Qualitative assessment of oesophageal cancer metabolic tumour volumes delineated by artificial intelligence

Author names

Dr Craig Parkinson[a]

Mr Shailen Sobhee[b]

Mr Walter Riviera[b]

Mr Salvatore Berenato[c]

Mr Costas Stylianou[b]

Prof. Tom Crosby[c]

Dr Kieran Foley[c]

Dr Emiliano Spezi[a]

Affiliation

[a]School of Engineering, Cardiff University

[b]Intel Corporation

[c]Velindre NHS Trust

Corresponding Author

Dr Craig Parkinson

Background

Incidence of oesophageal cancer is rising. Radiotherapy is increasingly used to treat this poor prognosis disease but requires significant resources to plan treatment. Therefore, automated methods would be preferred. Quantitative analysis of artificial intelligence (AI) algorithms is often reported, but qualitative evaluation is lacking. We investigated observers ability to differentiate manual versus a fully automated AI algorithm for outlining metabolic tumour volume (MTV) using a Turing test, including inter and intra-observer variability.

Method

Five radiologists (Ob1 to Ob5) independently observed 580 contours. 256 contours were delineated using a U-Net deep learning (DL) model and 324 were delineated manually. Observers decided whether the contour had been created with a DL method, manually, or if they were unable to tell. Of the 580 contours, 37 contours were repeated twice. Observers were blinded to the method and presented with a co-registered PET/CT, with a contour overlay. CT imaging was windowed to a window width of 330 Hounsfield units (HU) and window centre of -10 HU.

Results

Overall, Ob1 to Ob5 correctly identified 165 (28.4%), 199 (51.6%), 190 (32.8%), 181 (31.2%) and 193 (33.3%) out of 580 cases, respectively. Ob1 to Ob5 identified 202 (78.9%), 199 (77.7%), 159 (62.1%), 189 (73.8%) and 143 (55.9%) of 256 DL contours as being manually delineated. In repeat imaging, Ob1 changed opinion in 9 cases, Ob2 10 cases, Ob3 10 cases, Ob4 7 and Ob5 8 cases. On average observers changed opinion in 9 cases (21.6%) with a minimum of 7 (18.9%) cases and a maximum of 10 cases (27.0%). Observers on average identified 178.4 (69.6%) of the DL contours as being delineated manually (range; minimum 143 cases (55.8%) and maximum of 202 (78.9%) cases).

Conclusion

We have shown that Turing tests provide an additional method for qualitative evaluation that complements quantitative metrics, to assess AI algorithm performance in outlining metabolic tumour volumes. In our study, observers were unable to confidently determine the delineation method suggesting a strong performance of the AI algorithm. However, observer selection is subject to inter and intra-observer variability and potentially impacted by clinical experience.

Impact statement

Qualitative assessment of AI-based delineation algorithms is vitally important for clinical acceptance.