European Radiology (2019) 29:1391–1399
https://doi.org/10.1007/s00330-018-5695-5

COMPUTED TOMOGRAPHY

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

Leo Joskowicz[1] · D. Cohen[1] · N. Caplan[2] · J. Sosna[2]

Received: 26 May 2018 / Revised: 9 July 2018 / Accepted: 31 July 2018 / Published online: 7 September 2018

Abstract

Purpose To quantify the inter-observer variability of manual delineation of lesions and organ contours in CT to establish a reference standard for volumetric measurements for clinical decision making and for the evaluation of automatic segmentation algorithms.

Materials and methods Eleven radiologists manually delineated 3193 contours of liver tumours (896), lung tumours (1085), kidney contours (434) and brain hematomas (497) on 490 slices of clinical CT scans. A comparative analysis of the delineations was then performed to quantify the inter-observer delineation variability with standard volume metrics and with new group-wise metrics for delineations produced by groups of observers.

Results The mean volume overlap variability values and ranges (in %) between the delineations of two observers were: liver tumours 17.8 [-5.8,+7.2]%, lung tumours 20.8 [-8.8,+10.2]%, kidney contours 8.8 [-0.8,+1.2]% and brain hematomas 18 [-6.0,+6.0] %. For any two randomly selected observers, the mean delineation volume overlap variability was 5–57%. The mean variability captured by groups of two, three and five observers was 37%, 53% and 72%; eight observers accounted for 75–94% of the total variability. For all cases, 38.5% of the delineation non-agreement was due to parts of the delineation of a single observer disagreeing with the others. No statistical difference was found for the delineation variability between the observers based on their expertise.

Conclusion The variability in manual delineations for different structures and observers is large and spans a wide range across a variety of structures and pathologies. Two and even three observers may not be sufficient to establish the full range of inter-observer variability.

Key Points

- *This study quantifies the inter-observer variability of manual delineation of lesions and organ contours in CT.*
- *The variability of manual delineations between two observers can be significant. Two and even three observers capture only a fraction of the full range of inter-observer variability observed in common practice.*
- *Inter-observer manual delineation variability is necessary to establish a reference standard for radiologist training and evaluation and for the evaluation of automatic segmentation algorithms.*

Keywords Humans · Observer variation · Reproducibility of results

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-018-5695-5) contains supplementary material, which is available to authorized users.

✉ Leo Joskowicz
josko@cs.huji.ac.il

[1] The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, 9190401 Jerusalem, Israel

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

Abbreviations

CT	Computed tomography
HD LED	High-definition light-emitting diode
IRB	Institutional review board
MDCT	Multiple-detector computed tomography

Introduction

In cross-sectional imaging, contour delineation and volumetric measurements of anatomical structures in volumetric scans

are currently considered part of the standard of care in clinical practice [1–4]. True volumetric measurements have been shown to provide superior discriminating power as compared to linear measurements and to approximate volumetric measurements based on them [5]. However, volumetric measurements require the delineation of the contours of the structure of interest in each slice of the volumetric scan. Manual contour delineation is tedious, time-consuming, error-prone and requires expertise—it is currently only performed by experienced radiologists and technicians when deemed necessary [6]. Several computer-based software tools are available and are sometimes used for automatic contour delineation and volumetric measure of a variety of anatomical structures and pathologies [7, 8].

Manual structure delineation is currently required to establish a reference standard for volumetric measurements used in clinical decision making and for the evaluation of automatic segmentation algorithms. However, manual delineation is subject to inter-observer variability, and depends on objective and subjective factors. Objective factors include the imaging and data processing protocols, the scan resolution and contrast, the partial volume effect, structure location, contour characteristics and neighbouring structures. Subjective factors include the observer manual steadiness, the hand–eye coordination skills, and the attentiveness, thoroughness, expertise and knowledge of the observer. Automatic structure delineation is subject to algorithm and programming bias, i.e., what radiologist knowledge was elucidated and how it was encoded, or what tagged datasets were used to train a classification convolutional neural network. Whether it was produced manually of automatically, a reference standard of the structure contour is required to determine the validity and accuracy of a given delineation. However, a true reference standard with known volumes can only be available for phantom studies. A surrogate might be the mean contour computed from manual delineations by several observers [9]. Since the mean contour is also subject to inter-observer variability, it is also important to quantify it.

The purpose of this study was to quantify the inter-observer variability of manual delineations of a few common lesion and organs in CT slices for a group of independent observers. The hypothesis was that there can be significant variability between the delineations of two randomly selected observers and that the variability of manual delineations for different structures and observers is large and spans a wide range across a variety of structures and pathologies.

