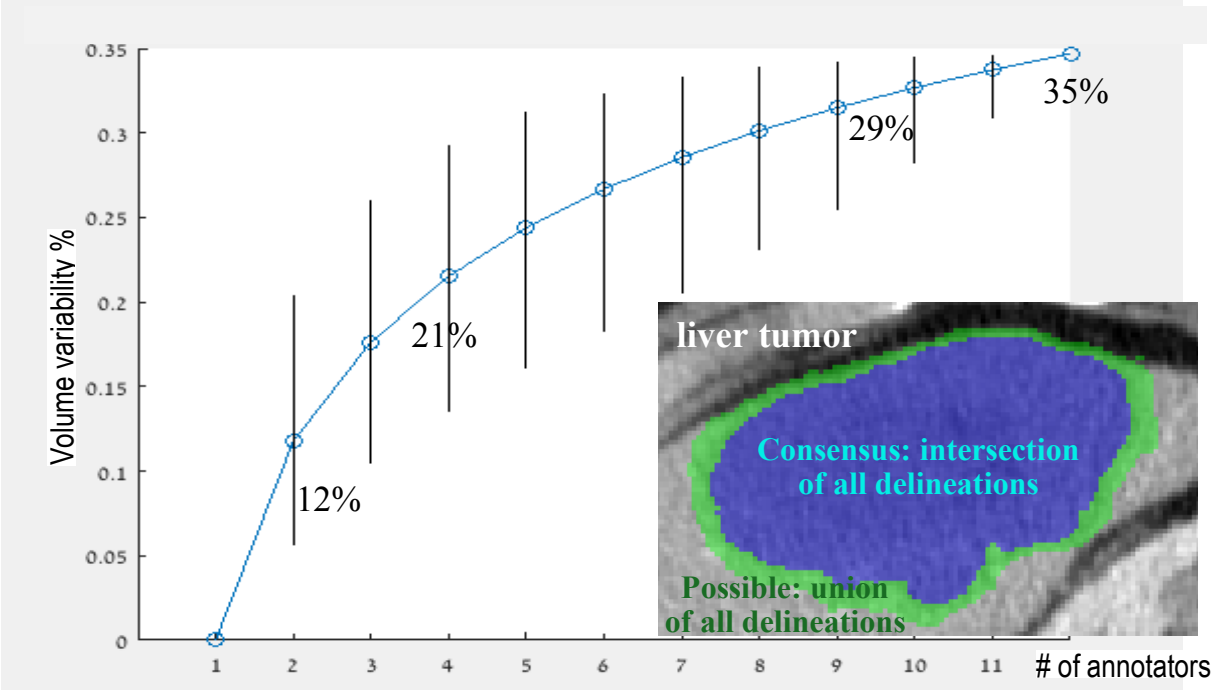

Fig. 2: Graph showing the variability of the liver tumor delineation volume in a sample CT slice (insert, lower right) as a function of the number of observers. The volume variability is defined as the difference between the union (possible, green) and the intersection (consensus, blue) delineation sets of voxels inside the contour computed from k = 1,2,...,11 observers delineations divided by the mean tumor volume from 11 observer delineations. The vertical bars indicate the observer variability of the k^{th} subgroup -- maximum and minimum variability of k observers out of 11.