An evaluation of AI auto-segmentation for Head & Neck cancer

A thesis submitted to the University of Manchester for the degree of Doctor of Clinical Science in the Faculty of Biology, Medicine and Health

2022

Simon W P Temple
School of Medical Sciences

Contents

Word Count: 23,591

List of Figures

List of Tables

Abstract

Purpose

To investigate performance of multiple commercial AI auto-segmentation systems for head and neck (H&N) radiotherapy treatment planning, to inform on associated quality assurance (QA) requirements, and to investigate patient views on the use of such technology in the planning of their own radiotherapy treatment.

Methods

Four commercial AI auto-segmentation systems were used to generate contours for five commonly used H&N organs at risk (OAR) using 50 H&N patient datasets. Resulting contours were compared to gold standard contours using multiple similarity metrics. One commercial system was used to generate four common H&N OARs on 500 patient datasets. Auto-segmented OARs were compared to manually-created clinical contours using Dice Similarity Coefficient (DSC) and failure rates were identified using previously calculated expected DSC values. An existing standardised patient questionnaire was distributed to cancer patients who were receiving radiotherapy at the Clatterbridge Cancer Centre between November 2021 and March 2022.

Results

Overall performance differences between commercial systems were found to be statistically insignificant for all comparison metrics. For the 500 patient study, true failure rates for the four OARs investigated were 0.4% for brainstem, 2.2% for mandible, 1.4% for left parotid and 0.8% for right parotid. The patient questionnaire results showed that there was a moderately negative patient view towards the use of AI in radiotherapy.

Conclusions

Comparable levels of performance were observed between all systems. This indicates that AI-based auto-segmentation products are developing at a similar pace in terms of the quality of contours produced. The true failure rate for AI auto-segmentation systems in the H&N region for the OARs investigated is extremely low and it is therefore advised that QA of resulting auto-segmented OARs should utilise automated methods. There are clear patient concerns around the use of AI in radiotherapy and therefore both staff and patient education is required.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other University or other institute of learning.

Copyright Statement

Dedication and Acknowledgements

I dedicate this thesis to my wife, Ronette. I could not have asked for a more loving, considerate and understanding wife during my time working on the DClinSci, and your ongoing support has been invaluable throughout. Thank you for everything.

I would also like to thank and acknowledge the significant support I received from my local supervisor, Carl Rowbottom, who was always willing to assist and advise, often at extremely short notice.

I would like to thank my external supervisor, Rob Chuter, who offered advice whenever needed, and I would like to thank Janette Simpson and Jennifer Cadwallader for their assistance distributing questionnaires to patients.

Finally, I would like to thank my colleagues from the physics department at the Clatterbridge Cancer Centre, who have enabled me to carry out this research and have supported me throughout.

The Author (statement for examiners)

The author is a state registered clinical scientist (CS15210) and a Medical Physics Expert (297) who has been working as a Consultant Clinical Scientist at the Clatterbridge Cancer Centre NHS Foundation Trust since February 2020.

Academic qualifications:

BSc (Hons) in Physics, University of Sheffield, 1993

MSc in Medical Physics, University of Leeds, 2003

Previous research experience includes, but is not limited to:

MSc Dissertation Title: *Validation and Optimisation of Leaf-Setting Strategies for Conformal Radiotherapy.*

Other publications as a named author:

Adams, E. et al. (2020). PD-0193: Validation of a multi-centre knowledge-based planning model for radiotherapy of cervical cancer. Radiotherapy and Oncology, 152, pp.S96–S97.

Ng, W.L. et al. (2015). Volumetric modulated arc therapy in prostate cancer patients with metallic hip prostheses in a UK centre. Reports of Practical Oncology and Radiotherapy, 20(4), pp.273–277. [online]. Available from: http://dx.doi.org/10.1016/j.rpor.2015.03.006.

Fenwick, J.D. et al. (2004). Geometric leaf placement strategies. Physics in Medicine and Biology, 49(8), pp.1505–1519.

Conference attendance:

In May 2022 the author attended the European Society for Therapeutic Radiology and Oncology (ESTRO) annual congress meeting in Copenhagen and presented work based on the research carried out in study A of this thesis in the form of a poster discussion (Appendix D).

DClinSci:

This thesis forms part of the Doctor of Clinical Science (DClinSci) along with a number of taught components which are summarised in Appendix A.

The taught modules were provided by the University of Liverpool, the University of Manchester and the Alliance Manchester Business School (Postgraduate Diploma in Leadership and Professionalism for Healthcare Sciences).

Submission Format

This thesis is presented in journal format in accordance with the University of Manchester requirements.

The decision to use journal format followed discussions with supervisors where it was agreed that this research topic was well suited to be presented in this format. Use of the format also aligns well with HSST research learning objectives.

According to journal format guidelines it comprises three main sections:

The first section provides an introduction to the use of Artificial Intelligence based software in healthcare and specifically radiotherapy, a review of the existing literature and a discussion of gaps in current knowledge.

The second section comprises three research studies presented in a format suitable for publication. Studies are linked by the theme of the use of AI auto-segmentation software in radiotherapy treatment planning.

The third section presents a critical analysis of the work undertaken in the three studies.

1. Introduction

In the modern-day National Health Service (NHS) the importance of Quality Improvement (QI) in order to improve the efficiency and safety of clinical processes cannot be overstated. Any research which leads to the implementation of such improvements is therefore of great value.

This principle is equally applicable to the field of radiotherapy, and Towards Safer Radiotherapy (The Royal College of Radiologists, 2008) recommends that 'changes should be introduced wherever and whenever appropriate to improve the effectiveness and efficiency of the radiotherapy department'.

In the area of radiotherapy treatment planning there is significant scope to further improve efficiency through automation of the complex routine tasks that form part of the treatment planning process. These tasks can consume a large amount of time for a number of different health professionals, including clinical oncologists, physicists and radiographers.

Worldwide, an estimated 562,328 people were diagnosed with head and neck (H&N) cancer in 2020, and in the same year an estimated 277,597 people worldwide died from the disease (Cancer.Net, 2022). In the UK there are over 4000 deaths from H&N cancer each year and (Cancer Research UK, 2022).

Radiotherapy is a common form of treatment for H&N cancer, with between 43% and 85% of patients receiving this treatment as part of their primary cancer treatment (Cancer Research UK, 2022).

Radiotherapy for the H&N anatomical site can be particularly time-consuming in relation to treatment planning due to the complex target volumes and abundance of nearby organs at risk (OAR) (van der Veen, 2017), and the introduction of any additional automation for this site is therefore likely to be particularly beneficial in terms of both efficiency and safety.

Figure 1 illustrates a typical H&N radiotherapy planning pathway. For each patient the entire process can take several days, and each individual task can occupy a significant amount of staff time.

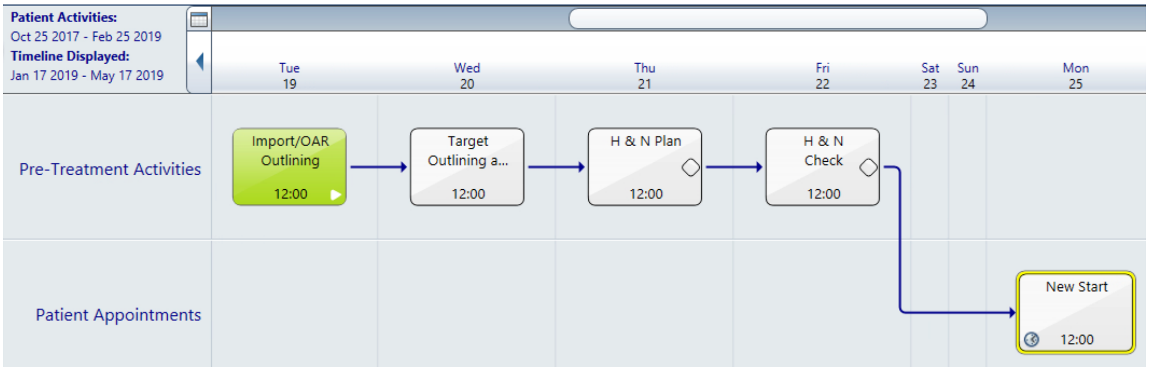
Patient Activities:
Oct 25 2017 - Feb 25 2019
Timeline Displayed:
Jan 17 2019 - May 17 2019

Fig. 1. The head and neck radiotherapy planning pathway

An important stage of the process involves comprehensive and accurate delineation of nearby OARs onto CT scans, and this can be especially time consuming due to the manual nature of the methods involved (Fritscher et al., 2014).