Materials and methods

A group of 11 radiologists produced a total of 3193 delineations of the contours of liver tumours (896), lung tumours (1,085), kidney contours (434) and brain hematomas (497) on 490 slices of clinical CT scans. The CT slices were selected from 18 clinically indicated studies (cases) performed on 64–256-slice MDCT scanners (Philips Healthcare): liver tumours (5, all slices), lung tumours (5, all slices), kidneys (6, selected slices) and brain hematomas (2, all slices). The studies were chosen to represent pathologies with varying delineation difficulty due to fuzzy boundaries (liver tumours), varying tumour contrast (lung tumours, brain hematomas) and normal anatomical structures (kidneys). IRB waiver for informed consent was obtained. The CT resolutions of the scans including the liver tumours, lung tumours and left kidney contours was $512 \times 512 \times 350–449$, $0.68–0.98 \times 0.68–0.98 \times 1.5–3.3$ mm^3 with slice spacing of 1.5 mm for 14 scans and 3.3 mm for 2 scans and $512 \times 512 \times 110$, $0.49 \times 0.49 \times 1.5$ mm^3 for the brain hematomas. The studies of the lung and liver tumours as well as the kidney contours were performed after the intravenous injection of 90–120 cc of iodinated contrast injection (Iomeron 300, Bracco, UK). Portal phase imaging (70-s delay) was used for the chest and liver abnormalities and the nephrographic phase (90-s delay) was used for kidney delineation. Brain imaging for hematoma evaluation was performed without contrast.

The radiologists recruited for the study had varying degrees of expertise: 4 were radiology residents, 2 were mid-career radiologists and 5 were experts with 10+ years of experience. One expert radiologist served as the clinical coordinator (NC); a graduate computer science student (DC) served as the technical coordinator. The manual delineation tasks were performed according to pre-established delineation guidelines. To avoid commercial bias, the manual delineations were performed with the ITK-SNAP v3.4 open software [10] on an HD LED screen.

The following five-phase manual delineation protocol was established: 1) preparation, in which the datasets were collected, the contrast adjustment was set for all slices to ensure uniformity and to avoid bias, the delineation criteria were established, i.e., what to include/not include, and the software use guidelines and work schedule were developed; 2) recruitment of the observers with a 1-hr tutorial session by the technical coordinator; 3) first blinded delineation round of all datasets; 4) delineations evaluation and second delineation round, and; 5) analysis of the results.

The radiologists annotated all slices in 8 hours within 2 weeks. All but two radiologists reported that the allotted delineation time, ~1 minute per slice, was sufficient. The other two reported that the actual delineation time was about 50% longer than anticipated. The delineations were verified by the clinical coordinator to ensure that they had no errors and that they conformed with the delineation protocol. This approach was chosen to ensure that the variability was indeed due to personal differences in delineations and not due to a misunderstanding of the guidelines. For each delineation and its corresponding CT slice, the delineation contour was superimposed on the original image. The coordinator then

either approved the delineation or recorded the probable cause of inconsistency, and corrected the errors when these were small and few. For the remaining delineations, the coordinator requested the radiologist that produced it to repeat the delineation based on the recorded probable cause. A total of 216 (7%) delineations required revisions. The main reasons for revisions were: 1) the inclusion of fat, blood vessels or parts of adjacent structures (51 slices); 2) the exclusion of structure parts due to fuzzy boundaries, or contrast/texture similar to that of adjacent structures (67 slices) and; 3) miscellaneous mistakes, e.g. lack of proper understanding of the guidelines, poor hand–eye coordination and other unknown errors (98 slices). The resulting set of 3193 delineations was free of errors, representing valid structure delineations. Figure 1 shows representative examples of delineations.

A comprehensive comparative analysis of the delineations was then performed to characterise the inter-observer delineation variability. We used standard volumetric metrics and new group-wise volumetric metrics for delineations produced by groups of observers. Volumetric metrics were chosen to compare the delineations because they are of clinical relevance and because they are frequently used to evaluate the results of automatic segmentation algorithms.

The Supplement lists the metrics definitions and formulas. Briefly, the volume of a delineation in a slice is the number of voxels corresponding to the image pixels on or inside the delineation contour times the volume of a voxel. The volume of delineations in multiple slices is the sum of the delineation volumes of each slice. The volume difference between two delineations is the difference between their volumes. The volume overlap similarity of two delineations (Dice index) is the ratio between the volume of the voxels on or inside both delineations and the mean of the individual delineation volumes (a number between 0 and 1). The volume overlap difference is 1 minus the volume overlap similarity. When the delineations are free of errors, it corresponds to the delineation volume overlap variability between the delineations. The mean delineation volume of a group of delineations is the mean of the single delineation volumes.

We introduce new group-wise metrics to further compare groups of delineations. A delineation d_k of a structure in a volumetric scan is the set of voxels on the structure boundary. Given a set $D_n = \{d_1, \ldots, d_n\}$ of n delineations of the same structure, we define three sets: $Consensus(D_n)$, the set of voxels that are included in all delineations, $Possible(D_n)$, the set of voxels that were included in at least one delineation and $Variability(D_n)$, the set of voxels for which delineations differ - the set difference between $Possible(D_n)$ and $Consensus(D_n)$. Figure 2 and Fig. 1A(d) in the Supplement illustrate these concepts.

