Comparative evaluation of autocontouring in clinical practice

Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test

Mark J. Gooding,[a)] Annamarie J. Smith, Maira Tariq, Paul Aljabar, and Devis Peressutti
Mirada Medical Ltd, Oxford Centre for Innovation, New Road, Oxford, OX1 1BY, UK

Judith van der Stoep, Bart Reymen, Daisy Emans, Djoya Hattu, Judith van Loon, Maud de Rooy, Rinus Wanders, Stephanie Peeters, Tim Lustberg, Johan van Soest, Andre Dekker, and Wouter van Elmpt
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Dr Tanslaan 12, 6229ET Maastricht, The Netherlands

(Received 29 May 2018; revised 10 September 2018; accepted for publication 10 September 2018; published 12 October 2018)

Purpose: Automated techniques for estimating the contours of organs and structures in medical images have become more widespread and a variety of measures are available for assessing their quality. Quantitative measures of geometric agreement, for example, overlap with a gold-standard delineation, are popular but may not predict the level of clinical acceptance for the contouring method. Therefore, surrogate measures that relate more directly to the clinical judgment of contours, and to the way they are used in routine workflows, need to be developed. The purpose of this study is to propose a method (inspired by the Turing Test) for providing contour quality measures that directly draw upon practitioners' assessments of manual and automatic contours. This approach assumes that an inability to distinguish automatically produced contours from those of clinical experts would indicate that the contours are of sufficient quality for clinical use. In turn, it is anticipated that such contours would receive less manual editing prior to being accepted for clinical use. In this study, an initial assessment of this approach is performed with radiation oncologists and therapists.

Methods: Eight clinical observers were presented with thoracic organ-at-risk contours through a web interface and were asked to determine if they were automatically generated or manually delineated. The accuracy of the visual determination was assessed, and the proportion of contours for which the source was misclassified recorded. Contours of six different organs in a clinical workflow were for 20 patient cases. The time required to edit autocontours to a clinically acceptable standard was also measured, as a gold standard of clinical utility. Established quantitative measures of autocontouring performance, such as Dice similarity coefficient with respect to the original clinical contour and the misclassification rate accessed with the proposed framework, were evaluated as surrogates of the editing time measured.

Results: The *mis*classification rates for each organ were: esophagus 30.0%, heart 22.9%, left lung 51.2%, right lung 58.5%, mediastinum envelope 43.9%, and spinal cord 46.8%. The time savings resulting from editing the autocontours compared to the standard clinical workflow were 12%, 25%, 43%, 77%, 46%, and 50%, respectively, for these organs. The median Dice similarity coefficients between the clinical contours and the autocontours were 0.46, 0.90, 0.98, 0.98, 0.94, and 0.86, respectively, for these organs.

Conclusions: A better correspondence with time saving was observed for the misclassification rate than the quantitative contour measures explored. From this, we conclude that the inability to accurately judge the source of a contour indicates a reduced need for editing and therefore a greater time saving overall. Hence, task-based assessments of contouring performance may be considered as an additional way of evaluating the clinical utility of autosegmentation methods. © *2018 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.13200]

Key words: assessment, autocontouring, editing time, organs-at-risk, Turing test

1. INTRODUCTION

Numerous methods for automatic contouring (segmentation) of structures within imaging have been proposed for a wide range of medical applications. However, there is a need to adequately assess the performance and added value of such tools.[1] In this paper, we focus on radiotherapy applications, where health organs to be spared during treatment, known as organs-at-risk (OARs), are outlined as part of the treatment planning process. However, the assessment method presented may have broader relevance.

Published evaluations of autocontouring (AC) systems broadly fall into three categories; Attempts to quantify clinical impact (e.g., time saving assessment as performed in Ref.

[2]), quantitative measures compared to "ground truth" (e.g., DICE measurement as performed in Ref. [3]), and subjective assessment of benefits (e.g., assessment of an expectation of clinical time saving as performed in Ref. [4]).

One of the purposes of autocontouring is to save time; therefore, the most straightforward way to measure this effect is to investigate the amount of time taken by editing autocontours to a clinically useable standard compared to performing fully manual contouring, for example, as assessed in Ref. [2]. However, the different actions of initial contouring and of editing may entail the use of different tools,[5] making a specific assessment of the impact of autocontouring more challenging. Similarly, the use of different editing tools makes it difficult to compare across studies.[6] Recontouring the same cases may also introduce bias since familiarity with a case may reduce editing time. Furthermore, manually drawn contours may also be considered to require editing when reviewed by other experts.[7] Finally, the process of extra repeated manual contouring of multiple cases solely for the assessment of time taken is time consuming and not feasible for large patient cohorts in busy radiotherapy departments. This reduces the numbers of cases that can be considered in any study.

An alternative approach is to assess the impact contour errors have on the resulting treatment plan.[8] Comparisons of plans, created with autocontours or manual clinical contours, indicate that the position of contours affects the treatment plan created, and could therefore impact clinical outcomes. Such assessment does not reflect a real clinical workflow whereby the autocontours will be reviewed and edited as necessary; rather it highlights and quantifies the impact that errors could have on dose if erroneous contours are not corrected. Furthermore, it must be recognized that dosimetric differences would also result from variability in manual clinical contours.[9]

The most common approach to evaluation is comparison with "ground truth" clinical contours using quantitative measures. A wide range of quantitative measures for assessing autocontours against ground truth have been proposed, based on position, distance, volume overlap.[10] However, the clinical utility of such surrogates has been questioned[9,11] as the ground truth itself is subjective, for example, even quantitative measures between two sets of manual contours will not give a perfect score — reflecting inter- and intraobserver variability. Furthermore, good scores can result from a range of differences, some of which would fall within human subjectivity and some of which are more clearly wrong. Therefore, although a bad score would indicate a bad contour, a good score does not necessarily imply good contouring and a perfect score is not achievable as a result of the variability affecting how the contours are judged. Quantitative scores are also be affected by the geometrical properties of the organ being delineated, for example, the sensitivity of Dice similarity coefficient (DSC) to contouring errors depends on the size and shape of the segmented structure.[12]

Subjective qualitative assessment,[4,11,13] whereby the results of contouring, whether manual or automatic, are graded by clinicians, also has drawbacks; Clinical contouring staff may admit that they will change their opinion of their own contouring and may edit their own work if they view it at a later point. Therefore, editing alone does not mean a contour is bad. Additionally, when human observers are aware that a contour was generated by a machine, this awareness itself can introduce bias during qualitative contour assessment, and can influence their decision to edit or accept a contour.