Automation of any stage of the planning process can produce efficiency savings, and methods of automation of the outlining process already exist, with commercial solutions having been in clinical use for a number of years (Daisne et al., 2013).

In recent years the use of Artificial Intelligence (AI) based software in healthcare has increased rapidly, and this thesis investigates a relatively new method of automatic segmentation of contours using such AI-based software.

In addition to the process of generating contours, methodologies to validate the quality of resulting contours have been investigated. Further to this, views of radiotherapy patients on the use of AI-based software for their own treatment planning have been collected.

The first section of the thesis reviews the associated existing literature. The review is presented in three distinct sections, covering the following three subject areas:

- Current state of AI auto-segmentation for radiotherapy treatment planning in H&N.
- Methods of Quality Assurance (QA) of AI auto-segmentation software.
- Patient views on the use of AI in radiotherapy treatment planning.

Hypotheses, along with research aims, are described for each subject area, and three studies prepared for journal submission are presented.

The first study evaluates multiple commercial AI-based systems for the auto-segmentation of H&N OARs using a common patient cohort.

The second study investigates failure rates of a commercial AI-based auto-segmentation system in order to inform on appropriate QA.

The third study looks at the use of a validated patient questionnaire to develop an understanding of patient views on the use of AI-based software in radiotherapy.

Four appendices are included:

- The first appendix is a list of all taught modules undertaken as part of the DClinSci.
- The second appendix is a patient questionnaire which was distributed as part of the research carried out in study C.
- The third appendix is a business case for the purchase of a commercial AI auto-segmentation system. This has been produced to satisfy the innovation requirement of the DClinSci.
- The fourth appendix is a poster, which was presented by the author at ESTRO 2022 relating to work carried out as part of study A (Chapter 3).

All text has been written by the author and details of co-author contributions have been included under an ‘author contributions’ section at the end of each study chapter.

2. Literature Review

2.1. Introduction

This research focusses on the subject of H&N OAR auto-segmentation for radiotherapy treatment planning using AI-based software. A deductive approach has been used, and by reviewing the available literature the current state of knowledge in this area has been assessed, gaps in this knowledge identified, and hypotheses and aims of the research defined.

2.2. Literature review search methodology

The state of current available literature was established primarily using the University of Manchester Library online search facility. Further articles were discovered using references obtained from papers highlighted in the initial search, and by additional searches using PubMed and Google Scholar. Methodologies such as PRISMA (Liberati et al., 2009) were also considered when carrying out this search in order to produce results of sufficient quality.

Searches were carried out using combinations of a number of terms relating to the topics in question. Terms used for study A (Chapter 3) and study B (Chapter 4) were "*machine learning*", "*deep learning*", "*artificial intelligence*", "*segmentation*", "*autosegmentation*", "*auto-segmentation*", "*delineation*", "*atlas based*", "*validation*", "*quality assurance*", "*automated*". Some terms were combined with "*radiotherapy*", "*radiation oncology*", "*radiation therapy*", or "*head and neck*" to further screen articles.

For study C (Chapter 5), terms used were "*patient*", "*public*", "*views*", "*perceptions*", "*perspective*" and these terms were combined with "*machine learning*", "*deep learning*", "*artificial intelligence*" and "*AI*".

Key papers were then chosen and critically assessed following initial review of abstract and discussion sections.

2.3. Auto-segmentation in radiotherapy treatment planning

The delineation of OAR and target volumes is a key aspect of the H&N treatment planning process. Manual delineation of these structures can be extremely time consuming. For example Harari et al. (2010) estimated the average contouring time to be 2.7 hours for a typical H&N IMRT treatment. In addition to this significant delineation time, the process is also prone to large intra- and inter-observer variability (Peng et al., 2018). The introduction of any process which leads to a reduction in delineation time and/or an improvement in contour consistency would therefore be extremely beneficial. For example, time savings which lead to reductions in treatment planning timescales and consequently earlier treatment start dates may result in improved clinical outcomes due to a reduced risk of local recurrence (Chen et al., 2008).

The most effective early methods of automatic segmentation were atlas-based (ABAS), making use of algorithms to extrapolate from an atlas of training examples. ABAS commercial products have been clinically implemented for a variety of anatomical sites (Escande et al., 2016). Although ABAS methods have delivered a reasonable degree of success, studies have concluded that the structures produced using such methods are very much a 'starting point' (Sharp et al., 2014). In the area of H&N, Thomson et al. (2014) evaluated a commercial ABAS system and concluded that '*improvements in automatic segmentation of H&N OARs would be worthwhile and are required before routine clinical implementation*'.

More recently, interest in the use of AI has surged. AI has been defined as '*human intelligence exhibited by machines*' (Artificial intelligence, 2019). Machine learning (ML) is a subset of AI and can be defined as the use algorithms to parse data, learn from this data, and then make a prediction or determination about something in the world.

Deep learning (DL) is a further subset of ML. With the rapid increase in the power of computer chips, statistical models known as artificial neural networks have been developed. These models process data in a similar way to the human brain. Very recent advances in computer power have facilitated the ‘stacking’ of these neural networks on top of each other in connected ‘layers’ (Schmidhuber, 2015). In recent years engineers have been able to create neural networks that are up to 100 layers ‘deep’ (Bini, 2018). This is where the phrase ‘deep learning’ originates. The result of this is the ability to cope with data that is increasingly complex.

The potential for the use of DL in healthcare is wide-ranging and the UK National Health Service (NHS) is actively encouraging the adoption of this technology (Health Education England, 2018). DL applications in healthcare and specifically in the field of radiotherapy are, however, still in their infancy. In 2018 Boon et al. assessed the role of AI in clinical oncology and concluded that it is likely to continue to evolve at a rapid pace with wide-ranging application to oncology and radiotherapy.

Meyer at al. (2018) reviewed DL research works which they determined could be used at stages of the radiotherapy workflow. The literature was classified into seven different categories relating to the different radiotherapy steps, and they concluded that several different DL methods could indeed be applied to radiotherapy. They did, however, state that at the time of publication DL in radiotherapy is still at the ‘prehistory’ stage.

The first commercial use of DL in radiotherapy is in the area of auto-segmentation, and as of April 2022 there are already more than ten commercial auto-segmentation systems available.

2.3.1. Review of current literature

The literature search carried out revealed that studies relating to the auto-segmentation of H&N structures using DL are limited in number, although the frequency of such publications is increasing rapidly, reflecting the growing popularity of this technology.

Table 1 gives an overview of H&N DL studies found in the existing literature.

Table 1. Summary of Deep Learning Auto-Segmentation studies for Head and Neck

Author	OARs	DL software type	Number of comparison datasets	Comparison metrics used
Chu et al., 2016	Unknown	In-house	N/A	N/A
Ibragimov and Xing, 2017	spinal cord, mandible, parotid glands, submandibular glands, larynx, pharynx, eye globes, optic nerves, and optic chiasm	In-house CNN	50	DSC
Nikolov et al., 2018	Brain, brainstem, cochlea, lacrimal glands, lens, lungs, mandible, optic nerves, orbits, parotids, spinal-canal, spinal-cord, submandibular glands	In-house 3D U-Net	24	Surface DSC
Van der Veen et al., 2019	Brainstem, cochlea, oesophagus, glottic larynx, mandible, oral cavity, glottic larynx, parotids, pharyngeal constrictor muscles, submandibular glands, spinal cord	In house CNN	15	DSC, ASSD (Average Symmetric Surface Distance)
Brouwer et al., 2020	carotid arteries, arytenoids, brainstem, buccal mucosa, cerebellum, cerebrum, cricopharyngeal inlet, cervical esophagus, glottic area, mandible, extended oral cavity, parotid glands, pharyngeal constrictor muscles, spinal cord, submandibular glands, supraglottic larynx, thyroid gland.	Commercial	103	Mesh vertices
Van Dijk et al., 2020	parotid glands, submandibular glands, thyroid gland, arytenoids, buccal mucosa, extended oral cavity, pharyngeal constrictor muscle, cricoid, supraglottic area, glottic area, cervical esophagus, brainstem, cerebellum, cerebrum, spinal cord, mandible, carotid arteries	Commercial	104	DICE, absolute mean and max dose differences
Brunenberg et al., 2020	Parotids, submandibular glands, thyroid gland, buccal mucosa, extended oral cavity, pharynx constrictor muscle, cricoid, supraglottic area, glottic area, brainstem, mandible	Commercial	58	DSC, HD95

As early as 2016 Chu et al. proposed a study to apply ML to automate segmentation of H&N contours on radiotherapy planning CT and MRI scans. This research was in partnership with Google's DeepMind Health, and involved the development of an algorithm using ML techniques. It was a retrospective non-interventional study using a sample size of approximately 700 patient cases and results were not published, but provides an indication of when DL technology was first used for auto-segmentation.