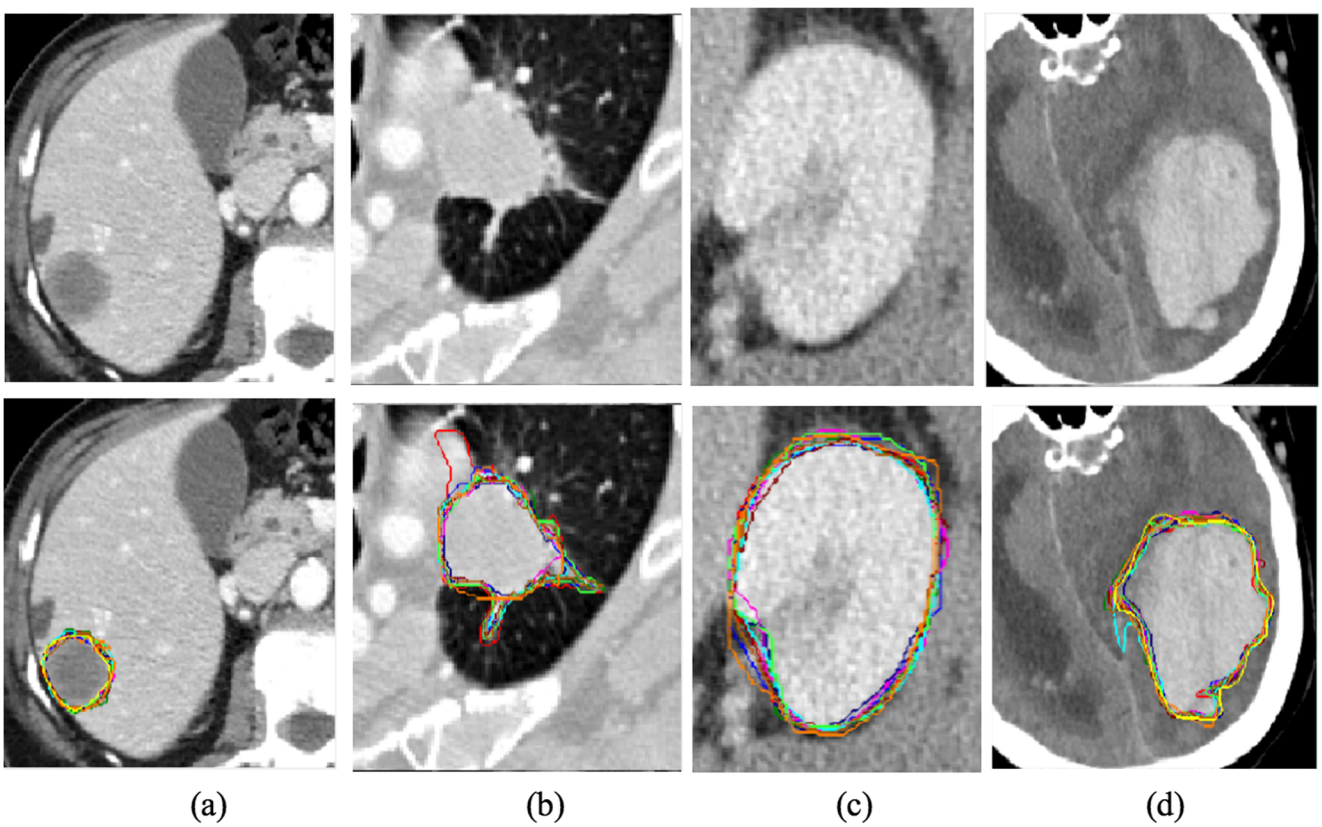

Fig. 1 Top: representative axial CT slices for each of the study structures: (**a**) liver tumour; (**b**) lung tumour; (**c**) kidney contour; (**d**) brain hematoma. Bottom: corresponding manual delineations of 11 observers superimposed on the slices; each colour corresponds to the delineation of one radiologist