This work is, therefore, motivated by the continuing need to identify a measure that represents a good surrogate for contouring quality, as well as being directly relevant to the contour-related tasks in the clinical workflow. It is also desirable that the surrogate measure be readily calculated, requiring less effort than is required by a full assessment of impact of workflow efficiency, and that it can be obtained easily across multiple institutions. In this study, we propose and evaluate a framework based on a variation on the Turing test for assessing contour quality. This assessment is built on the assumption that if a clinical observer is not able to distinguish between autocontours and those produced manually by an expert, then they are likely to edit the autocontours less as they will consider them as adequate for clinical use as those of the expert.

2. MATERIALS AND METHODS

2.A. The imitation game

The Imitation Game, sometimes referred to as the Turing test, was a proposal to test if "machines can think", which is reframed as "can machines display intelligence via imitation".[14] While the original proposal is quite complex, a common formulation of the game requires an interrogator to communicate electronically with a single subject and to judge whether that subject is a human or a machine, for example, The Leobner Prize.[15] 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 discussion regarding whether the Turing test is sufficiently complete as a demonstration of intelligence,[16] indistinguishability from human behavior can itself be regarded as a performance criterion.[17,18]

This question of imitation is applied in the proposed framework for assessing autocontouring, where we can test whether autocontouring is sufficiently similar to manual contouring by an expert to be indistinguishable when judged by a blinded interrogator. In this case, autocontouring might be deemed to be acceptable, or at least of the same quality as the human standard.

To investigate if the "Imitation Game" approach is a useful method of assessment of the clinical utility of contours, it was compared to commonly used quantitative measures of contouring quality as a surrogate of the editing time required to adapt autocontours to a clinically useable quality.

2.B. Imaging data and contours

Twenty stage I–III NSCLC patients were selected from routine clinical practice and their CT scans were used to delineate OARs. The following OARs were delineated: left lung, right lung, heart, mediastinum envelope, spinal cord, esophagus. Each OAR was delineated manually by a senior radiotherapy technician (Observer 1) with 10+ yr of experience, specialized in the thorax region using institutional guidelines, and independently by a further two also with 5+ yr of experience (Observer 2 and 3), using the standard contouring tools (Eclipse, version 11.0, Varian, Palo Alto, USA) available within the institution. The clinical contouring workflow was predominantly 2D contouring, as is common practice in radiotherapy, with the addition of some semiautomatic 3D threshold-based tools available within Eclipse.

A commercial atlas-based autocontouring product (Mirada RTx 1.6 and Workflow Box 1.4, Mirada Medical Ltd., Oxford, UK) was used to automatically generate contours for the 20 test patients, using 10 atlases.[4,19] These atlases consisted of stage I NSCLC patients with minimal geometric distortions and small lesion volumes, collected from the same institution and contoured by the same senior radiotherapy technician (Observer 1), and carefully inspected by a radiation oncologist for correctness.

The resulting autocontours were then manually edited by each technician to match clinical contouring guidelines. The time for the existing clinical contouring workflow (which may include semi-automatic contouring tools within Eclipse) and for editing autocontours was recorded for Observer 1, as previously reported in Ref. [20].

2.C. Quantitative contour analysis

For comparison, the initial contours of Observer 1 were treated as the “ground truth”. The following quantitative metrics, Dice similarity coefficient (DSC), average distance (AD), and 95% Hausdorff distance (HD), were computed between all of the other sets of contours, whether manual, autocontours, or edited autocontours, and this ground truth.

2.D. Implementation of the contouring imitation game

A website was setup, as shown in Fig. 1, in which reviewers were shown a single CT slice at a time, together with a contour corresponding to an organ, with the CT case, organ, slice, and method of contour creation being chosen at random. For each slice/contour reviewed, the user was asked “How was this contour drawn?” and given the options to select “By a human” or “By a computer”. The image display could be adjusted to standard window/levels and a magnifying tool was provided to enable inspection of the image/contour in detail. Performing assessment for a single slice at a time enables the estimation of the proportion of slices in which autocontouring is deemed indistinguishable.

For each case, the autocontours to be assessed and the original manual contours from one of the technicians (excluding Observer 1’s “ground truth” contours) were available for assessment. Only slices where contours from both methods of creation were available were considered in this study.

The website was setup to facilitate multiuser assessment of the contours, allowing the reviewers to perform assessment at a time convenient to them. Results were recorded in a database for later analysis. In addition to documenting the details of each slice presented and the reviewer’s choice for each question, the time taken to make the assessment was also noted. To allow unpressured assessment, observers were not aware that the time taken was being recorded.

Four radiation oncologists and four radiotherapy technicians, all specialized in thoracic radiotherapy, participated in the imitation game. Of these, two of the radiotherapy technicians had participated in the original contouring. Each reviewer assessed 50 randomly selected slice/contour combinations, from the pool of approximately 16,020 slices/contours. These figures and the random nature of the selection meant that it was rare for the same slice/contour to be presented for assessment more than once (4 out of 400 assessments).

