MIRADA
The Imaging Software People

Assessment of thoracic auto-contouring using a modified Turing Test

M. J. Gooding

Introduction

A wide range of auto-contouring methods have been proposed within radiotherapy and, with each technical improvement, the clinical benefit must be evaluated.

Perhaps the most common approach, with the least clinical burden, is quantitative assessment of the difference between the result and the ground truth contours (as in [1]). Unfortunately, while this approach demonstrates technical improvement, it does not prove clinical benefit. The most direct assessment of the impact of auto-contouring on clinical practice is in the comparison of the time required to adjust the auto-contours for clinical use against the time required to create contours in the current clinical practice (as in [2]). However, this approach also has the highest clinical burden. A less time consuming approach is scoring by clinicians of the utility of contours (as in [3]), however this approach remains subjective and qualitative.

Critically, all of these approaches to validation are hindered by inter-observer variability. For example, a clinician may still request editing to be carried out when reviewing the manual contours of another clinician (as in [4]).

To overcome this limitation, instead of trying to answer the question “is this contour perfect?”, the problem can be reframed as “is this contour as good as one drawn by a clinical expert?”. Therefore, we used an evaluation framework for assessing contour quality based on a variation of the Turing Test.

The imitation game

In 1950, Alan Turing sought to answer the question "Can machines think?" [5]. In his famous paper, he introduced the Imitation Game, sometimes referred to as the Turing Test, and we have reformulated the original question in our work. While the original proposal is quite complex, it has been popularized as the more simple formulation where an interrogator communicates blindly with a single subject and attempts to determine whether the subject is a human or a machine, such as the approach described in the film *Ex Machina* [6]. It is assumed that the machine has performed well if the interrogator makes an incorrect identification as often as they make a correct one.

While there is debate regarding whether the Turing Test is sufficient as a demonstration of intelligence [7], indistinguishability from human behavior can itself be regarded as a performance criterion [8, 9].

In the same way, our goal in assessing auto-contouring is to establish if the system is performing to at least the same standard as a clinical expert. i.e. Are machine-generated contours indistinguishable from human contouring?

Method

Implementation

There are many ways in which such an experiment could be conducted, with a number of design choices. To facilitate multi-center evaluation, a website (www.autocontouring.com) was set up to perform this experiment. Participants, the ‘interrogators’, are shown individual slices and contours on this website and asked a number of questions outlined in the following section. Single slices were chosen to allow faster assessment by participants and a greater number of samples to be assessed, compared with assessing full structures in 3D. Contours presented to a participant were either from an auto-contouring system or expert clinical contours, and the participant was blinded to this information.

Variations of the question

In the closest sense to the original Turing Test question, the participants are asked "*How was this contour drawn?*" with a choice of two answers: "*By a machine*" or "*By a human*". If the auto-contouring system is indistinguishable from human contouring, a correct identification rate of 50% is expected.

However, such a question only reveals that the interrogator cannot tell the source, and does not assess quality. Therefore, the question "You have been asked to QA these contours for clinical use by a colleague. Would you…" is also asked, with answer options ranging from "Require them to be corrected; there are large, obvious, errors" through to "Accept them as they are; the contours are very precise". This question is framed within the context of clinical QA to acknowledge that minor differences of opinion in contouring may or may not have clinical significances. The question is not whether the contour is perfect, but whether it is good enough for clinical use or requires additional work.

In both of the previous questions, only a single contour is presented to the participant. In this context, it is possible that the judgement as to the source is harder than if both contours are shown. To allow for this, a third question type is shown whereby the participant is shown two contours and asked to judge "Which contour is best?". Again, the question is not about which contour is correct but rather which is preferred, to account for differences in what is considered correct by different clinicians.

Thoracic contour evaluation with DLCExpert^{TM}

The initial evaluation conducted on the website includes imaging data from 40 clinical thoracic cases. Contours were assessed for six organs (lungs, spinal canal, heart, mediastinum envelope and esophagus).

Three sources of contours were included in the evaluation: clinical contours, atlas-based auto-contouring (Mirada Workflow Box, v1.4, Mirada Medical Ltd, Oxford, UK) and DLCExpert (Mirada Medical Ltd, Oxford, UK). The atlas-based contouring used 20 carefully curated atlases, while the DLCExpert was trained on 450 cases. All cases were provided by the same clinical institution and created using the same contouring guidelines.