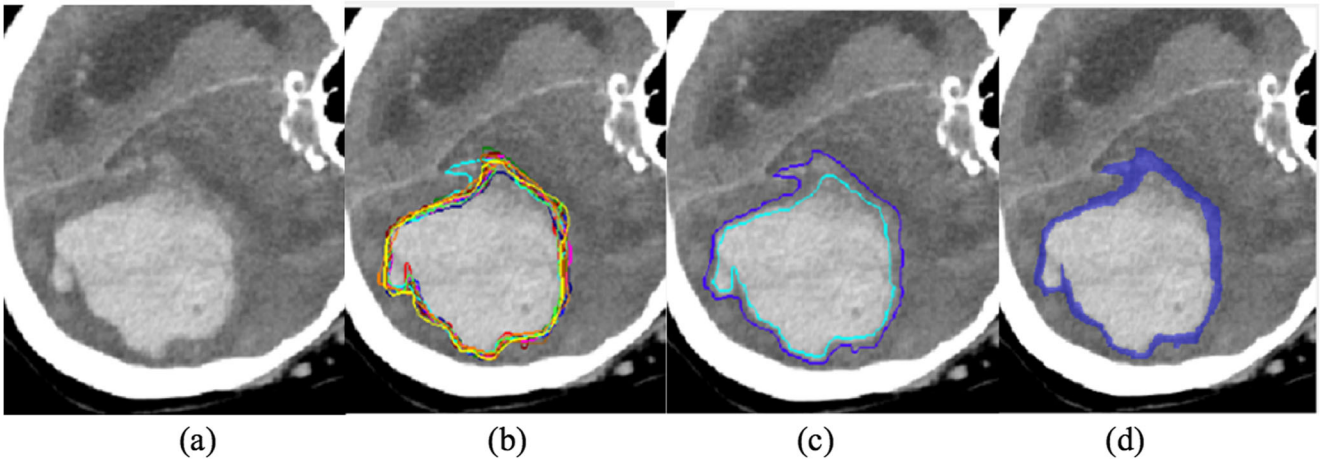

Fig. 2 Illustration of the *Possible*, *Consensus* and *Variability* sets: (**a**) representative slice of a CT scan slice showing a brain hematoma; (**b**) 10 manual delineations of the brain hematoma, $D_{10} = \{d_1, \ldots, d_{10}\}$; (**c**) contours of the *Possible*(D_{10}) set (dark blue) and *Consensus*(D_{10}) set (light blue) overlaid on the original axial CT slice; (**d**) *Variability*(D_{10}) set (translucid purple) superimposed on the original axial CT slice

Note that by definition, all these volume sets depend on the number of delineations n. More delineations (higher values of n) yield a more comprehensive and, hence, more accurate characterisation of the actual delineation volume overlap variability. The volume sets sizes increase monotonically as the number of delineations grows. The mean volume overlap variability of a group of delineations is the mean of the volume of the group delineation variability divided by the mean delineation volume. Comparison differences were determined with the Student's t test. Significance was determined for $p < 0.05$.

The analysis was performed along six categories: 1) Manual tracing variability, to quantify the delineation volume overlap variability due to manual tracing and hand-eye coordination; 2) Pairs of observers, to determine the delineation volume overlap variability between two delineations of the same structure by two observers; 3) Case and structure type, to determine the delineation volume overlap variability for each case and for each structure type; 4) Expertise of observers, to determine the delineation volume overlap variability by observer expertise; 5) Groups of observers, to determine the delineation volume overlap variability for groups of delineations; 6) Non-agreement between observers, to determine the delineation volume overlap variability that results from delineation disagreement between observers. All analyses were performed with standard and custom code in MATLAB R2017b (The MathWorks Inc.). Note that one observer did not complete all the delineation tasks, so some results are reported for 10 observers while others for 11 observers. For the full details of the study, see [11].

Results

Manual tracing For each case, we computed the minimum delineation variability of D_n and identified the CT slice for which the delineation volume overlap variability is the smallest (Appendix A4). For each structure type, we computed the mean, minimum and maximum values for the slice. The mean volume overlap variability values and ranges (below and above the mean) in percentages were: liver tumour 27.1% [-13.3, +13.8], lung tumour 26% [-13.4, +13.2], kidney contour 16.0% [-7.7, +8.3] and brain hematoma 27.5% [-12.4, +15.1]. The kidney cases had the lowest variability—about ±2 pixels from the mean delineation contour—because the kidney boundary appeared clearly in the images.

Pairs of observers For each case, we computed the mean, minimum and maximum volume difference and volume overlap variability between any two randomly selected delineations. Figure 3 shows the results. The mean delineation volume difference values and ranges (below and above the mean) in percentages were: liver tumour 7.0% [-3.0, +7.0], lung tumour 10.2% [-6.2, +10.8], kidney contour 4.1% [-1.1, +0.9] and brain hematoma 9.0% [-5.0, +5.0]. The mean volume overlap variability values and ranges (below and above the mean) in percentages were: liver tumours 17.8% [-5.8, +7.2], lung tumours 20.8% [-8.8, +10.2], kidney contours 8.8% [-0.8, +1.2] and 18% [-6.0, +6.0]. The largest variability of a single case for each structure type was: liver tumour 40%, lung tumour 37%, kidney contour 13% and brain hematomas 31%.

Case and structure type For each case, we computed the volume overlap variability between the individual delineations and the delineation volume of the mean delineation of all observers in each axial CT slice. We grouped the delineation volume variability results for individual cases and for structure types (Table 1). The mean volume overlap variability values and ranges (below and above the mean) in percentages were: 7% [-3, +2] for liver tumours, 9% [-4, +4] for lung tumours, 4% [0, +1] for kidney contours and 7% [-2, +2] for brain hematomas.