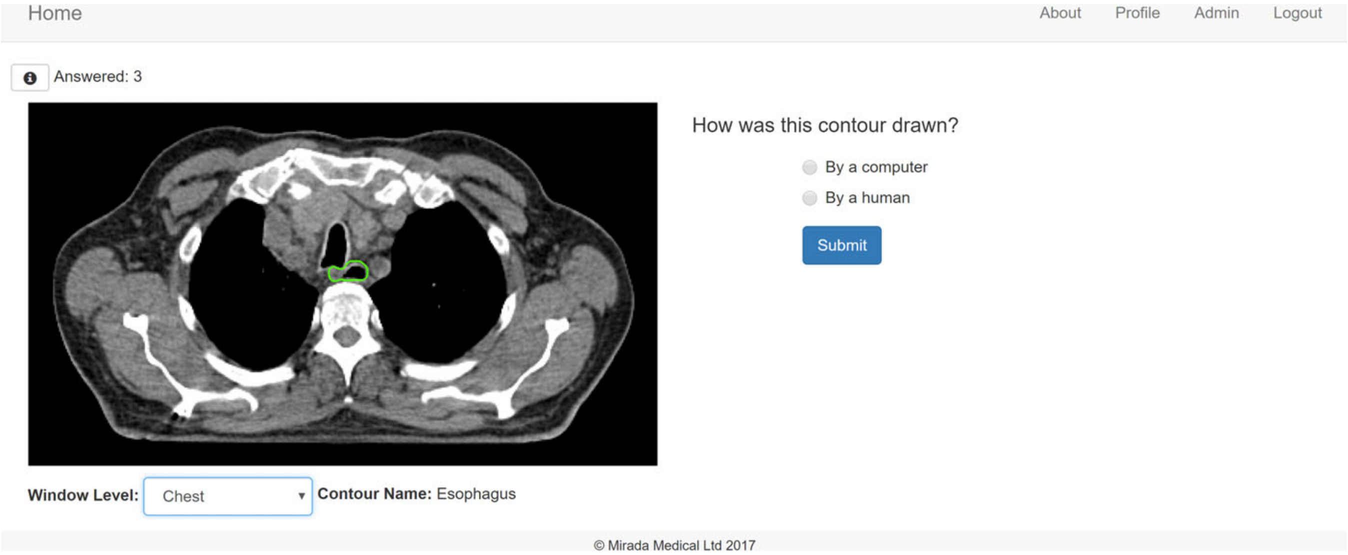
3. RESULTS

3.A. Quantitative contour analysis

The results of the quantitative assessment using DSC, AD and 95% HD are shown in Tables I, II and III respectively. To evaluate differences in quantitative measures, a ranked Wilcoxon test was performed. Statistical significance was assumed for P-values lower than 0.05. Bonferroni correction was used to correct for multiple tests. Further results are included in the Supporting Information.

For both lungs, observers 2 and 3 had a statistically significant greater agreement with the ground truth contouring than the atlas-based autocontouring, with higher DSC scores and lower distance measures as indicated in Tables I–III. This is illustrated in Figs. 2 and 3, which shows box-plots for the DSC and AD, respectively, for the right lung for each method. After editing, the interobserver error was similar, but showed greater conformity to the atlas, with no significant difference being found between observers for any of the metrics when comparing the contours to the edited autocontours of Observer 1. Although the difference in DSC appears small (approximately 0.02), this is a result of the large organ size. This difference corresponds to a more substantial difference in the AD (approximately 0.7 mm) and the 95%HD (approximately 2.3 mm).

In contrast, the heart shows more similar quantitative between observers and between the atlas and GT. Greater conformity was observed after editing of autocontouring with higher DSC scores and lower distances between the autocontours edited by Observer 1 and both the unedited autocontours and the those edited by the other observers. This is shown in Fig. 4, which shows box-plots of DSC for the heart, and Fig. 5, which shows the box-plots of AD. A similar finding was also observed for the mediastinum envelope.



Answered: 3

How was this contour drawn?

○ By a computer

○ By a human

Window Level: Chest

Contour Name: Esophagus

© Mirada Medical Ltd 2017

FIG. 1. Screenshot of the user interface used for the Imitation Game test. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE I. DSC scores for secondary observers and autocontouring, before and after editing of autocontouring.

Reference contour	Observer 1			AC editing — Obs 1 AC		
Comparison contour	Observer 2	Observer 3	Autocontouring	AC edited — Obs 2	AC edited — Obs 3	Autocontouring
Right lung	0.998 ± 0.001^{3}	0.999 ± 0.001^{3}	0.981 ± 0.006^{1,2}	0.998 ± 0.003	0.999 ± 0.002	0.998 ± 0.004
Left lung	0.999 ± 0.000^{3}	0.999 ± 0.001^{3}	0.957 ± 0.086^{1,2}	0.994 ± 0.017	0.992 ± 0.026	0.977 ± 0.078
Heart	0.871 ± 0.040	0.893 ± 0.029	0.898 ± 0.037	0.946 ± 0.032	0.92 ± 0.036	0.941 ± 0.039
Mediastinum envelope	0.919 ± 0.017	0.927 ± 0.017	0.926 ± 0.030	0.964 ± 0.022	0.955 ± 0.026	0.959 ± 0.027
Esophagus	0.755 ± 0.036^{3}	0.752 ± 0.058^{3}	0.462 ± 0.151^{1,2}	0.766 ± 0.056^{6}	0.761 ± 0.049	0.484 ± 0.196^{4}
Spinal cord	0.784 ± 0.038^{3}	0.783 ± 0.058^{3}	0.854 ± 0.027^{1,2}	0.934 ± 0.044	0.930 ± 0.036	0.928 ± 0.039

"Observer . . ." indicates original manual contours by the numbered observer. "Autocontouring" denotes unedited autocontours. "AC Edited — Obs . . ." indicates results for autocontours edited to a clinically acceptable standard by the numbered observer. The numbered superscript (from 1 to 6) indicates column of data (excluding the organ label column) to which statistical significant difference was found, that is, a super script 2 would indicate statistical difference to the results of the DSC comparison of the contours of Observer 1 (as reference contour) and Observer 3 (as comparison contour).

TABLE II. AD scores for secondary observers and autocontouring before and after editing of autocontouring.

Reference contour	Observer 1			AC editing — Obs 1 AC		
Comparison contour	Observer 2	Observer 3	Autocontouring	AC edited — Obs 2	AC edited — Obs 3	Autocontouring
Right lung	0.0652 ± 0.0375^{3}	0.0615 ± 0.0401^{3}	0.765 ± 0.191^{1,2}	0.0919 ± 0.104	0.0695 ± 0.0778	0.0728 ± 0.119
Left lung	0.0385 ± 0.0141^{3}	0.143 ± 0.435	2.12 ± 5.57^{1}	0.199 ± 0.359	0.273 ± 0.726	1.28 ± 4.79
Heart	3.88 ± 1.37	3.35 ± 0.794	3.32 ± 1.42	1.76 ± 1.1	2.49 ± 1.12	2.04 ± 1.38
Mediastinum envelope	2.77 ± 0.720	2.63 ± 1.25	2.70 ± 1.36	1.40 ± 0.946	1.77 ± 1.01	1.62 ± 1.22
Esophagus	1.53 ± 0.457^{3}	1.57 ± 0.418^{3}	4.59 ± 3.01^{1,2}	1.43 ± 0.335^{6}	1.53 ± 0.417^{6}	4.79 ± 3.57^{4,5}
Spinal cord	6.05 ± 2.49^{3}	5.77 ± 3.80^{3}	1.28 ± 0.405^{1,2}	0.800 ± 0.975	0.845 ± 0.977	0.757 ± 0.913

"Observer . . ." indicates original manual contours by the numbered observer. "Autocontouring" denotes unedited autocontours. "AC Edited — Obs . . ." indicates results for autocontours edited to a clinically acceptable standard by the numbered observer. The numbered superscript (from 1 to 6) indicates column of data (excluding the organ label column) to which statistical significant difference was found, that is, a super script 2 would indicate statistical difference to the results of the DSC comparison of the contours of Observer 1 (as reference contour) and Observer 3 (as comparison contour).