In 2017 Ibragimov and Xing made use of deep convoluted neural networks (CNN) to develop a system to segment a number of different H&N OARs using CT datasets. Nine different OARs were segmented and the study concluded that the AI system segmented seven of the nine structures with similar or superior performance to three current commercial non-AI auto-segmentation software packages. Segmentation of submandibular glands and optic chiasm was found to be inferior, this being attributed to poorly recognisable boundaries on a CT image. A suggestion was made that MRI images may be required for accurate delineation of some structures, and this highlights the fact that no matter how effective auto-segmentation using DL becomes, there will always be a limitation relating to the quality of the CT scan in use. Optimisation of the image quality for planning CT scans is therefore an important consideration for all methods of contouring, and the importance of MR for delineation of some structures should be noted. A future direction of research in auto-segmentation is therefore likely to be using MR images, and studies have already been published investigating this (Hague et al., 2021).

In 2018 Nikolov et al. demonstrated a DL architecture which they concluded achieved '*performance similar to experts*'. The model used was trained using 663 datasets and then applied to a test set of 24 CT scans. Twenty one OARs in the H&N region were selected and findings were that 19 of the 21 OARs were segmented by the model to '*near expert radiographer level performance'*.

Performance was quantified using a new performance metric introduced by the authors and named the surface Dice-Sorenson Coefficient (surface DSC), a variation of the standard volumetric DSC (Dice, 1945), which was determined to be better suited for the presented use case. This metric was conceived to be more sensitive to errors of clinical significance for radiotherapy planning due to the potentially large effect of small differences in border placement.

The two OARs which the model segmented with inferior performance were brainstem and right lens. The authors suggested a number of possible reasons for this. One issue with brainstem was the difficulty in determining its superior extent, where it transitions into brain. Problems with the lens segmentation were attributed to the small size of this structure and difficulty in visualising borders on the CT scan. This is a comprehensive publication, which provides a good insight into levels of performance associated with DL contouring software in 2018, and introduces a new and potentially more useful performance metric.

Cardenas et al. (2018) specifically looked at using DL to automatically delineate high risk Clinical Target Volumes (CTV) in oropharyngeal cancer patients. An algorithm was developed which made use of deep auto-encoders, and a probability threshold selection function based on DSC was utilised to improve generalisation of the predicted volumes. The study used data from 52 patients who had previously been treated with curative-intent IMRT, and concluded that the generated CTVs provided close agreement with physician manual contours and that the algorithm could be implemented clinically with minor or no changes to the contours produced. An interesting proposal of the use of this technology was that it could be utilised in the peer-review process, highlighting other possibilities for the use of AI auto-segmentation beyond creation of structures for planning. The ability to produce accurate CTVs in addition to OARs also has the potential to save significant oncologist time.

Commercial DL systems

In terms of clinical evaluation of commercial DL software, the existing literature is still relatively sparse. In 2018 Lustberg et al. investigated the time savings produced with lung OAR contouring when using both ABAS and DL-based commercial software (Mirada DLC Expert^{TM}, Mirada Medical Ltd, Oxford, UK) to generate contours. The study used 20 CT scans and found that with a median manual contouring time of 20 minutes, there was a 7.8 minute time saving with the use of ABAS software, and a saving of 10 minutes with DL-based software. This time saving also highlights the amount of time required to manually correct the contours produced by the software. The DL software outperformed the ABAS software, but still required 10 minutes of time for contour adjustment to produce clinically acceptable structures.

One point of note with this publication is that 20 patients were used to generate the ABAS model, but contours from 450 lung patients were required to train the DL model. This is a considerable number of patients, and there are therefore significant time resource implications for the creation of DL models, which should be considered when undertaking such a task. This may not be a concern if models are created and provided by manufacturers, but will be more relevant if manufacturers are providing ‘custom’ models which utilise local patient data. It should also be noted that the DL software used in this study was a ‘prototype’ version, in the early stages of development, indicating that there was still considerable scope for improvement with this new technology.

There are fewer existing studies which evaluate the performance of DL systems for H&N OAR segmentation. In 2020 Brouwer et al. evaluated head and neck contours produced by Mirada DLC Expert^{TM}. This study investigated the manual adjustment required when used in clinical practice, and results were also compared to previously-reported data regarding inter-observer variation.

Evaluation of the degree of manual adjustment required made use of mesh vertices, which is an interesting and unusual methodology. The metric used was defined as the shortest surface displacement from the auto-segmented contour to the final edited contour and was calculated with the aid of deformable registration software. Whilst this method may have its merits, it makes this study difficult to compare to other studies which use more common metrics to evaluate contour differences, such as DSC and Hausdorff distance (HD).

Results showed that for the majority of structures the median values for editing and for adjustment were within 2mm. Reasons for the need to adjust auto-segmented contours were partly attributed to interpretation of delineation guidelines. There was also an observation that the DLC model under-segmented some contours, such as parotid glands, and it was hypothesised that this may be caused by inter-observer variability in training datasets, causing the model to average data and leading to a 'shrinking' effect. It should be noted that the model utilised for this study was produced in 2019 and has since been updated and improved several times by the manufacturer.

Overall, this study provided some useful information regarding differences between AI auto-segmented and human expert contours in H&N, but used non-standard comparison metrics and did not inform on any associated time savings or the clinical significance of contour differences.

Another study carried out in 2020 by Van Dijk et al. compared Mirada DLC Expert^{TM} to ABAS for a number of commonly used H&N OARs. Comparison metrics of DSC and HD were used, along with mean and maximum dose differences. The study also evaluated contouring time, inter-observer variation, and carried out a qualitative evaluation using a Turing test.

Results showed that DLC was significantly superior to ABAS and that clinical use of DLC would reduce overall delineation time compared to ABAS. Comparisons between DLC and human experts suggested that DLC was approaching similar levels to that of inter-observer variability. The subjective evaluation did, however, conclude that manual contours were still preferable to DLC, indicating that further improvements to the DL model are still required.

Overall this was a comprehensive piece of research which provided useful information regarding the state of AI auto-segmentation in 2019/2020. The use of mean and maximum dose difference was questionable as the method used involved recalculating dose using the original clinical treatment plan. In reality the use of different OAR contours would result in the creation of a different treatment plan, which calls into question the usefulness of such a dosimetric comparison.

A further study carried out by Brunenberg et al. in 2020 performed an independent validation of the model used by van Dijk et al. on a set of 58 H&N cancer patient datasets. This work evaluated contours both quantitatively (using DSC and HD) and also qualitatively using a Turing test.

Quantitative results were comparable to those obtained by van Dijk for glandular OARs and mandible, but for some other OARs, such as those in the aerodigestive tract, scores were substantially lower than those obtained in the original study. It was suggested that differences could be at least partly attributed to inter-observer variation in relation to the reference datasets used in this study, as both studies used the same delineation guidelines. This study can be used as a good example of the importance of having a robust reference dataset when validating auto-segmentation software.

With regard to a comparison of multiple commercial AI auto-segmentation systems no existing research was found, highlighting a clear gap in the current literature. In 2021 Robert et al. described the methodology used by three different French radiotherapy centres to clinically deploy three different CE-marked commercial auto-segmentation systems for a variety of anatomical sites. This study provided useful information for other centres to draw on when commissioning such products, but although three different commercial systems were referenced, there was no direct comparison between systems other than a table summarising system properties. Harrison et al. (2022) also list names and manufacturers of some commercial AI auto-segmentation systems, but again do not compare their relative performance. It is important to note that systems comparisons are difficult without the availability of a common patient dataset.

2.3.2. Study A hypothesis, aims and objectives

The following hypothesis was made regarding the first research study presented in this thesis:

- Multiple commercial AI auto-segmentation systems will produce results of comparable quality for H&N OARs.

The aim of this area of the research was to compare multiple commercial AI auto-segmentation systems for delineation of commonly used H&N OARs using a common validation patient dataset.

Objectives of study A will be as follows:

- To produce contours for a selection of commonly used head and neck OARs on 50 patient datasets using each commercial system.
- To assess the quality of contours produced by each system by comparing with 'gold standard' contours using multiple similarity metrics.
- To provide recommendations relating to the clinical introduction of AI auto-segmentation software in H&N radiotherapy planning.