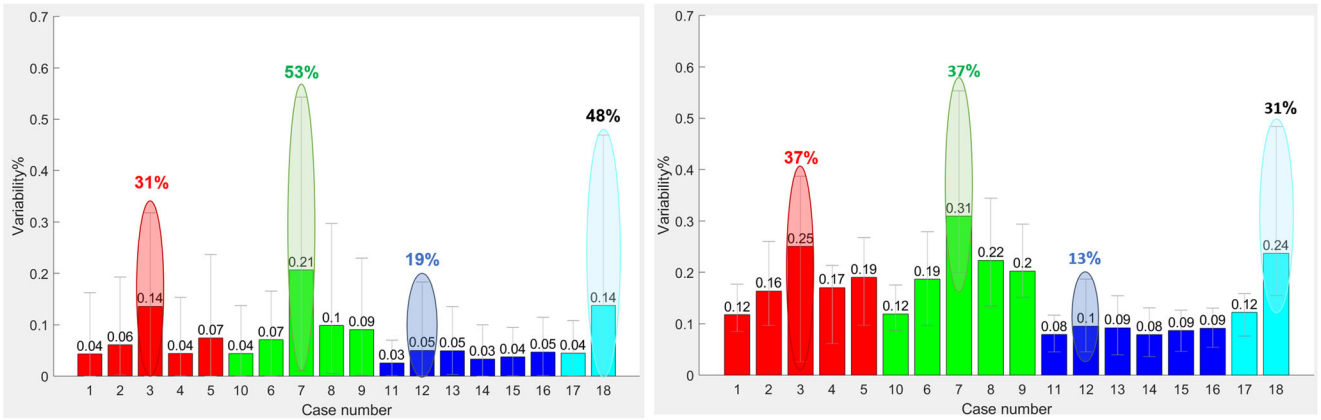

Fig. 3 Pairs of observers: (**a**) delineation volume difference and (**b**) delineation volume overlap difference variability for any pair of observers per case. The horizontal axis shows the liver tumours (red), lung tumours (green), kidney contours (dark blue) and brain hematomas (light blue). For each case, the mean (number above the bar) and the minimum and maximum range (grey line) are shown

Expertise of observers For each observer and structure type, we computed the volume overlap variability between the individual delineations volume and the mean delineations volume (Table 2). For liver tumours, the mean volume overlap variability values and ranges (below and above the mean) in percentages were: 6% [-1, 0] for radiology residents, 7% [-1, +1] for mid-career radiologists and 7% [-2, +2] for experts. For lung tumours, it was 7% [-1, +1] for residents, 9% [-4, +3] for mid-career radiologists and 9% [-2, +1] for experts. There was no statistical difference ($p < 0.05$) between the groups.

Groups of observers For each case, we computed the mean, minimum, maximum and *Consensus* and *Possible* volume overlap variability between groups of k observers for all values $2 \leq k \leq 10$.Figure 4 shows the results for lung tumours. For $k = 1$, there is no variability; for $k = 2$, the variability is that of all pairs of observers (45 = 10 choose 2 possible pairs), as reported in item 2 above. We repeated this procedure for $k = 3$ (triples of observers), $k = 4$, etc. For $k = 10$ observers, there is a single group, so the minimum, maximum and mean variability are the same. The mean volume overlap variability values and ranges (*Consensus* and *Possible* sets) in percentages for all observers were: liver tumours 51% [-24, +27], lung tumours 56% [-25, +31], kidney contours 25% [-12, +13] and brain hematomas 53% [-24, +29]. As expected, the *Possible* volume increases, the *Consensus* volume decreases and the *Variability* volume and the maximum and minimum ranges decrease as the number of delineations increases. The trends are similar for all structures (Fig. 4).

Non-agreement between observers We measure the non-agreement between the delineations of two observers as the delineation volume overlap variability resulting from the delineation differences between them. Non-agreement between delineations was defined as the percentage of the variability for which $m = 1,...,5$ delineations differ from the others (Fig. 5a, b). The non-agreement of one delineation with the rest contributed a mean of 37% of the volume overlap variability for liver tumours, 36% for lung tumours, 41% for kidney contours and 41% for brain hematomas. The non-agreement of two delineations contributes a mean of 24% of the variability for liver tumours, 22% for lung tumours, 22% for kidney contours and 22% for brain hematomas. The non-agreement contribution for three, four and five delineations then decreases to 17,

Table 1 Delineation volume overlap variability: delineation volume overlap variability (percentage) for all observers between the individual delineation volume and the delineation volume of the mean of all delineations per structure type and case. Mean values and ranges (below and above the mean) are shown

	Structure type																	
	Liver tumours			Lung tumours						Kidney contours						Brain hematomas		
Case #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
VODV (%)	5	6	10	7	8	9	13	9	9	5	3	4	4	3	3	4	5	9
Mean range (%)	7 [-3, +2]			9 [-4, +4]						4 [-1, +1]						7 [-2, +2]		
	6.5 [-4, +4]																	

VODV volume overlap difference variability

Table 2 Inter-observer delineation volume overlap variability: delineation volume overlap variability (percentage) between the individual delineation volume and the delineation volume of the mean of all delineations per structure type and radiologist experience. Mean values and ranges (below and above the mean) structures are shown

	Radiology residents		Mid-career radiologists				Expert radiologists					All mean (range)
Observer #	r1	r2	m1	m2	m3	m4	e1	e2	e3	e4	e5	
Liver tumours	7	6	6	6	8	7	6	5	9	7	8	7 [-3, +2]
Lung tumours	8	7	8	6	13	9	11	9	8	9	–	9 [-4, +4]
Kidney contours	4	2	4	2	3	–	4	2	4	4	3	4 [-1, +1]
Brain hematomas	7	11	7	6	5	–	5	6	7	7	7	7 [-2, +2]
Mean per radiologist	6.5	6.5	6.3	5.0	7.3	8	6.5	5.5	7.0	6.8	6.0	6.5 [-4, +4]
Mean per expertise	6.5 [-2, +4]		6.4 [-3, +2]				6.4 [-4, +4]					

15 and 7% for all structures, respectively. Trends for all structures are similar (Table 3).