15 clinical experts from five different institutions participated via the website. Each participant answered 100 questions, with the slice, contour source and question type chosen at random.

Results

The results of the Turing Test-style question suggest that the participants are largely guessing the source of the contours. Results by method of contour creation are shown in Figure 1. For clinical contours, the participants were correct approximately half the time, with a slight bias towards deciding a computer generated the labels (51%). For both auto-contouring methods, the correct identification rate was slightly higher, at around 60%, suggesting that in a few of the slices, the source of the contour was more apparent.

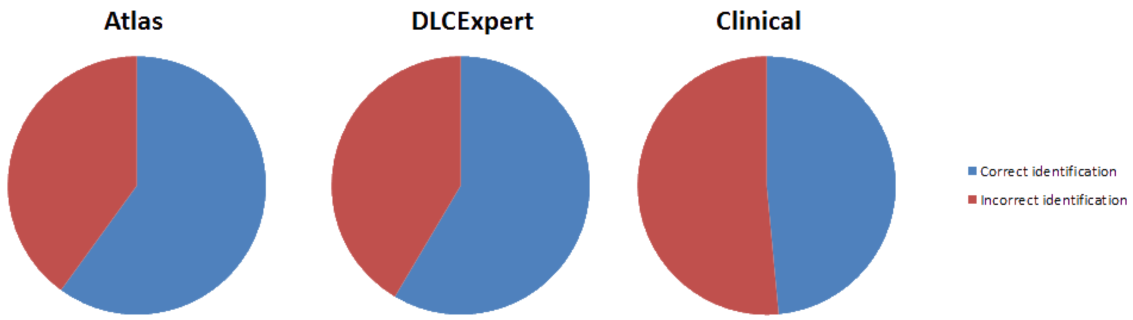

Figure 1: Identification of source of contours by blinded participants. The participants were marginally more likely to say a contour was computer generated, even for clinical contours. Participants were slightly more likely to correctly identify automatically generated contours.

When asked to consider whether the contours are acceptable, the original clinical contours were only accepted about 70% of the time, as shown in Figure 2. The institution of the participant made a noticeable difference. Participants from same institution that generated the contours were willing to accept 76% of the contours, with 50% being considered perfect. However, participants from other institutions were only willing to accept 65% of the contours, with only 30% being considered perfect, as shown in Figure 3. This difference in acceptance rate also has an impact on the acceptance of auto-contours. Contours generated by DLCExpert were considered more acceptable than those generated by atlas-based auto-contouring, although the acceptance rate again depended on the institution of the reviewing participant. Overall, the acceptance rate of DLCExpert was similar to that of the clinical contours.