These objectives are considered to be both measurable and realistic given the availability of the required software and suitable H&N patient numbers treated at the centre. Patient numbers used will be sufficient to obtain good statistical power.

2.4. Quality Assurance and failure rates of AI auto-segmentation software

Any piece of software used for medical purposes is classed as a medical device and is subject to associated regulations. In addition to these regulations, international guidelines exist which advise on appropriate QA for treatment planning systems. For example, the American association of physicists in medicine (AAPM) radiation therapy committee task group 53 published guidance around such QA in 1998. The report states that '*inaccuracy in definition of the anatomical model of the patient may be one of the largest sources of uncertainty in the entire RTP process*'. This statement highlights the importance of adequate commissioning and QA of any auto-segmentation software that is in clinical use.

QA for auto-segmentation software can be separated into QA that is required to commission a system for clinical use and to routinely check system performance, and QA that is required for every individual patient and contour that is generated (patient-specific QA). Due to the nature of such software, and variations in individual patient anatomy, current practice is to manually check (and adjust if necessary) resulting contours for every patient, and this checking of contours for every individual patient can be a time consuming process (Brouwer et al., 2020).

As the quality of contours produced by AI auto-segmentation software improves, it may be that such patient-specific QA is no longer required, but the existing literature suggests the technology has not yet advanced to this level (van Dijk et al., 2020 and Claessens et al., 2022).

2.4.1. Review of current literature

Contour validation

A summary of existing literature relating to automatic contour QA is shown in table 2. Studies can be classed as using one of two methods for QA, these being data abstraction and use of a secondary auto-segmentation system.

In terms of methods to validate the contours produced by auto-segmentation software, Valentini et al. (2014) discuss how evidence can be obtained. They separate the issues into three distinct areas, namely ontology, performance evaluation and benchmark evaluation, and produce a set of recommendations covering these areas.

Table 2. Summary of automatic contour QA studies

Author	Anatomical site investigated	Software development type	QA Methodology
Altman et al., 2015	Head and Neck	In-house	Data abstraction
Chen et al., 2015	Head and Neck	In-house	Data abstraction
Court et al., 2018	All	Unknown	Secondary Atlas Based
Hui et al., 2018	Thoracic	In-house	Data abstraction
Shah et al., 2018	Pelvis	In-house	Data abstraction
Rhee et al., 2019	Head and Neck	In-house	Secondary AI
Men et al., 2020	Lung	In-house	Secondary AI
Claessens et al., 2022	Pelvis	Commercial	Secondary AI
Du et al., 2022	Head and Neck	In-house	Data abstraction

In the context of this paper, ontology refers to a form of dictionary containing the information required to be able to delineate a structure i.e. delineation guidelines along with associated anatomical and pathological information. The importance of ontology is discussed, and concerns are raised around the existence of multiple endorsed atlases for each anatomical site, highlighting the importance of understanding the choice of ontology used by auto-segmentation systems and how this ontology is propagated when used clinically.

Regarding performance evaluation, a suggestion is made that a combination of conformation scores, metric elements and clinical risk assessment could be used to produce a new class of performance indices, and the paper also stresses the importance of measuring the time aspect of the process.

For benchmark evaluation the importance of having a 'gold standard' structure set is discussed, and suggestions are made around how such a structure set may be produced, for example by combining the knowledge of multiple expert users.

This paper adds valuable knowledge to the literature regarding the different types of evidence that may be required to evaluate such systems. The paper was published in 2014 and there are now more recent international consensus guidelines (Mir et al., 2020) which could be incorporated into these recommendations to modernise them. These new international guidelines are also likely to reduce issues encountered by the study in relation to ontology. In addition, the use of newer comparison metrics such as surface DSC and APL, which are better indicators of time savings (Vaassen et al., 2020) could be incorporated into any updated recommendations.

Vandewinckele et al. (2020) produced guidance around the implementation and QA of AI models in radiotherapy. For case-specific QA, recommendations include '*to keep a log of all corrections required*' and '*to keep track of poorly performing cases*'. The theory behind this is that the information can be used to further improve the model. Realistically it may be impractical to log all corrections made, as there are likely to be small edits for almost every patient. The suggestion to keep a log of poorly performing cases is more realistic, and where a commercial system is being used the author would suggest that this information is routinely fed back to manufacturers.

Discussions around the advantages and disadvantages of commonly used similarity metrics are included in this work, and the important point is made that achieving an accuracy comparable to intra and inter-observer variability is a good indication that a system produces contours of high accuracy.

This paper also advises that every automatically generated contour should be reviewed, corrected if required, and approved by a human. This is consistent with current thinking regarding the checking of auto-segmented contours but is clearly a time-consuming process. There is mention of possible automated methods that can be used to detect outliers. For example, a paper published by Chen et al. in 2015 which makes use of geometric attribute distribution models to identify unusual structure attributes is referenced, along with a paper from Court et al. (2018) where the possibility of using a second auto-segmentation system for QA is discussed. Both of these options may be useful for automatic identification of errors in auto-segmented structures.

Claessens et al. (2022) similarly suggest that use of a second independent DL model may be an appropriate method of QA. This study assessed prostate OARs, but the methodology used could equally be applied to H&N OARs. Nine different quantitative comparison metrics were utilised, and in addition the metrics were used as input features for a ML classifier to determine segmentation quality from the primary model.

The required adaptation was classed as either minor or major, and the resulting ML classifiers successfully highlighted all cases where a major adaptation was required. There were also, however, a number of false negatives in the results, for example with the prostate structure 50% of minor cases were assigned to the wrong class by the ML classifiers. This level of performance would still be clinically useful, and highlights the potential benefit of using such automated checking.

Further suggested discussion points regarding this research are the technical similarity and the relative performance of the two models used for producing and checking clinical structures. If systems use the same underlying technology there may be a risk that they will exhibit the same limitations, even if they are independent in terms of source datasets. This could in theory mean there is a risk that a gross contouring error would not be identified by the checking system. Also, in clinical use it would make sense from an efficiency point of view to use the highest performing system to generate, rather than check, contours. In this study a commercial system (Mirada DLC Expert^{TM}) was used as the checking system and no indication was given about the relative performance of the two systems.

In 2022 Harrison et al. published a study which provided an overview of auto-segmentation techniques, and specifically focussed on the use of ML and DL. The study discussed QA options in terms of the evaluation of system performance using simultaneous truth and performance-level estimation (STAPLE) and similarity and truth estimation for propagated segmentations (STEPS). STAPLE (Warfield, Zou and Wells, 2004) is a method of fusing together multiple segmentations using an expectation-maximisation (EM) algorithm and can be used to combine contours produced by multiple clinicians. STEPS is a newer algorithm and an extension to the STAPLE algorithm, which is also used to combine multiple segmentations and can produce superior results (Cardoso et al., 2013). Use of such high quality QA datasets is likely to produce more robust QA results and is an important recommendation.

The study by Harrison et al. (2022) also discusses options regarding comparison metrics, but does not discuss patient specific QA beyond suggesting that clinical use of auto-segmentation is normally used in combination with manual editing by clinicians, highlighting the limited research around individual patient QA.

An interesting paper published by Rhee at al. in 2019 looked at detection of errors for a multi-ABAS system which was in clinical use, using a locally developed auto-contouring tool based on CNN. Errors were classified as contours needing either minor or major edits, as described by Cardenas et al. (2017) in a study looking at peer review QA. A major error was defined as one which could potentially affect patient outcomes, and a minor error was defined as where contours required more elective or stylistic changes.

This categorisation was determined by a human expert, and it could therefore be argued that this is quite subjective, and a more scientific approach should be used for such definitions. The work by Rhee did also employ a more scientific approach by quantifying the relationship between DSC results and physician scores using receiver operating characteristic (ROC) curves.

The CNN-Based tool was demonstrated to be effective at detecting major errors in the majority of OARs, and this approach could potentially be adopted for the checking of a commercial AI auto-segmentation system, although consideration would need to be given to the risk of false negatives if both clinical and QA systems used the same underlying technology (DL). It is important to note that this tool was locally developed and is therefore not available widely.

The majority of studies which look at auto-segmented contour validation effectively use the oncologist as a QA tool. This is clearly a very inefficient process and is also likely to introduce inter-observer inconsistencies due to the likelihood of oncologists modifying contours. Furthermore, if oncologists are required to check all auto-segmented contours then efficiency savings are inevitably reduced. There is therefore a clear gap in the existing literature relating to alternative QA options.