Discussion

Our evaluation of the manual tracing variability indicates that the mean smallest delineation volume overlap variability due to manual tracing and hand–eye coordination is 16% for a CT with resolution 0.75 × 0.75 × 1.5 mm^3. In our study, it corresponds to the healthy kidney contour, which requires minimal medical knowledge to delineate it. This establishes the ground-truth minimal delineation variability for the reference standard and for segmentation algorithms validation. Note that it may be a better surrogate measure for observer variability than one derived from a phantom scan, which would quantify the manual delineation variability due to CT scan resolution and hand–eye coordination and will leave out the sources of variability that are of clinical importance, e.g., structure characteristics, neighbouring structures, radiologist expertise and knowledge of complex medical pathologies.

Our evaluation of the delineation volume overlap variability between randomly selected pairs of observers indicates that the variability may differ significantly from one pair of observers to another. For the kidney contours, it is about half of that of the other structures. This is due to the size of the kidney and its clear and unambiguous contours in the CT scans. Overall, the volume overlap variability range between two observers is very wide, from 5 to 57%. Note that the volume overlap difference is on average ×2.35 larger than the volume difference. This is explained by the fact that the volume difference does not take into account the location of the voxels, while the volume overlap difference does (Fig. 1A, Supplemental Material).

We note that the delineation variability between the radiologists based on their expertise is not statistically significant. This means that a reliable structure delineation can be obtained from less experienced radiologists followed by validation and correction of a few delineations (7% in our study) by a more experienced radiologist. This suggests that manual structure delineation with minimal training may be outsourced and then revised by a radiologist. Note that for liver tumours, lung tumours and brain hematomas, there is no observer bias across all cases. However, there is a small bias for kidney contours, most likely due to the manual tracing skills of each observer.

With regards to groups of delineations, we observe that the volume overlap variability for a large group of delineations is wide and differs significantly between structures. This means that measures derived from the delineation of a single observer or from a segmentation algorithm may exhibit these wide variability ranges. The delineation volume overlap variability increases as a function of the number of observers in groups. On average, only 37% of the volume overlap variability is captured by the delineations of two observers. It increases to about 53, 72 and 85% for 3, 5 and 8 delineations, respectively, reaching 100% for all delineations. After nine delineations, the contribution of each additional delineation is less than 5%. Therefore, to establish a proper observer variability reference, delineations from multiple different observers may be necessary.

When analysing delineation agreement between observers, we note that ~40% of the delineation volume variability results from the non-agreement of one observer with the others, and 20% from the non-agreement of two observers with the others. This means that the delineation volume overlap variability is

Table 3 Non-agreement between annotators: volume overlap variability in percentages that results from delineation differences between observers. The discrepancy between delineations is the percentage of the variability for which $m = 1,...,5$ observers differ with the others

Structures	Disagreement between observers (#)				
	One	Two	Three	Four	Five
Liver tumours	37	24	17	15	7
Lung tumours	36	22	17	18	8
Kidney contours	41	22	17	14	7
Brain hematomas	41	22	17	17	7
All structures	38.5	22.8	17.0	15.7	7.3

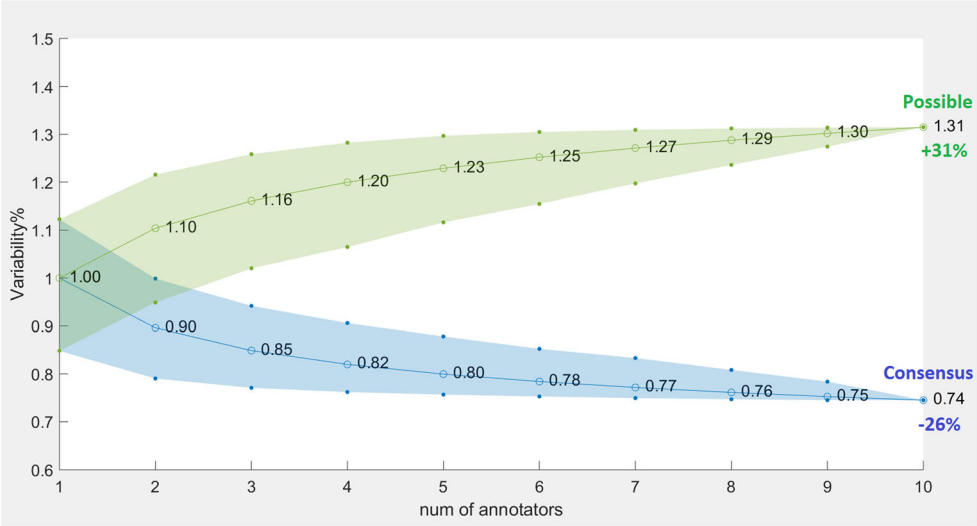

Fig. 4 *Variability* volume (vertical axis) as a function of the number of observers (horizontal axis) for lung tumours. The upper and lower areas (green and blue) show the mean *Possible* and *Consensus* volume variability (mean, unfilled circle, maximum and minimum filled circles). The variability volume was normalised to 1 (no variability) to allow comparisons across studies

significantly influenced by a single observer, while ~7% of the variability is caused by the disagreement between half of the observers. This indicates that the number of observers involved in determining the mean delineation matters, and that two and even three observers may not be sufficient to establish the full range of inter-observer variability.