The elongated structures gave contrasting results. For the esophagus, the interobserver measurements showed significantly better agreement (higher DSC and lower distance measures as shown in Tables I–III) than was observed with the atlas, both when creating a new contour and when editing the results of autocontouring. However, for the spinal cord, significantly better agreement was observed between the ground truth and the atlas than between the human observers based on the initial manual contouring. Again, editing of autocontours improved conformity among the observers. Visual

TABLE III. 95% HD for secondary observers and autocontouring before and after editing of autocontouring.

Reference contour	Observer 1			AC editing — Obs 1 AC		
Comparison contour	Observer 2	Observer 3	Autocontouring	AC edited — Obs 2	AC edited — Obs 3	Autocontouring
Right lung	0.138 ± 0.112^{3}	0.134 ± 0.113^{3}	2.41 ± 1.02^{2,3}	0.405 ± 0.691	0.251 ± 0.492	0.337 ± 0.796
Left lung	0.0912 ± 0.00825^{3}	0.0943 ± 0.00886^{3}	9.35 ± 25.2^{2,3}	0.861 ± 1.28	1.28 ± 3.61	6.30 ± 23.7
Heart	11.5 ± 3.53	12.8 ± 3.73	10.2 ± 4.51	7.14 ± 3.72	11.4 ± 4.75	8.31 ± 4.43
Mediastinum envelope	9.97 ± 2.79	8.87 ± 4.36	9.50 ± 5.09	6.42 ± 3.43	8.10 ± 5.09	7.84 ± 4.90
Esophagus	5.38 ± 3.64^{3}	4.87 ± 2.41^{3}	16.4 ± 11.9^{1,2}	4.67 ± 1.39^{6}	5.17 ± 1.84^{6}	17.3 ± 14.5^{4,5}
Spinal cord	48.3 ± 17.8^{3}	43.7 ± 24.2^{3}	5.00 ± 3.49^{1,2}	3.47 ± 4.75	4.06 ± 4.59	2.92 ± 1.15

"Observer . . ." indicates original manual contours by the numbered observer. "Autocontouring" denotes unedited autocontours. "AC Edited — Obs . . ." indicates results for autocontours edited to a clinically acceptable standard by the numbered observer. The numbered superscript (from 1 to 6) indicates column of data (excluding the organ label column) to which statistical significant difference was found, that is, a super script 2 would indicate statistical difference to the results of the DSC comparison of the contours of Observer 1 (as reference contour) and Observer 3 (as comparison contour).

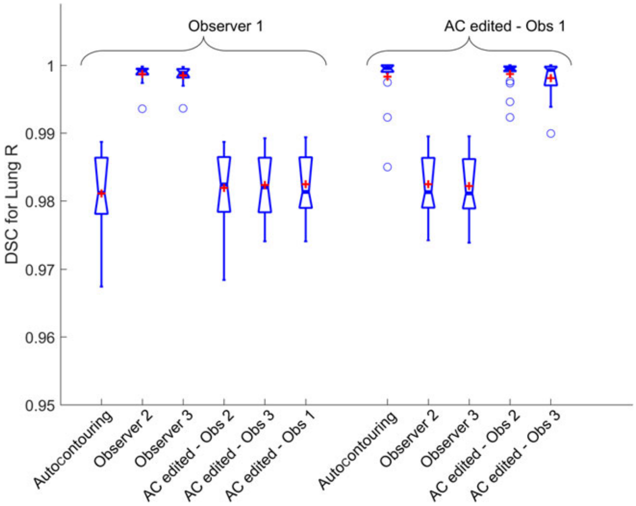


FIG. 2. DSC values for the right lung for the various resultant contours. The overhead braces indicate the reference against which the DSC was calculated. "Autocontouring" denotes the values for unedited autocontours. "AC edited — . . ." indicate the measurements after editing of the autocontours to a clinically acceptable standard by the numbered observer. "Observer . . ." are the values for the manual contours of the numbered observer. The crosses represent mean values. Circles indicate outliers. [Color figure can be viewed at wileyonlinelibrary.com]

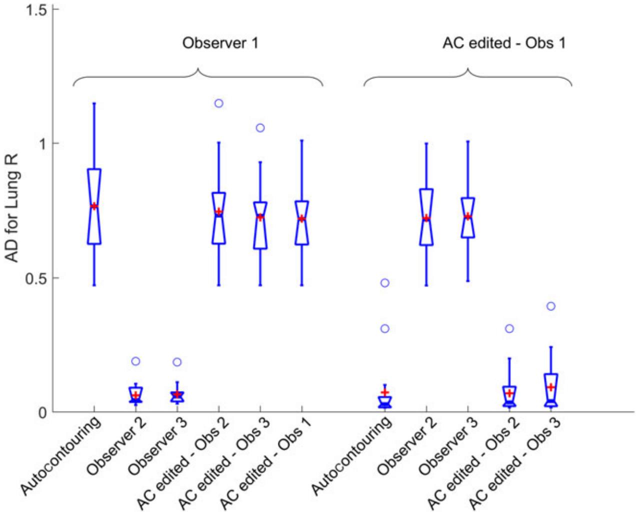


FIG. 3. AD for the right lung for the various resultant contours. The overhead braces indicate the reference against which the AD was calculated. "Autocontouring" denotes the values for unedited autocontours. "AC edited — . . ." indicate the measurements after editing of the autocontours to a clinically acceptable standard by the numbered observer. "Observer . . ." are the values for the manual contours of the numbered observer. The crosses represent mean values. Circles indicate outliers. [Color figure can be viewed at wileyonlinelibrary.com]

inspection, as illustrated in Fig. 6, suggests that the main variation occurs in the extent of the cord contoured.

Full quantitative results are provided as Supporting Information in Data S1 (QuantitativeResults.xls) and are summarized in Data S2 (QuantitativeSummary.xls).