Failure rates of modern AI auto-segmentation systems

To determine the most appropriate method of QA for such systems it is important to understand the likelihood of failure. Information regarding failure rates of modern commercial AI auto-segmentation systems would therefore be extremely useful to inform future practice regarding QA requirements.

Despite an extensive literature search on this subject no existing studies were found which provided such information. Some studies did look at the amount of manual adjustment required after auto-contouring, which suggests a significant percentage of structures do require adjustment, but no detail on whether further manual adjustment produced a clinically significant difference was available.

Further to this, when using similarity metrics such as DSC to compare differences between human expert contours and auto-segmented contours, and when evaluating inter-observer variability, DSC values are frequently of comparable magnitude (Wong et al., 2020), indicating that modern auto-segmentation systems may already have reached human levels of performance.

2.4.2. Study B hypothesis, aims and objectives

The following hypothesis has been made regarding the second research study presented in this thesis:

- Gross failure rates of commercial AI auto-segmentation systems are low and human inspection is therefore an inefficient method of QA.

The aim of this study was to assess failure rates of a commercial AI auto-segmentation system for head and neck OARs and to consider implications for appropriate QA methods for such systems.

Objectives of study B will be as follows:

- To produce expected mean and standard deviation values for the metric of DSC for commonly used H&N OARs when compared to ground truth contours.
- To auto-segment these H&N OARs for a very large patient cohort using a commercial AI auto-segmentation system. This large patient cohort is required to obtain sufficient statistical power if failure rates are low, as hypothesised.

- To determine failure rates of the auto-segmentation for multiple OARs and identify any common modes of failure.
- To recommend appropriate QA methodologies based on knowledge of true failure rates.

2.5. Patient views on the use of Artificial Intelligence in radiotherapy treatment planning

The importance of patient involvement in their own healthcare is now widely recognised as being extremely important (Brett et al., 2014). This involvement may be related to direct decision making during their own care, or it may be related to being involved in wider decisions which shape healthcare services.

The use of AI in healthcare is no exception to this, and it is important that patients have sufficient levels of knowledge such that they are not worried, and that they understand the benefits of using this technology.

2.5.1. Review of current literature

Research into both healthcare professional and patient opinions relating to the use of AI in healthcare is extremely limited (Shinners et al., 2020 and Yakar et al., 2022). Although the focus of the study in this thesis will be on the views of patients on the use of AI in healthcare, it is also important to have an understanding of healthcare professional views on AI, as these key stakeholders are likely to be an important source of information for patients on the subject.

Healthcare professional perceptions of AI

In 2020 Shinners et al. published a paper which explored the understanding of healthcare professionals of AI. This work involved carrying out an integrative review of the existing literature from the time period 2010-2018 and identified a single study which met all inclusion criteria. This highlights a clear gap in the literature that existed at this time.

Following on from this previous study, in 2022 Shinners et al. published work evaluating and piloting a questionnaire designed to explore the perceptions of healthcare professionals of AI. The questionnaire was named the Shinners Artificial Intelligence Perception (SHAIP) questionnaire, and had been developed in a previous study (Shinners et al., 2021). It made use of exploratory factor analysis to group items together and resulted in a ten item questionnaire grouped into two factors, namely 'Professional impact of AI' and 'Preparedness for AI'.

Approximately 3000 Australian healthcare employees from varied roles were invited to take part in the study and 252 responses were received. Their findings were that use of AI influenced the perception of both of the factors, with those who routinely used AI strongly understanding that it would impact their role, but also feeling more prepared for its use. This highlights the importance of educating staff who have not yet directly encountered AI in their role in order to prepare them for the future.

Coppola et al. (2021) carried out a study to gather Italian radiologist opinions on AI. They surveyed 1032 radiologists using an online method, and some of the key findings were:

- 73% believed the use of AI would result in a lower diagnostic error rate.
- 60% believed that there was a risk of a poorer reputation for radiologists when compared to non-radiologists.
- 78% of respondents did not have any ethical concerns over the use of AI.
- 89% were not afraid that they might lose their job.
- 77% were favourable to the adoption of AI tools in radiological practice.

This research concluded that there is a mainly positive attitude towards AI among a profession that, some have predicted, may be significantly affected (Malamateniou et al., 2021), and that the main concern was around professional reputation.

With regard to the views of radiation oncology professionals, Wong et al., (2021) used a questionnaire to gather information about views of radiation oncologists, radiation therapists and medical physicists in Canada. A 29 Likert-scale questionnaire was developed and distributed electronically. There were 159 respondents and results showed that the majority did not feel well versed with AI knowledge, and only 20% felt comfortable with AI systems performing without human interference. There was, however, a strong overall feeling that patients would positively benefit from the use of AI in radiation oncology (87%). This research further supports the view that more staff education is required around AI.

Patients' views on AI in healthcare

Regarding views around AI in healthcare of the public and of patients, a small number of studies were identified when searching the literature.

A study by Fast and Horvitz in 2016 focussed on views expressed about AI in the New York Times over a thirty year period. It reported that hopes for the use of AI in healthcare have increased significantly in recent years, and that there is now more optimism than pessimism. The study did, however, identify some specific concerns which appeared to be growing. These concerns included loss of control, ethical issues and the potential negative impact of AI on work due to the displacement of human workers. Whilst the increasing levels of optimism are a positive finding of this study, the growing levels of concern that were identified are more worrying, and these concerns will need to be addressed in order to successfully implement the use of AI in healthcare.

Research carried out by Longoni, Bonezzi and Morewedge in 2019 looked at patient receptivity to medical AI. This work discussed how patients may *directly* drive the adoption of AI, for example when patients interact with autonomous tools themselves, and also how they may *indirectly* determine the adoption of AI in healthcare, which is likely to be the case where healthcare professionals are mediators. Use of AI auto-segmentation falls more into the *indirect* category, although it may be argued that it is even further removed than this, and currently most patients are unlikely to be aware of the use of such systems at all.

The paper proposed that patients would be less likely to utilise healthcare delivered by AI providers than they would be to utilise healthcare delivered by comparable humans. Results indicated that there was strong resistance to the use of AI compared to a human provider, even if the AI offered superior performance. Reasons for this were identified as a perception that AI could not take into account the ‘uniqueness’ of individual cases. This study highlights the need to improve patient receptivity of AI in medicine.

In 2020 Pezzo and Beckstead published a commentary on the research from Longoni, Bonezzi and Morewedge (2019), where they point out that a statistical analysis of the results demonstrates that people did in fact prefer AI when it outperformed humans. This important piece of information was not part of the conclusion in the original paper.

In 2022 Yakar et al. published the results of an online survey which looked at the Dutch general population’s views on AI in medicine. The study focussed on radiology, robotic surgery and dermatology, and the final sample size was 1909 individual responses. Five point Likert scale questions were used to gather information around personal demographics and trust in the three different domains being studied.

Overall, the research found that the general population is quite distrustful of AI in medicine, although certain demographic groups such as educated males of a Western background had more trust than most. This is another study which highlights the need to improve public perception of AI in medicine.

Lennartz et al. (2021) also surveyed patients to ascertain views around the use of AI in different aspects of the medical workflow. The survey was taken by 229 patients who were scheduled for imaging scans at a diagnostic imaging centre in Cologne, Germany. The questionnaire comprised of five subsections covering areas such as confidence in physician vs AI, opinion of human control of AI and acceptance of AI for diagnosing and treating diseases of different severity. A combination of Likert scale and binary answer questions were utilised.

Findings indicated that patients had significantly more confidence in physicians than AI for all almost all clinical tasks, although use of AI under physician supervision was considered to be more acceptable. Once again this highlights the current levels of scepticism in the general public when it comes to the use of AI in medicine.

Patients' views on AI in Radiology

Radiology is one of the more commonly surveyed areas in relation to patient views on AI, possibly due to the publication of studies which gained notable prominence in the media (McKinney et al., 2020).

In 2019 Haan et al. published a qualitative study designed to ascertain an understanding of patient knowledge levels of AI in radiology and to identify associated domains. The study involved semi-structured face to face interviews and a small financial incentive was utilised, and the outcome of the study was that there were diverse levels of knowledge among patients, and that there was a lack of understanding of staff roles, and of how AI might be used in radiology.

The study identified six key domains, namely 'proof of technology', 'procedural knowledge', 'competence', 'efficiency', 'personal interaction', and 'accountability'. It was suggested that these six domains could provide a framework for patient education and future research relating to the clinical implementation of AI systems in radiology. Just twenty participants were involved in this study, which is a relatively small number, and therefore future work may benefit from larger patient numbers in order to provide greater statistical power.