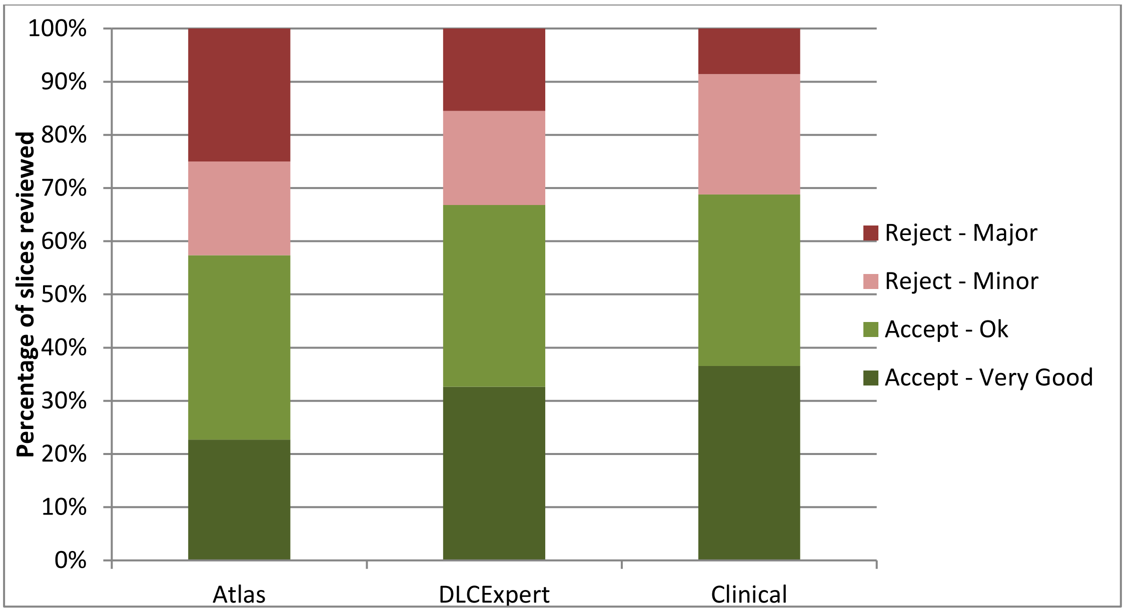

Figure 2: Levels of acceptance based on method of contour generation. Contours generated by DLCExpert achieved levels close to those for clinical contours

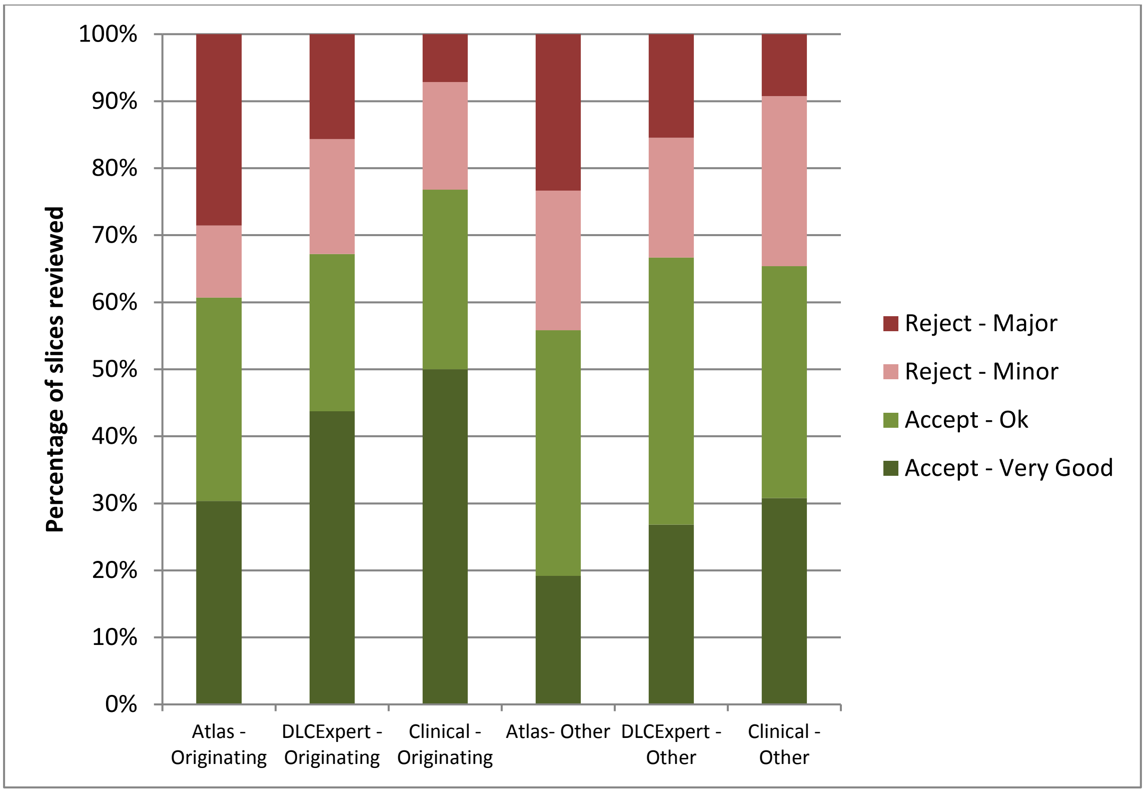

Figure 3: Acceptance levels grouped by method and reviewing institution. Acceptance is higher when participant's institutions match the origin of the contours. The lower acceptance by other institutions demonstrates that contouring style affects acceptance rates. Acceptance rates for DLCExpert remains close to the rates for clinical contours.