Our results indicate that the non-agreement between observers follows a similar pattern for all four structures. Figure 5 shows that the non-agreement variability between observers monotonically increases with the number of observers at the same rate—the variability interval is different (Fig. 5a) but the trend is similar (Fig. 5b).

Some of the practical implications of our study are as follows. Radiologists should be aware that there may be significant differences in structure contour delineations and volumetric measurements derived from them. The differences might stem from objective and subjective factors, e.g., the imaging and data processing protocols, the scan resolution and contrast, the structure location and contour characteristics and the observer's hand–eye coordination skills. Of note is that the delineation volume overlap difference between two randomly selected observers may be significant. While more accurate and less variable than volume estimations derived from linear measurements, delineation variability should be taken into account when making clinical decisions based on volumetric measurements and when evaluating automatic segmentation tools. Moreover, when assessing computerised tools against a reference standard, the number of readers should be reported.

Several observer contour delineation variability studies are reported in the literature. Meyer et al [12] describe a study that identified the source of variability in lung MDCT nodule

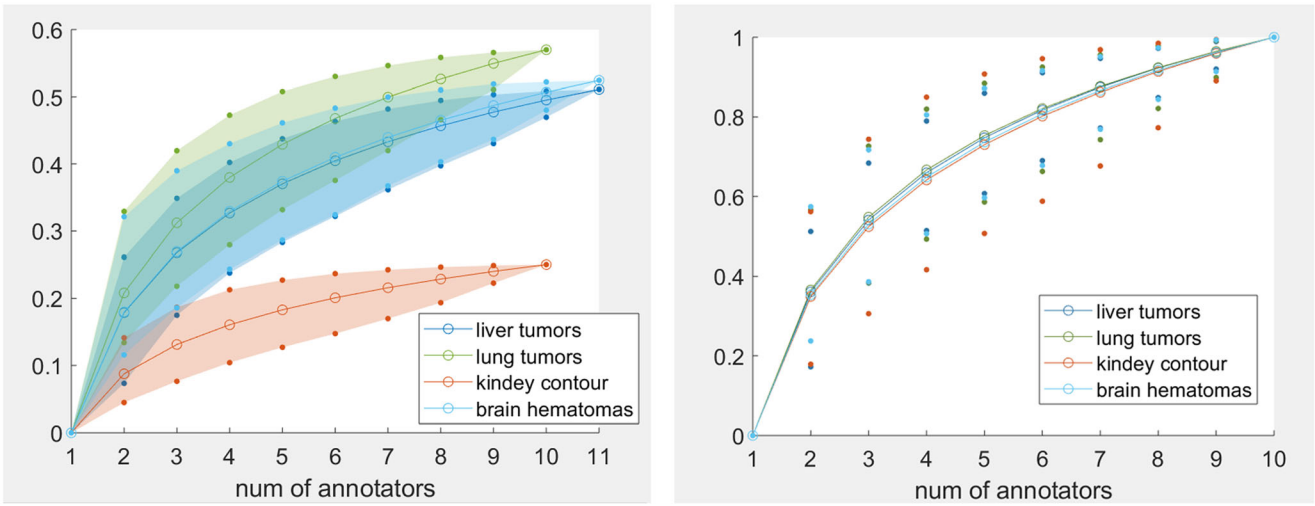

Fig. 5 Convergence rates of the delineation volume overlap variability for all four structures (one colour for each structure): (**a**) sum of the *Possible* and *Consensus* volume variability as a function of the number of annotators. The upper and lower areas (green and blue) show the *Possible* and *Consensus* (mean, unfilled circle, maximum and minimum filled circles); (**b**): Normalised *Variability* ratio, which represents the variability volume to the interval [0, 1]

annotations. It reports that radiologists are the major source of variability regardless of the manual delineation tools used. Bø et al [13] investigated the intra-observer delineation variability of low-grade gliomas. Gurari et al [14] evaluated the performance of trained experts, crowd-sourced non-experts, and algorithms in the delineation of structures of six datasets. Irshad et al [15] compared the concordance of annotations from four types of observers: crowd-sourced annotations, research fellows, experts and automated methods. None of these studies quantifies the delineation variability for more than four observers. Helm et al [16] evaluated the variability of 29,000 crowd-sourced annotations for non-medical images. Valindria et al [17] described a method to predict automatic segmentation performance without ground truth.

Possible limitations to our study include heterogeneity of pathologies and structures on one side and the relatively small number of studies evaluated. However, the total of 3193 delineations of 490 slices is a large enough sample to derive statistically significant comparisons. The datasets and observers come from the same institution. The scans were acquired under similar conditions on the same scanners. Radiologists come from the same institution, so there may be a hidden group bias. The number of datasets for brain hematomas was only two, so the corresponding findings may not be indicative. The validation stage was performed by a single expert radiologist. The variability results do not distinguish between the three sources of delineation variability: scan-specific, manual tracing and, to a lesser degree, medical knowledge.