Following on from this study, in 2020 Ongena et al. made use of the six domains that had been identified by Haan et al. (2019) to develop a questionnaire designed to measure patient acceptance of the implementation of AI in radiology. This research utilised exploratory factor analysis to identify five factors relating to the use of AI in radiology. The resulting questionnaire used five point Likert scales and was distributed to 155 patients.

Findings from the completed questionnaires were that patients are not overly optimistic about the ability of AI systems to perform tasks that are currently performed by humans, and a strong desire for patients to maintain human interaction was identified. This lack of optimism supports the findings of other studies, and once again highlights the need to educate and, if possible, change patient views on the use of AI in medicine in order to increase acceptance in the future.

Patients' views on AI in radiotherapy

A literature search carried out in April 2022 did not find a single study relating to patient views on the use of AI in radiotherapy.

There is therefore a clear gap in the availability of literature relating to this subject, and this will be investigated in this thesis by gathering patient opinions regarding the use of AI and computer automation in their own radiotherapy treatment planning.

2.5.2. Study C hypothesis, aims and objectives

The following hypothesis has been made regarding the third research study presented in this thesis:

- Patients have significant concerns about the use of AI-based software in their own radiotherapy treatment.

The aim of this study was to use a standardised patient questionnaire to develop an understanding of patient views on AI and its use in their own radiotherapy treatment process.

Objectives of study C were as follows:

- To adapt questions from an existing, validated questionnaire on AI in healthcare to the field of radiotherapy.
- To survey a sufficiently large number of patients to obtain meaningful views.
- To provide recommendations based on the results obtained.

2.6. Conclusions and research aims

Hypotheses, aims and objectives have been described for three research studies relating to the use of AI-based software in the field of radiotherapy.

All three studies address clear gaps in the existing literature and findings are expected to lead to new knowledge that can be utilised to improve radiotherapy processes and ultimately produce tangible clinical patient benefits.

It can be hypothesised that benefits might include:

- Reduction of waiting times between referral and radiotherapy treatment due to increased efficiency in treatment planning processes (Kosmin et al., 2019).
- Improved clinical outcomes as a result of increased consistency in OAR outlining (Sherer at eal., 2021).
- Quality and efficiency improvements in the area of adaptive planning, where treatment plans are ‘adapted’ to cope with patient changes such as tumour regression, local inflammation, changes in weight, and alterations in tissue distribution (Chen et al., 2014).
- Increased patient understanding, and consequently reassurance, around the use of AI in their treatment planning processes (Longoni et al., 2019).
- To allow healthcare professionals to spend time working on other areas of patient care due to the time-savings produced by this new technology.

2.7. References

‘Artificial intelligence’ (2019).Wikipedia. Available at: https://en.wikipedia.org/wiki/Artificial_intelligence (Accessed: 20 February 2019).

Altman, M.B. et al. (2015). A framework for automated contour quality assurance in radiation therapy including adaptive techniques. *Physics in Medicine and Biology*, 60(13), pp.5199–5209.

Bini, S.A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? Journal of Arthroplasty, 33(8), pp.2358–2361. [online]. Available from: https://doi.org/10.1016/j.arth.2018.02.067.

Boon, I., Au Yong, T. and Boon, C. (2018). Assessing the Role of Artificial Intelligence (AI) in Clinical Oncology: Utility of Machine Learning in Radiotherapy Target Volume Delineation. Medicines, 5(4), p.131. [online]. Available from: http://www.mdpi.com/2305-6320/5/4/131.

Brett, J. et al. (2014). Mapping the impact of patient and public involvement on health and social care research: A systematic review. Health Expectations, 17(5), pp.637–650.

Brouwer, C.L. et al. (2020). Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy. Physics and Imaging in Radiation Oncology, 16(June), pp.54–60. [online]. Available from: https://doi.org/10.1016/j.phro.2020.10.001.

Cancer.Net (2002) Head and Neck Cancer: Statistics. Available at: https://www.cancer.net/cancer-types/head-and-neck-cancer/statistics (Accessed: 11 June 2022).

Cancer Research UK (2022). Head and neck cancers statistics. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers (Accessed: 11 June 2022).

Cardenas, C.E. et al. (2018). Auto-delineation of Oropharyngeal Clinical Target Volumes Using Three-Dimensional Convolutional Neural Networks. Physics in Medicine and Biology.

Cardenas, C.E. et al. (2018). Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. International Journal of Radiation Oncology Biology Physics, 101(2), pp.468–478. [online]. Available from: https://doi.org/10.1016/j.ijrobp.2018.01.114.

Cardoso, M. J. et al. (2013). STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical Image Analysis, 17(6), pp.671–684. [online]. Available from: http://dx.doi.org/10.1016/j.media.2013.02.006.

Chen, Z. et al. (2008). The relationship between waiting time for radiotherapy and clinical outcomes: A systematic review of the literature. Radiotherapy and Oncology, 87(1), pp.3–16.

Chen, H.C. et al. (2015). Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy. Medical Physics, 42(2), pp.1048–1059.

Chu, C. et al. (2016). Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans. F1000Research, 5(0), p.2104. [online]. Available from: https://f1000research.com/articles/5-2104/v1.

Claessens, M. et al. (2022). Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm. Physics in Medicine & Biology, 67(11), p.115014.

Court, L.E. et al. (2018). Radiation planning assistant - A streamlined, fully automated radiotherapy treatment planning system. Journal of Visualized Experiments, 2018(134), pp.1–9.

Daisne, J.F. and Blumhofer, A. (2013). Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation. Radiation Oncology, 8(1), pp.1–11.

Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published. Ecology, 26(3), pp.297–302. [online]. Available from: http://www.jstor.org/stable/1932409 .

'Deep learning' (2019).Wikipedia. Available at: https://en.wikipedia.org/wiki/Deep_learning (Accessed: 20 February 2019).

Du, D. et al. (2022). Automatic organ contour check: One essential step in autonomous treatment planning. Medical Dosimetry, 47(2), pp.197–201. [online]. Available from: https://doi.org/10.1016/j.meddos.2022.02.006.

Hague, C. et al. (2021). An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. Radiotherapy and Oncology, 158, pp.112–117. [online]. Available from: https://doi.org/10.1016/j.radonc.2021.02.018.

Hui, C.B. et al. (2018). Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. *Medical Physics*, 45(5), pp.2089–2096.

El Naqa, I. et al. (2018). On the Fuzziness of Machine Learning, Neural Networks, and Artificial Intelligence in Radiation Oncology. International Journal of Radiation Oncology*Biology*Physics, 100(1), pp.1–4. [online]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S036030161731060X.

Escande, A. et al. (2016). Comparison of Automated Atlas-Based Segmentation Software for Postoperative Prostate Cancer Radiotherapy. Frontiers in Oncology, 6(August), pp.1–6.

Fritscher, K.D. et al. (2014). Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Medical Physics, 41(5), pp.1–11.

Harrison, K. et al. (2022). Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clinical Oncology, 34(2), pp.74–88. [online]. Available from: https://doi.org/10.1016/j.clon.2021.12.003.

Health Education England. (2018). The Topol Review. Preparing the healthcare workforce to deliver the digital future. [online]. Available from: www.hee.nhs.uk.

Kosmin, M. et al. (2019). Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiotherapy and Oncology, 135, pp.130–140. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.03.004.

Liberati, A. et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Medicine, 6(7).

Longoni, C., Bonezzi, A. and Morewedge, C.K. (2019). Resistance to Medical Artificial Intelligence. Journal of Consumer Research, 46(4), pp.629–650.

Lustberg, T. et al. (2018). Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiotherapy and Oncology, 126(2), pp.312–317. [online]. Available from: https://doi.org/10.1016/j.radonc.2017.11.012.

‘Machine learning’ (2019).Wikipedia. Available at: https://en.wikipedia.org/wiki/Machine_Learning (Accessed: 20 February 2019).

Malamateniou, C. et al. (2021). Artificial intelligence in radiography: Where are we now and what does the future hold? Radiography, 27, pp.S58–S62. [online]. Available from: https://doi.org/10.1016/j.radi.2021.07.015.

McKinney, S.M. et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), pp.89–94. [online]. Available from: http://dx.doi.org/10.1038/s41586-019-1799-6.

Men, K. et al. (2020). Automated Quality Assurance of OAR Contouring for Lung Cancer Based on Segmentation With Deep Active Learning. *Frontiers in Oncology*, 10(July), pp.1–7.

Meyer, P. et al. (2018). Survey on deep learning for radiotherapy. Computers in Biology and Medicine, 98(May), pp.126–146. [online]. Available from: https://doi.org/10.1016/j.compbiomed.2018.05.018.

