Multi-centre evaluation of atlas-based and deep learning contouring using a modified Turing Test

M. Gooding[1], A. Smith[1], D. Peressutti[1], P. Aljabar[1], E. Evans[2], S. Gwynne[3], C. Hammer[4], H.J.M. Meijer[5], R. Speight[6], C. Welgemoed[7], T. Lustberg[8], J. van Soest[8], A. Dekker[8], W. van Elmpt[8]
[1]Mirada Medical Limited, Oxford, United Kingdom.[2]Velindre Cancer Centre, Clinical Oncology, Cardiff, United Kingdom.[3]South West Wales Cancer Centre, Clinical Oncology, Swansea, United Kingdom.[4]University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands. [5]Radboud University Medical Center, Department of Radiation Oncology, Nijmegen, The Netherlands. [6]Leeds Cancer Centre, Leeds, United Kingdom. [7]Imperial College Healthcare NHS Trust, Radiotherapy Department, London, United Kingdom.[8]MAASTRO Clinic, Department of Radiation Oncology, Maastricht, The Netherlands.

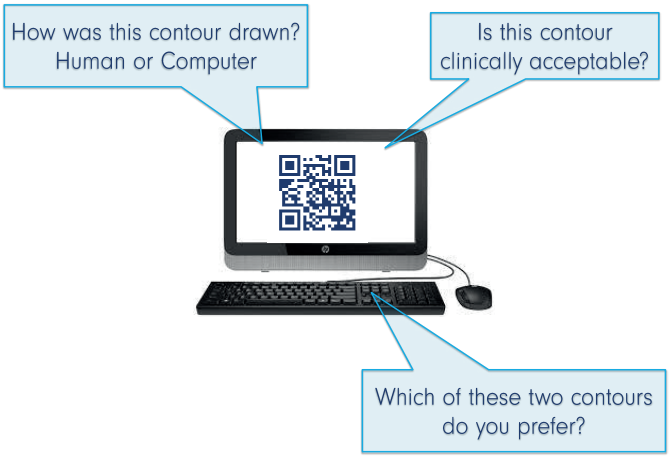
Objective

While quantitative assessment of autocontouring quality is useful, frequently used measures do not necessarily indicate clinical acceptability or benefit. In contrast, clinical based assessment metrics, such as time saved with autocontouring or subjective evaluations, are both time consuming to perform and difficult to implement in a multi-centre evaluation.

Inspiration is taken from the Artificial Intelligence community to propose an assessment method based on the "Turing Test". The objective of this study was to perform a multi-centre evaluation of two autocontouring methods using this approach.

Materials and Methods

A website was set up to facilitate multi-centre comparison, showing images and contours in a blinded fashion.



Use the QR-code to try it for yourself at
http://www.autocontouring.com

60 Clinical test cases from a single institution (40 thoracic, 20 prostate)

3 Types of contour
- Existing clinical
- Multi-atlas contouring (Mirada Workflow Box 1.4)
- Deep Learning Contouring (DLC) (Mirada DLCExpert™ prototype)

15 Clinical participants (clinicians, dosimetrist or technicians)

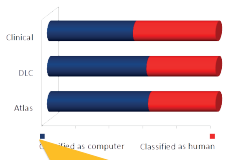
5 Institutions participated

100 Randomly chosen questions answer by each participant

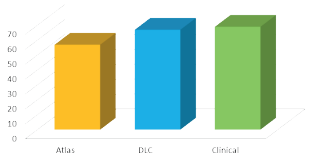
Results

Thoracic segmentation

Classification of contours



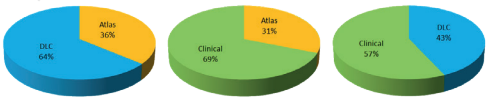
Acceptance of contours (%)



DLC acceptance rate close that of clinical contouring. Note: Only 70% of clinical contours were accepted

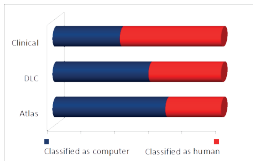
The source of the contours is often misidentified by the blinded participants

Contour preference in blind side-by-side comparison

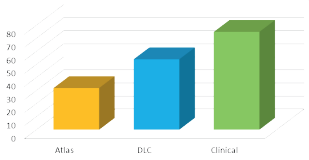


Prostate segmentation

Classification of contours

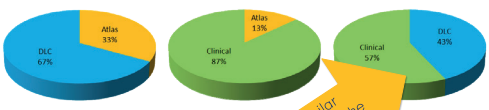


Acceptance of contours (%)



DLC shows increased clinical acceptance compared to atlas contouring

Contour preference in blind side-by-side comparison



DLC has similar preference rate to the clinical contouring

Conclusions

The Turing Test style assessment method provided an easy way to perform web-based multi-centre validation of autocontouring.

This study found that autocontours may be confused with clinical ones, when reviewed blindly. DLC showed increased clinical acceptance for prostate OAR contouring compared to atlas contouring. For thoracic imaging, DLC contours were accepted at a similar rate to clinical ones.

@SciChief
mark.gooding@mirada-medical.com