3.B. "Imitation game" assessment

The system recorded the responses given by each user during the assessment. Figure 7 shows the overall percentage of slices correctly and incorrectly identified as being drawn by a human or computer for each organ. The *mis*classification rates for each organ are: Esophagus 30.0%, heart 22.9%, left lung 51.2%, right lung 58.5%, mediastinum envelope 43.9%, and spinal cord 46.8%. Figure 7 also gives a more detailed breakdown by the method of creation of the contours, for example, human-created contours of the esophagus were misclassified in 29.3% of cases while autocontours of the esophagus were misclassified in 21.9% of cases. For the lungs, the misclassification rate is close to random selection (left, 51.2%; right, 58.5%) with a high proportion of the unedited atlas contours being misclassified as human (left, 36.7%; right, 50.0%). A similar finding is obtained for the mediastinum envelope and spinal cord. In contrast, the heart and esophagus are often correctly classified as being drawn by a human or computer. Table IV shows the misclassification rate by observer, together with the assessment time taken per slice. It is noted that the lowest misclassification was achieved by one of the radiotherapy technician's involved in the contouring time experiment.

Full raw results of the imitation game assessments are provided as Supporting Information in Data S3 (Imitation

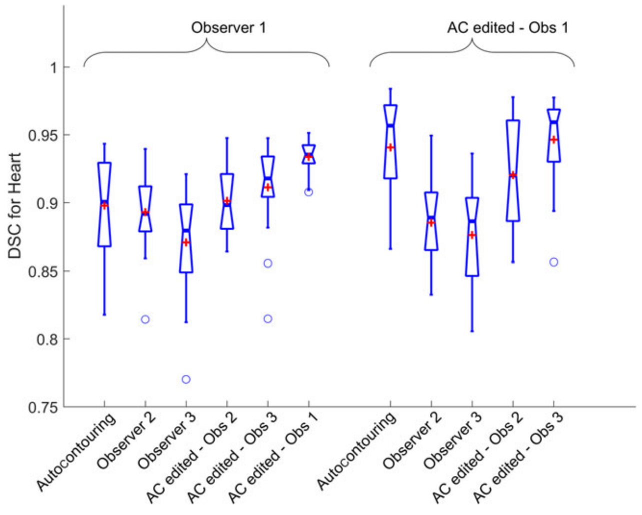


FIG. 4. DSC values for the Heart for the various resultant contours. The overhead braces indicate the reference against which the DSC was calculated. "Autocontouring" denotes the values for unedited autocontours. "AC edited — . . ." indicate the measurements after editing of the autocontours to a clinically acceptable standard by the numbered observer. "Observer . . ." are the values for the manual contours of the numbered observer. The crosses represent mean values. Circles indicate outliers. [Color figure can be viewed at wileyonlinelibrary.com]

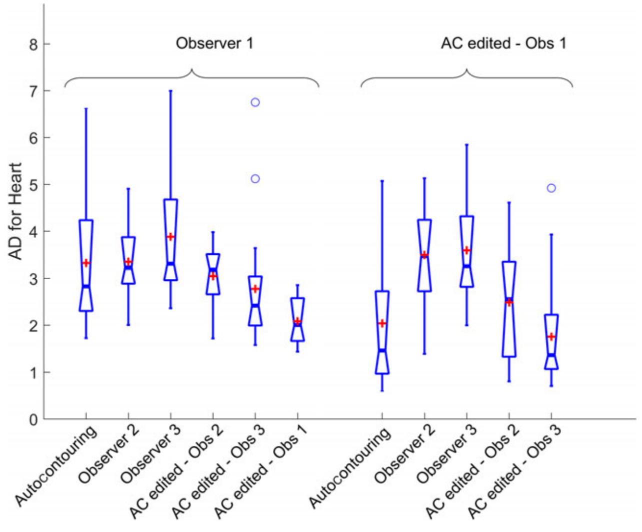


FIG. 5. AD values for the Heart for the various resultant contours. The overhead braces indicate the reference against which the AD was calculated. "Autocontouring" denotes the values for unedited autocontours. "AC edited — . . ." indicate the measurements after editing of the autocontours to a clinically acceptable standard by the numbered observer. "Observer . . ." are the values for the manual contours of the numbered observer. The crosses represent mean values. Circles indicate outliers. [Color figure can be viewed at wileyonlinelibrary.com]

GameResults.xls) and are summarized in Data S4 (Imitation GameSummary.xls).

3.C. Temporal assessments

Manual contouring, using the tools available at the institution, took on average 20.2 ± 2.3 min per case without autocontouring for the original contouring by Observer 1. Editing automatic contours took 12.6 ± 2.5 min per case by the same observer. The median times required for manual contouring and for editing automatic contours are shown for individual organs in Table V.

To evaluate differences in contouring time, a ranked Wilcoxon test was performed; *P*-values smaller than 0.05 were assumed to be statistically significant. Full analysis of the results of this time assessment have previously be reported in Ref. [20], but are reported here as the real measure of clinical impact for which approaches such as DSC or the misclassification rate seek to act as a surrogate.

The misclassification rate for an organ during the Imitation Game represents the inability to judge the source of a contour. A plot showing how time saving varies with respect to each of the surrogate measures (misclassification rate, DSC) is shown in Fig. 8 where the values are the median for each of the six organs. This gives some indication of how well the proportionate time saving might be predicted by each surrogate. It should be noted that the absolute editing time will vary according to the organ and observer.

For the Imitation Game, the mean time for each slice assessment was 18 s (range: 8–36 s), so that the full assessment took approximately 16 min per participant (range: 7 min 23 s to 31 min 32 s). Seven times (out of 800) that were over 2 min were excluded from the analysis, as it is assumed that the participant had left the website/assessment to attend to something else. No correlation was observed with the time taken to make the assessment and the level of accuracy of the assessment.

Full timing results are provided as Supporting Information in Data S5 (TimingResults.xls) and are summarized together with quantitative and imitation game results in Data S6 (SimpleSummary.xls).

4. DISCUSSION

The main purpose of this study was to evaluate whether the proposed Imitation Game method for assessing autocontouring is beneficial compared with standard assessment methods, rather than to actually assess the performance of any particular autocontouring method. As a consequence, more information was gathered about the performance of the atlas-based autocontouring system than might normally be considered in an evaluation.