When shown contours side-by-side and asked to express a preference, participants preferred clinical contours over atlas-based auto-contours about 69% of the time, as shown in Figure 4. Contours generated by DLCExpert were preferred about 43% of the time over clinically-produced contours. The preference for the auto-contours varied between organs, as shown in Figure 5. For the lung, DLCExpert generated contours were preferred more often than not when compared to the original clinical contours. For both methods, the contours of the spinal canal were considered preferable to the clinical contours about as often as not. The largest difference between the two methods can be seen for the esophagus, where clinical contours were preferred 90% of the time compared to atlas-contours. In contrast, DLCExpert contours were preferred 40% of the time to the clinical ones.

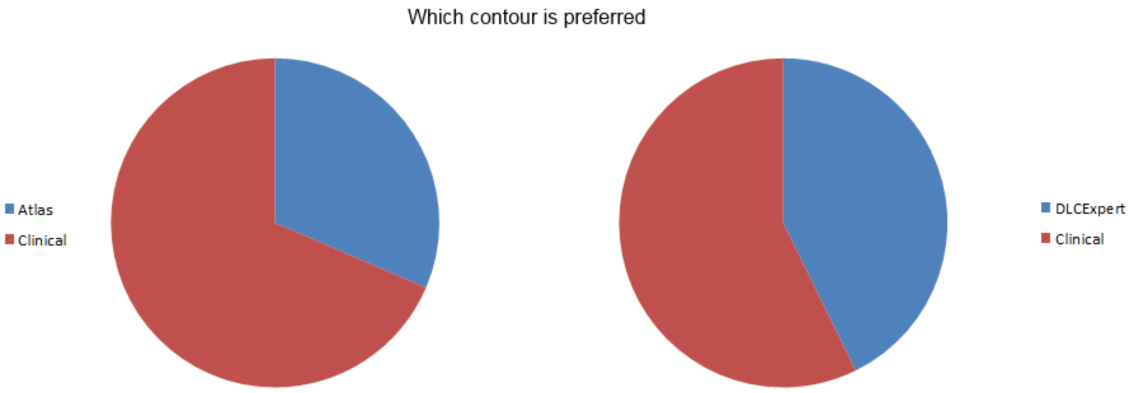

Figure 4: Preference in side by side comparison. Clinical contours are preferred over at Atlas-based auto-contouring ones about 2/3 of the time. DLCExpert contours are preferred to clinical ones almost as often as not.