On a broader context, we note that the important issue of what constitutes an acceptable variability in clinical practice is orthogonal to the findings or our study. Indeed, each clinical task and case may have a different threshold. For example, for the initial diagnosis of tumours in the liver, the most important clinical finding is the presence of the tumour and its approximate size, e.g., small, medium or large. On the other hand, in a lung tumour longitudinal study for radiotherapy response to treatment assessment, a more accurate measure of the volume difference between the baseline and the follow-up tumour is required. And even in this case, if the change is significant, e.g., when the tumour volume is doubled, then a lower accuracy may be acceptable.

In conclusion, our study indicates that the delineation volume overlap variability for different structures and observers is large and spans a wide range across a variety of structures. The delineation volume variability range for two observers is 5–57%, which is very wide. Two and even three observers capture only a fraction of the full range of inter-observer variability and may not be sufficient to establish a reliable reference standard for radiologist training and evaluation and for the evaluation of automatic segmentation algorithms.

Acknowledgements We thank Dr. Alexander Benstein and the team of radiologists of the Department of Radiology Hadassah Hebrew University Medical Center, Jerusalem, Israel, for their participation in the manual delineation project. We also thank Dr. Tammy Riklin-Raviv, Ben Gurion University of the Negev, for providing the brain CT scans and the brain hematoma delineations.

Funding This study has received partial funding from the Israel Ministry of Science, Technology and Space, grant 53681, 2016-19, and by the Oppenheimer Applied Research Grant, The Hebrew University, TUBITAK ARDEB grant no. 110E264, 2015-16.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Leo Joskowicz.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the institutional review board.

Ethical approval Institutional review board approval was obtained.

Methodology
- retrospective
- experimental
- performed at one institution

References

1. Nanda A, Konar SK, Maiti TK, Bir SC, Guthikonda B (2016) Stratification of predictive factors to assess resectability and surgical outcome in clinoidal meningioma. Clin Neurol Neurosurg 142:31–37
2. Vivanti R, Szeskin A, Lev-Cohain N, Sosna J, Joskowicz L (2017) Automatic detection of new tumors and tumor burden evaluation in longitudinal liver CT scan studies. Int J Comput Assist Radiol Surg 12(11):1945–1957
3. Bhooshan N, Sharma NK, Badiyan S et al (2016) Pretreatment tumor volume as a prognostic factor in metastatic colorectal cancer treated with selective internal radiation to the liver using yttrium-90 resin microspheres. J Gastrointest Oncol 7(6):931–937
4. Abbara S, Blanke P, Maroules CD et al (2016) SCCT guidelines for the performance and acquisition of coronary computed tomographic angiography: A report of the society of Cardiovascular Computed Tomography Guidelines Committee: Endorsed by the North American Society for Cardiovascular Imaging (NASCI). J Cardiovasc Comput Tomogr 10(6):435–449
5. Greenberg V, Lazarev I, Frank Y, Dudnik J, Ariad S, Shelef I (2017) Semi-automatic volumetric measurement of response to chemotherapy in lung cancer patients: How wrong are we using RECIST? Lung Cancer 108:90–95
6. Pupulim LF, Ronot M, Paradis V, Chemouny S, Vilgrain V (2017) Volumetric measurement of hepatic tumors: accuracy of manual

contouring using CT with volumetric pathology as the reference method. Diagn Interv Imaging S2211-5684(17):30282–30286
7. Cai W, He B, Fan Y, Fang C, Jia F (2016) Comparison of liver volumetry on contrast-enhanced CT images: one semiautomatic and two automatic approaches. J Appl Clin Med Phys 17(6):118–127
8. Haas M, Hamm B, Niehues SM (2014) Automated lung volumetry from routine thoracic CT scans: how reliable is the result? Acad Radiol 21(5):633–638
9. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23(7): 903–921
10. ITK-SNAP open software. http://www.itksnap.org/pmwiki/pmwiki.php. Accessed Jul 8 2018.
11. Cohen D (2017) Segmentation variability estimation in medical image processing: framework, method and study. MSc Thesis. The Hebrew University of Jerusalem Israel
12. Meyer CR, Johnson TD, McLennan G et al (2006) Evaluation of lung MDCT nodule annotation across radiologists and methods. Acad Radiol 13(10):1254–1265
13. Bø HK, Solheim O, Jakola AS, Kvistad KA, Reinertsen I, Berntsen EM (2017) Intra-rater variability in low-grade glioma segmentation. J Neurooncol 131(2):393–402
14. Gurari D, Theriault D, Sameki M, et al (2015) How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. Proc IEEE Winter Conference on Applications of Computer Vision, pp 1169–1176
15. Irshad H, Montaser-Kouhsari L, Waltz G et al (2015) Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. Pac Symp Biocomput, pp 294–305
16. Helm E, Seitel A, Isensee F et al (2018) Clickstream analysis for crowd-based objects segmentation with confidence. IEEE Trans Pattern Anal Mach Intell, to appear. https://doi.org/10.1109/TPAMI.2017.2777967
17. Valindria VV, Lavdas I, Bai W et al (2017) Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans Med Imaging 36(8):1597–1606