4.A. Comparison to quantitative assessment

While anatomically there is a correct definition of an organ, the processes of contouring on CT are subject to both inter- and interobserver variability. That Observer 1 can edit the autocontours until they believe them to be clinically acceptable, and yet not achieve a DSC of 1 against their own original contours demonstrates this. Therefore, the notion of

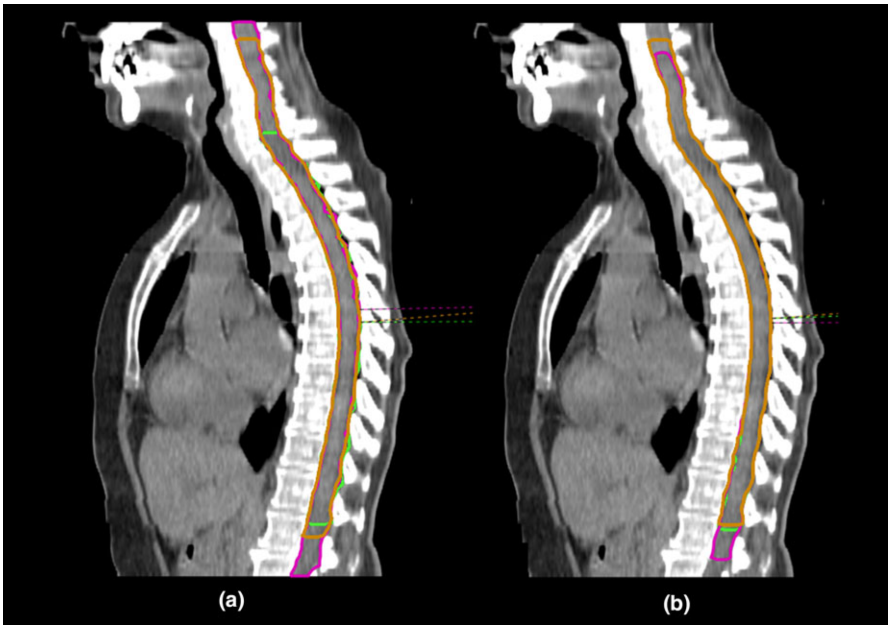


FIG. 6. Contouring of the spine manually (a) and after editing autocontouring (b). Contouring agreement is high between observers without autocontouring, except in the extent contoured. The agreement in the extent improves following autocontouring. [Color figure can be viewed at wileyonlinelibrary.com]

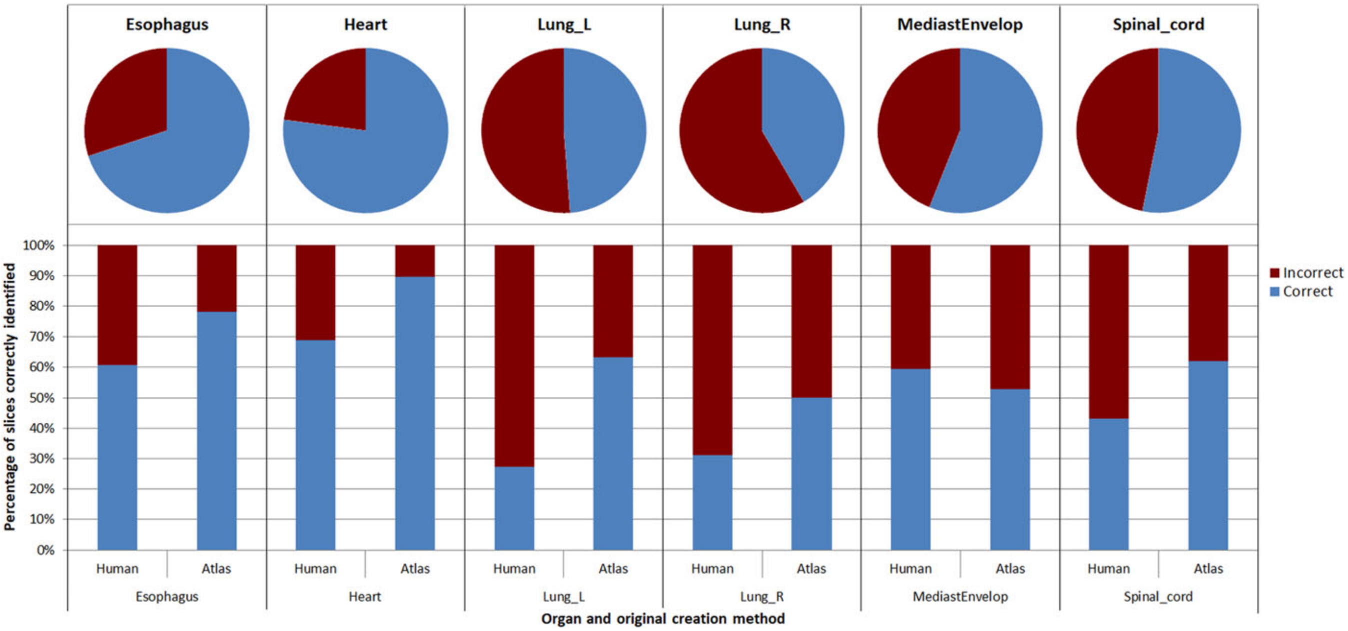


FIG. 7. Rates of correct and incorrect classification of contours as human or automatically generated during the Imitation Game test. Top: overall rates. Bottom: rates by contour origin. [Color figure can be viewed at wileyonlinelibrary.com]

a "ground truth" contour is flawed to some extent. Measures of autocontouring performance, such as DSC, average distance etc., against a manual ground truth, should consider interobserver variability in manual contouring as a standard against which variation can be measured, for example, see Ref. [7].

If DSC was the only quantitative measure considered in this study, then one might conclude that the performance of autocontouring for the lungs was poor in relation to interobserver variability and that the performance for the heart was acceptable being within human variation. However, the Imitation Game reveals the opposite to be the case; the autocontouring of the lung is difficult to distinguish from human contouring, while for the heart the participants could distinguish a difference more easily. This suggests that the lung performance would be deemed satisfactory while the heart

TABLE IV. Overall misclassification rate and mean ± SD assessment time per slice by observer.

Observer	Rad Onc 1	Rad Onc 2	Rad Onc 3	Rad Onc 4	RTT 1[a]	RTT 2	RTT 3	RTT 4[a]
Misclassification rate (%)	36	54	44	46	48	48	48	26
Assessment time per slice (s)	29.3 ± 22.3	18.2 ± 9.4	10.0 ±8.2	15.7 ±9.1	8.0 ±8.6	36.6 ±29.0	16.9 ±12.7	8.9 ±5.6

[a]Indicate observers involved in the original contouring.

TABLE V. Median times (and standard deviation) taken to produce contours manually and to edit automatically generated contours and the percentage of time saved.