Mir, R. et al. (2020). Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines: GHG OAR consensus contouring guidance. Radiotherapy and Oncology, 150, pp.30–39. [online]. Available from: https://doi.org/10.1016/j.radonc.2020.05.038.

Moher, D. et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ, 339(jul21 1), pp.b2535–b2535. [online]. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.b2535.

Nikolov, S. et al. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. , pp.1–31. [online]. Available from: http://arxiv.org/abs/1809.04430.

Rhee, D.J. et al. (2019). Automatic detection of contouring errors using convolutional neural networks. Medical Physics, 46(11), pp.5086–5097.

Robert, C. et al. (2021). Clinical implementation of deep-learning based auto-contouring tools–Experience of three French radiotherapy centers. Cancer/Radiotherapie, 25(6–7), pp.607–616.

Rohlfing, T., Russakoff, D.B. and Maurer, C.R. (2004). Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Transactions on Medical Imaging, 23(8), pp.983–994.

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. Neural Networks, 61, pp.85–117. [online]. Available from: http://dx.doi.org/10.1016/j.neunet.2014.09.003.

Shah, V.P. et al. (2018). Data integrity systems for organ contours in radiation therapy planning. *Journal of Applied Clinical Medical Physics*, 19(4), pp.58–67.

Sharp, G. et al. (2014). Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Medical Physics, 41(5), p.050902. [online]. Available from: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.4871620.

Sherer, M. V. et al. (2021). Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiotherapy and Oncology*, 160, pp.185–191. [online]. Available from: https://doi.org/10.1016/j.radonc.2021.05.003.

Shinners, L. et al. (2020). Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: An integrative review. Health Informatics Journal, 26(2), pp.1225–1236.

Shinners, L. et al. (2021). Exploring healthcare professionals' perceptions of artificial intelligence: Validating a questionnaire using the e-Delphi method. Digital Health, 7, pp.1–9.

Shinners, L. et al. (2022). Exploring healthcare professionals' perceptions of artificial intelligence: Piloting the Shinners Artificial Intelligence Perception tool. Digital Health, 8.

Sung, H. et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 71(3), pp.209–249.

Thomson, D. et al. (2014). Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiation Oncology, 9(1), pp.1–12.

Vaassen, F. et al. (2020). Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology, 13(November 2019), pp.1–6. [online]. Available from: https://doi.org/10.1016/j.phro.2019.12.001.

Wang, Y. et al. (2018). Automatic Tumor Segmentation with Deep Convolutional Neural Networks for Radiotherapy Applications. Neural Processing Letters, 48(3), pp.1323–1334. [online]. Available from: http://link.springer.com/10.1007/s11063-017-9759-3.

Warfield, S.K., Zou, K.H. and Wells, W.M. (2004). Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging, 23(7), pp.903–921.

Willems, S. et al. (2018). Clinical Implementation of DeepVoxNet for Auto-Delineation of Organs at Risk in Head and Neck Cancer Patients in Radiotherapy. In Lecture Notes in Computer Science. Springer International Publishing, pp. 223–232. [online]. Available from: http://link.springer.com/10.1007/978-3-030-01201-4.

Wong, J. et al. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiotherapy and Oncology, 144, pp.152–158. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.10.019.

Xing, L., Krupinski, E.A. and Cai, J. (2018). Artificial intelligence will soon change the landscape of medical physics research and practice. Medical Physics, 45(5), pp.1791–1793. [online]. Available from: http://doi.wiley.com/10.1002/mp.12831.

Yakar, D. et al. (2022). Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine. Value in Health, 25(3), pp.374–381. [online]. Available from: https://doi.org/10.1016/j.jval.2021.09.004.

3. Study A: An evaluation of multiple deep learning-based auto-segmentation systems for head and neck cancer

Simon Temple, Carl Rowbottom
The Clatterbridge Cancer Centre NHS Foundation Trust, Liverpool, UK.

Abstract

Purpose/Objective

Commercial software utilising deep learning can be used to automatically delineate organs at risk (OAR) on CT scans with the potential for significant efficiency savings in the radiotherapy treatment planning pathway, and simultaneous reduction of inter- and intra-observer variability.

Vendors of commercial systems often claim superiority of their own system in comparison to competitor systems. To date there has been limited research comparing multiple systems using multiple comparison metrics and a common patient cohort. This has been addressed in this study.

Materials/Methods

Four different deep learning-based auto-segmentation systems, which had been independently developed for commercial use, were used to create five commonly used head and neck (H&N) OARs (brainstem, spinal cord, mandible, left and right parotid), for 50 H&N patient datasets. All systems were running their latest available software version at the time of study (June 2021 – Sep 2021).

The resulting auto-segmented contours were compared to ‘gold standard’ clinical contours, created by Consultant Clinical Oncologists at our centre. All data used originated from patients entered into the PATHOS clinical trial. The associated trial protocol includes clear anatomical guidelines for OAR delineation and, in addition, trial entry involved pre-trial OAR outlining Quality Assurance, which all Oncologists were required to undertake. A sample of patient data was retrospectively reviewed during the trial, to provide further assurance around the quality of contours used.

Standard similarity metrics of 3D Dice Similarity Coefficient (DSC), Added Path Length (APL) and 2D 95% Hausdorff (HD) were utilised for the study.

Results

For all OARs and all systems tested, results obtained for 3D DSC, 2D 95% HD and APL correlate well within the range of other recent published studies.

Performance differences between the four systems varied with OAR. Overall differences between the systems were relatively small.

Using mean 3D DSC values for all structures, System 2 was found to perform most effectively, but the difference between the first and second highest performing systems for each OAR was not statistically significant.

Conclusion

Comparable levels of performance were observed between all four systems. This indicates that deep learning-based auto-segmentation products are developing at a similar pace in terms of the quality of contours produced.

It is therefore likely to be more beneficial to consider other factors such as financial cost and range of contours offered when considering the evaluation of such a system for clinical use.

Introduction

Modern radiotherapy treatment planning to allow delivery of techniques such as volumetric modulated arc therapy (VMAT) requires accurate delineation of both target volumes and organs at risk (OAR). Manual delineation of these structures can often be a time-consuming process (Thomson et al., 2014), and in addition is prone to both inter and intra-observer variability, even when published international consensus guidelines (ICG) are available.

For example, van der Veen et al. (2021) found statistically significant inter-observer variability between radiation oncologists who worked in the same centre and made use of the same ICG.

VMAT radiotherapy treatments for head and neck (H&N) cancer can be especially complex due to the relatively large number of critical OARs in this part of the body (van der Veen, 2017). Auto-segmentation for this anatomical site is therefore of particular interest, as there is the potential for significant time savings to be made in the delineation stage of the planning process, as well as a reduction in inter and intra-observer variability.

Historically auto-segmentation software has either used model-based approaches (Fritscher et al., 2007), atlas-based (Sims et al., 2009), or a combination of both (Fritscher et al., 2014). The quality of structures produced using such systems is limited (Teguh et al., 2011), and it has been debatable whether the benefits offered were sufficient to justify the financial and resource cost of clinical implementation. For example, in 2014 Thomson et al. evaluated a commercially available system that utilised the combined approach to outline H&N OARs, and did not recommend clinical implementation of the system until further improvements in accuracy were made, as the resulting contours were less accurate than manual segmentation and there were no associated time-savings.

In more recent years the effectiveness of artificial intelligence (AI) based software has improved significantly due to advancements in computer chip technology and power.

AI has been defined as ‘*human intelligence exhibited by machines*’ (Artificial intelligence, 2019). Machine learning (ML) is a subset of AI and can be defined as the use of algorithms to parse data, learn from this data, and then make a prediction or determination about something in the world.

Deep Learning (DL) is a further subset of ML. With the rapid increase in the power of computer chips, statistical models known as artificial neural networks (ANN) have been developed. These models process data in a similar way to the human brain. Very recent advances in computer power have facilitated the ‘stacking’ of neural networks on top of each other in connected ‘layers’ (Schmidhuber, 2015), allowing engineers to create neural networks that are up to 100 layers ‘deep’ (Bini, 2018). This is where the phrase ‘deep learning’ originates. The result of this is the ability to cope with data that is increasingly complex.

DL technology has evolved to the point where auto-segmentation performance is now superior to traditional atlas-based solutions (Lustberg et al., 2018) and there are currently multiple commercially available DL auto-segmentation systems. Several hundred datasets are typically required for initial model training of such systems (van Dijk et al., 2020).