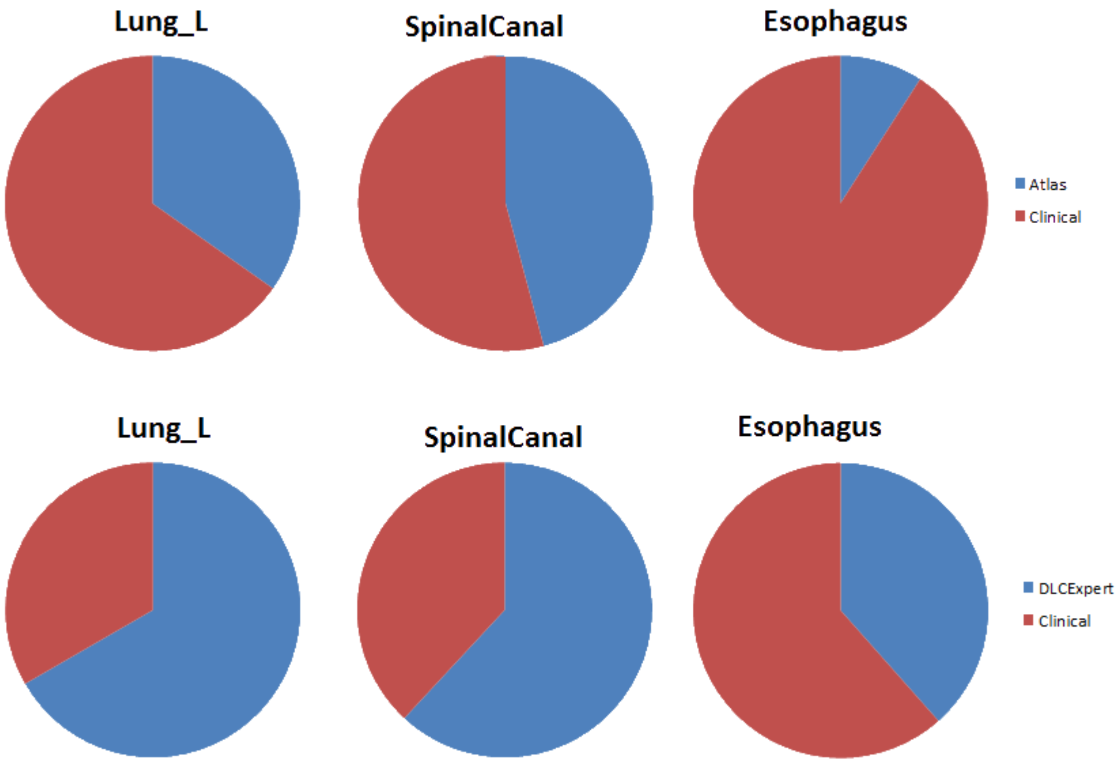

Figure 5: Preference of contours in side by side comparison for selected organs. For the left lung, DLCExpert contours are preferred to the clinically-drawn contours. Performance is similar for both auto-contouring methods for the spinal canal, with the contours being preferred as often as clinical ones. For the esophagus, the contours of DLCExpert are preferred almost half the time, while the Atlas-based auto-contours are generally less desired.

Discussion and Conclusions

Blind assessment of contours provides a method to assess the acceptance, or otherwise, of auto-generated contours within the context of the natural inter-observer variability of clinical contours. It is well known that clinically-drawn contours may be judged to require editing if reviewed at a later date. This study found that this may be true for as much as 30% of contours. Similarly, this work confirms previous observations that different institutions have different contouring styles.

In this study, DLCExpert was found to outperform atlas-based auto-contouring both in terms of clinical acceptance of contours and in the percentage of contours preferred over clinical ones.

Deep learning contouring (DLC) seeks to learn how to contour from clinical examples, and has the potential to generate contours that are much more closely aligned with the clinical standard. This study demonstrated that automatically-generated contours are getting closer to being indistinguishable from clinical contours. Participants were not able to reliably identify the source of the contours and were almost equally willing to accept contours generated by DLCExpert and clinical sources. For some organs assessed the DLCExpert*,* generated contours were actually preferred to contours that have been produced clinically.

Future work

This study is part of a larger ongoing study, which will consider further treatment areas and organs, to establish how close auto-contouring can get to clinical performance. We welcome clinical experts from all institutions to participate as observers in this study at www.autocontouring.com

To ensure the study is representative of multiple institutions, we are also seeking expert institutions to contribute cases for evaluation. Please contact Mirada Medical if you would be interested in taking part.

References

1. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Medical physics*. 2014;41(5):050902.
2. Reed VK, Woodward WA, Zhang L, et al. Automatic segmentation of whole breast using atlas approach and deformable image registration. *International Journal of Radiation Oncology* Biology* Physics*. 2009;73(5):1493-1500.
3. Greenham S, Dean J, Fu CKK, et al. Evaluation of atlas-based auto-segmentation software in prostate cancer patients. *Journal of medical radiation sciences*. 2014;61(3):151-158.
4. Teguh DN, Levendag PC, Voet PW, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *International Journal of Radiation Oncology* Biology* Physics*. 2011;81(4):950-957.
5. Turing AM. Computing machinery and intelligence. *MIND*. 1950;59(236):433-460.
6. Garland A. Ex Machina. Universal Pictures International, 2014.
7. Gunderson K. The imitation game. *Mind*. 1964;73(290):234-245.
8. Harnad S. The Turing Test is not a trick: Turing indistinguishability is a scientific criterion. 1992;3(4):9-10.
9. Saygin AP, Cicekli I, Akman V. Turing Test: 50 Years Later. *Minds and Machines*. 2000;40:463-518.