	Lung L	Lung R[a]	Heart[a]	Spinal cord[a]	Esophagus	M_Envelope[a]
Manual contouring	01:29 (00:24)	01:55 (00:23)	03:50 (00:33)	02:20 (00:27)	02:51 (00:30)	07:36 (01:13)
Editing autocontours	00:51 (00:51)	00:27 (00:34)	02:49 (01:09)	01:10 (01:11)	02:31 (00:51)	04:09 (01:12)
Time saved (%)	43	77	25	50	12	46

[a]Denotes organs for which the time saving was statistically significant ($P < 0.05$).

performance would not. This is in accord with more general observations that quantitative measures, such as DSC, are insufficiently informative, for example, by failing to differentiate systematic from random errors.[1]

The quantitative measures show that the postediting lung contours conform more to the original autocontours suggesting that the autocontours for the lung contours are relatively acceptable. The statistically significant difference prior to editing can be attributed to low interobserver variability resulting from the use of semiautomatic contouring tools in the used delineation software which leads to different, yet similarly acceptable, contours. Therefore, while statistically significant, these differences are considered clinically insignificant in the subjective judgment of the observers performing the contouring. However, after editing the heart autocontours, the performance appears to be the same, suggesting that the differences observed in the Imitation Game are clinically significant, but do not end up appearing statistically different according to the quantitative measures.

4.B. Comparison to workflow assessment

The key test of autocontouring is in its impact on clinical workflow; in particular, it is important to consider whether the use of autocontouring ultimately leads to time saving. In this work, the relationship between the time saving and potential surrogate measures (the Imitation Game misclassification rate and DSC) was explored in Fig. 8. While the Imitation Game does not consider the workflow directly, the assumption is that if the participant cannot tell the difference between the autocontour and a human-drawn contour, then it must be sufficiently good to require minimal editing since the manual contours were drawn to the clinical standard. For Observer 1, a statistically significant ($P < 0.05$) time saving was found for editing the autocontours of the right lung, spine, and mediastinum envelope compared with the routine clinical workflow. A 43% time saving was found in the left lung, although this was not statistically significant as a consequence of a single outlier as report in Ref. [20]. In the Imitation Game, these organs were the ones for which the participants were least able to identify the source. However, for the heart and esophagus, the organs for which the autocontouring failed to show an ability to imitate human contouring, little or no time saving was found in practice. This correspondence suggests that the original assumption is correct, and the inability to accurately judge the source of a contour indicates a reduced need for editing and therefore a greater time saving overall.

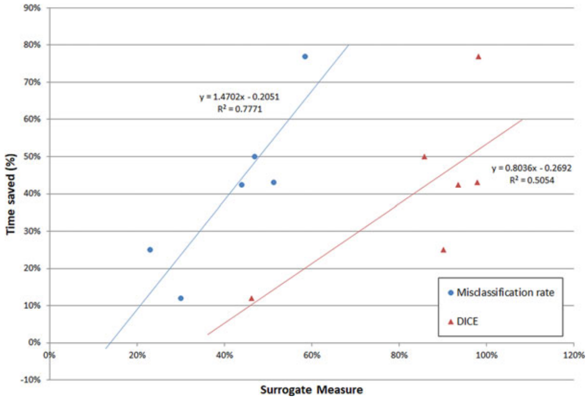


FIG. 8. Comparison of time saved when editing autocontours against two surrogate measures of contouring quality, misclassification rate during the IG test and DSC value. Each point represents the average values for an individual organ. [Color figure can be viewed at wileyonlinelibrary.com]

The data in Fig. 8 give an overview of how the surrogate measures of misclassification rate and DSC relate to time saving. The small number of data points prevents a rigorous statistical analysis but the data in Fig. 8 suggest that time saving is better predicted by Imitation Game misclassification as a surrogate than by DSC value. Absolute time taken by different observers will vary, and therefore the absolute relationship between time saving and misclassification is unlikely to be adequate to predict time saving for a particular autocontouring method. However, the relationship may indicate that an autocontouring system demonstrating greater misclassification is likely to save greater time when editing contouring. This relates to the notion of "task-" or "application-based" assessment of autocontours, that is, the measurement of performance in a way that more directly relates to how the contours will be used in a clinical workflow. Measures that more strongly predict time saving can be viewed as more valuable in this context.

On a separate point, a general bias that participants have for classifying a contour's source as a human (or machine) during the Imitation Game should also be investigated in future. This could then be used to generate significance levels for the observed classification rates, for example, through permutation/simulation tests.

4.C. Design decisions

In setting up the Imitation Game website, there were a number of design decisions with the potential to affect the user acceptance and the quality of results.

Perhaps the most important decision was to show participants single slices of single organs. While the majority of clinical users will be very familiar with viewing and editing contours in an axial orientation, on a slice-by-slice basis, all contouring software tools allow scrolling between slices, and many also offer orthogonal views. Allowing the users to scroll, or perform a 3D review of the contours, would enable a reviewer to fully assess a given set of contours and to judge their origin. While ultimately we would like to achieve autocontouring of a standard to pass such a test, this method of assessment has its drawbacks; it is more time consuming both to assess a whole case in 3D and also to get a sufficiently large number of data samples. A large number of assessments is required to enable quantitative assessment, therefore, the decision was made to perform single slice assessments. The proportion of correctly identified slices can be estimated by performing multiple random assessments. Assuming that editing is carried out on axial slices, this proportion can be expected to be related to the editing time, as was found to be the case.

A further consideration relates to organs, such as the spinal cord and esophagus, which span a large number of slices and where contours from both methods (manual and automatic) are not available for every slice. We decided in this survey to only show slices during the Imitation Game where both methods provided contours. This may introduce a small degree of bias if one method provides, for example, more slices than the other, but this was deemed preferable as the set of slices provided to the system was matched across both methods.

Furthermore the slice-by-slice assessment, combined with the random selection of organ and slice, meant that it was unlikely that there was any repeat assessment of contours in this experiment. Consequently, this study did not consider the consistency of assessment by observers (either intra- or interobserver). Such a study is an avenue for further investigation.

4.D. Alternative questions

Verbal feedback from participants indicated that they felt that the question was not the right question, since, in clinical practice, they are less concerned about the origin of contours than they are regarding the correctness of the contour. Ultimately, contours which are more precise and reliable than human contouring would be desirable, even if such contours are distinguishable from human contouring by their better quality.

This would suggest that it would be better to ask a question with the format:

You have been asked to review these contours for clinical use by a colleague. Would you:

- *Require them to be corrected; There are large, obvious, errors*
- *Require them to be corrected; There are minor errors that need a small amount of editing*
- *Accept them as they are; There are minor errors but these are clinically not significant*
- *Accept them as they are; The contours are very precise*

Such questions attempt to assess the clinical acceptance of contours. The question above has been carefully phrased to focus on QA for review purposes, which is already performed by some institutions, so as to allow some element of acceptable clinical variability. While similar questions on whether contours need editing have previously been asked, they present a challenge in that a clinical end user may often even wish to edit their own or their colleagues' contours.[19] Asking such a question in a blinded manner, as in the Imitation Game test, would allow assessment of acceptance level compared to clinical contouring and would mitigate any bias due to knowing the contouring source. A challenge with this method of assessment may also occur where multi-institutional participation is sought from practitioners with different contour styles. The reference clinical acceptability of manual contouring may be heavily affected by the institution of the participant,[4] and therefore some level of adjustment/interpretation would be required to understand the absolute performance. The Imitation Game avoids this problem as the user is not asked whether they like the contour, and therefore they are able to indicate that they think it was a human even if they think it is incorrect.

An alternative would be to ask an A/B comparison type question, whereby the user is shown images from two methods of contour creation as asked

Which contour (set of contours) do you think are best?

- *Set A*
- *Set B*

This question lacks the judgment of the clinical acceptability of the contouring source, but has the potential to mitigate bias introduced by institutional variation. If both contours of the direct comparison are intended to represent the same structure as contoured by a single institution, then the method of autocontouring should have a similar contouring style as the manual contouring. Thus, a participant may judge the better accuracy of the contour, even if they deem neither acceptable according to their institutional guideline. In the case where the autocontouring imitates the human-level performance, the decision should be random leading to a 50% preference toward either method. However, this approach would also enable the autocontouring to be preferred.

These alternatives will be investigated in the future research, for exmaple, to correlate the results with the results obtained in the Imitation Game. The evaluation website can be accessed at www.autocontouring.com, where these questions have been included.

5. CONCLUSIONS

A method of assessing autocontouring based on the Imitation Game, or Turing test, has been proposed. In this initial investigation, this approach was found to be able to discriminate between contouring methods for organs where standard quantitative approaches would indicate no difference. It was also observed that this approach is able to mitigate against differences that might occur as a result of normal contouring variation within the range of clinical acceptable contours, something that might be inadvertently highlighted by quantitative assessment. At the same time, the clinical burden of conducting such an experiment was found to be much lower than performing an assessment of time taken for contour correction. Finally, the initial results comparing the proposed measure with times saved suggest that it may act as a surrogate for a task-focused assessment of contouring quality. While we have focused in this work on a radiotherapy application the Imitation Game test may be applied to the assessment of contours in other areas, such as, for example, segmentation for radiological quantification,[21] for ROI detection[22] or for cell classification.[23]

ACKNOWLEDGMENTS

This work was supported by Eureka Eurostars (E!9297) via InnovateUK (600277).

CONFLICTS OF INTEREST

M.J. Gooding and P. Aljabar are current employees of Mirada Medical. A.J. Smith, M. Tariq, and D. Peressutti were employees of Mirada Medical at the time of writing.

[a] Author to whom correspondence should be addressed. Electronic mail: mark.gooding@mirada-medical.com.

REFERENCES

1. Valentini V, Boldrini L, Damian A, Muren L. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. *Radiother Oncol*. 2017;112:317.
2. Reed VK, Woodward WA, Zhang L, et al. Automatic segmentation of whole breast using atlas approach and deformable image registration. *Int J Radiat Oncol Biol Phys*. 2009;73:1493–1500.
3. Schreibmann E, Marcus DM, Fox T. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. *J Appl Clin Med Phys*. 2014;15:22–38.
4. Gooding MJ, Chu K, Conibear J, et al. Multicenter clinical assessment of DIR atlas-based autocontouring. *Int J Radiat Oncol Biol Phys*. 2013;87:S714–S715.
5. Sjöberg C, Lundmark M, Granberg C, Johansson S, Ahnesjö A, Montelius A. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. *Radiat Oncol*. 2013;8:229.
6. Delpon G, Escande A, Ruef T, et al. Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy. *Front Oncol*. 2016;6:178.
7. 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. *Int J Radiat Oncol Biol*. 2011;81:950–957.
8. Voet PW, Dirkx ML, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJ. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis *Radiother Oncol*. 2011;98:373–377.
9. Mattiucci GC, Placidi L, Boldrini L, et al. A dosimetric analysis of Dice index and Hausdorff distance in H&N: which index can evaluate autocontouring software? *Radiother Oncol*. 2014;111:S52.
10. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41:050902.
11. Gautam A, Weiss E, Williamson J, et al. Assessing the correlation between quantitative measures of contour variability and physician's qualitative measure for clinical usefulness of auto-segmentation in prostate cancer radiotherapy. *Med Phys*. 2013;40:90.
12. Rohlfing T, Brandt R, Menzel R, Maurer C. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*. 2004;2004:1428–1442.
13. Greenham S, Dean J, Fu CKK, et al. Evaluation of atlas-based auto-segmentation software in prostate cancer patients. *J Med Radiat Sci*. 2014;61:151–158.
14. Turing AM. Computing machinery and intelligence. *MIND*. 1950;59:433–460.
15. Shieber SM. Lessons from a restricted Turing test; 1994. arXiv preprint cmp-lg/9404002.
16. Gunderson K. The imitation game. *Mind*. 1964;73:234–245.
17. Harnad S. *The Turing Test is not a trick: Turing indistinguishability is a scientific criterion*. 1992;3:9–10.
18. Saygin AP, Cicekli I, Akman V. Turing test: 50 years later. *Mind Mach*. 2000;40:463–518.
19. Larrue A, Gujral D, Nutting C, Gooding MJ. The impact of the number of atlases on the performance of automatic multi-atlas contouring. *Phys Med: Eur J Med Phys*. 2015;31:e30.
20. Lustberg T, van Soest J, Gooding MJ, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol*. 2018;126:312–317.
21. Aguilar C, Edholm K, Simmons A, et al. To develop an algorithm to segment and obtain an estimate of total intracranial volume (tICV) from computed tomography (CT) images. *Eur Radiol*. 2015;25:3151–3160.
22. Tai S, Chen Z, Tsai WT. An automatic mass detection system in mammograms based on complex texture features. *IEEE J Biomed Health Inform*. 2014;2014:618–627.
23. Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review

—current status and future potential. *IEEE Rev Biomed Eng*. 2014;7:97–114.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Original quantitative results of pairwise comparison of contours per structure and per test case

Data S2. Summary of quantitative results of pairwise comparison of contours per structure

Data S3. Original results per question from the Imitation Game website

Data S4. Results for the Imitation Game summarized by participant, by organ. and overall

Data S5. Original timing results per structure and per test case

Data S6. Summary of timing, quantitative, and Imitation Game results by organ