The aim of this study was to compare the quality of contours produced by four different DL-based auto-segmentation systems to those produced by a trained human expert. To the authors knowledge there has not yet been a publication that compares multiple DL-based auto-segmentation systems using a common patient cohort.

Methods

Systems

This study evaluates four different DL-based auto-segmentation systems, which have been independently developed for commercial use. The systems evaluated were:

Varian AI Segmentation v2.0, Varian Medical Systems Inc., Palo Alto, CA, USA

Limbus Contour v1.4.1, Limbus AI, Regina, Saskatchewan, Canada

Siemens AI-Rad Companion Organs RT VA30, Siemens Healthcare GmbH, Erlangen, Germany

Mirada DLC Expert^{TM} v2.6.0, Mirada Medical Ltd, Oxford, UK

All systems were running their latest available software version at the time of study (June 2021 – Sep 2021). For the purposes of this research, systems have not been identified in the results and they will be known as Systems 1, 2, 3 and 4 (in no particular order). OARs outlined by the different systems are shown in table A1.

Table A1. Overview of H&N structures offered by systems

Structure	**Limbus Contour**	**Mirada DLC Expert**	**Varian AI Segmentation**	**Siemens Organs RT**
Carotid L+R		✓		
Brachial Plexus L+R	✓	✓		✓
Brain	✓	✓	✓	✓
Brainstem	✓	✓	✓	✓
Buccal Musosa L+R		✓		
Chiasm	✓	✓	✓	✓
Cerebellum		✓		
Cerebrum		✓		
Cervical Oesophagus		✓		
Cochlea L+R	✓	✓	✓	
Cricopharyngeal Inlet		✓		
Glottic area		✓		✓
IAM L+R		✓		
Lacrimal L+R	✓	✓	✓	
Larynx	✓	✓		?
Lens L+R	✓	✓	✓	✓
Lips	✓	✓	✓	✓
Lymph Nodes Neck L+R	✓			
Mandible	✓	✓	✓	✓
Oesophagus	✓		✓	✓
Optic Nerve L+R	✓	✓	✓	✓
Oral Cavity	✓	✓	✓	✓
Orbit L+R	✓	✓	✓	✓
Parotid L+R	✓	✓	✓	✓
Pharynx Constrictors	✓	✓		
Pituitary		✓		
Spinal Canal		✓	✓	
Spinal Cord	✓	✓	✓	✓
Submandibular L+R	✓	✓	✓	✓
Supraglottic		✓		
Thyroid	✓	✓		
Trachea	✓	✓	✓	

Contour selection

Five commonly used H&N OARs were utilised for the study, namely brainstem, spinal cord, mandible, left and right parotid. These OARs were chosen because they were used for the vast majority of H&N VMAT plans at the author's Institution, and also because they present varying levels of difficulty for auto-segmentation systems due to their anatomical location and their composition. For example, parotid glands are soft tissue and can vary significantly in size and shape between patients, but the mandible is a bony structure which, in theory, should be more straightforward to auto-segment due to the higher electron density compared to the surrounding soft tissues, and the reduced variation in size and shape.

Patient Selection

Fifty anonymised patient data sets were used for this retrospective study. All data originated from patients who were enrolled in the PATHOS clinical trial (Owadally et al., 2015).

This patient cohort was selected because the associated trial protocol included clear anatomical guidelines for OAR delineation and, in addition, trial entry involved pre-trial OAR outlining Quality Assurance, which all Oncologists were required to undertake. A review carried out by Vinod et al. (2016) showed that guidelines and teaching can significantly reduce inter-observer variability. A sample of patient data was retrospectively reviewed during the study to provide further assurance around the quality of contours used. For the purpose of this research the contours were deemed to be of 'gold standard' when comparing to automatically generated contours.

Comparison Metrics

After reviewing the literature to identify commonly used metrics, and considering locally available tools, the similarity metrics of 3D Dice Similarity Coefficient (DSC) (Dice, 1945), Added Path Length (APL) (Vaassen et al., 2020) and 2D 95% Hausdorff Distance (HD) (Huttenlocher et al., 2021) were utilised for the study.

3D DSC is a measure used to indicate the spatial overlap between two delineations, yielding a value of 1 in case of perfect overlap, and a value of 0 if no overlap. It is the most commonly used metric in structure comparison studies.

Studies have shown that values obtained for DSC when investigating inter-observer variability are in the region of 0.8 for brainstem, spinal cord and parotid structures and 0.9 for mandible (Stelmes et al., 2021 and Nelms et al., 2012).

The following formula is used to calculate 3D DSC:

$$\left(\frac{2\;X\;TP}{2\;X\;TP + FP + FN} \right)$$

Where the True Positive Volume (TP) is defined as the volume correctly identified as being present in both the reference and test structures.
The False Positive Volume (FP) is defined as the volume incorrectly identified by the test structure as being present in the reference structure.
The False Negative Volume (FN) is defined as the volume incorrectly identified by the test structure as not being present in the reference structure
Figure A1 provides an illustration of these volume.

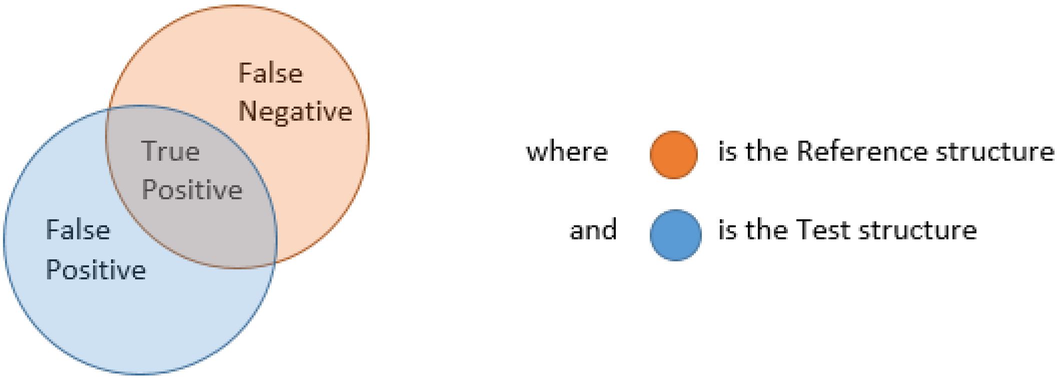
False Negative
True Positive
False Positive
where ● is the Reference structure
and ● is the Test structure

Fig. A1. Illustration of regions used to calculate 3D DSC

APL is the length of structure that must be drawn to correct the difference between Reference and Test structures. APL is calculated slice by slice in 2D and is then summed over all slices. It is important to note that the length of structure to be deleted is not included in this correction. APL uses a pre-defined tolerance value, e.g. 1mm, below which structures are considered to be identical.

APL was chosen for this study because it is a relatively new metric that can be used as a surrogate for time saving when compared to manual delineation. For example Vaassen et al. (2020) compared multiple evaluation measures and found that APL offered the highest correlation with absolute adaptation time of all metrics.
Typical values for APL vary considerably between structures due to this being an absolute distance measurement, the magnitude of which will therefore normally be greater for larger structures. A literature search conducted in early 2022 could not identify any other publications which made use of APL for H&N OARs.

The HD is a different type of metric which evaluates the distance between boundaries. It is defined as the greatest shortest distance between two structures, calculated in 2D for each CT slice. Menze at al. (2015) recommend use of the more robust 95% HD for structure comparison, due to the susceptibility of the HD metric to small outlying sub-regions. The 2D 95% HD is the 95th percentile of the distance and is calculated using the whole volume.
Typical values of 2D 95% HD when investigating inter-observer variability are 4mm for brainstem, 3mm for spinal cord and 6mm for parotids (Wong et al., 2020).

All comparison metrics were generated using the Mirada Medical Ltd Contour Insights^{TM} Tool.

Statistical Analysis

A Shapiro-Wilk test was used to test for normality, and differences between systems were assessed using a two-sided, paired Wilcoxon signed rank test, with a significance level a = 0.05 and a power of 90%.

Results

Results of segmentations for all structures and systems are presented in tables A2 to A4. In each table the highest performing system for each structure is highlighted in bold.

The Wilcoxon signed rank test showed that differences between the highest and second highest performing systems were rarely statistically significant, with the only exception being the mandible structure, where system 1 demonstrated superior performance.

Firstly, considering the 3D DSC metric (table A2), it can be observed that the highest performing system varies depending on the OAR being evaluated. For example, system 2 produces superior results for brainstem and parotid, but systems 1 and 3 produce the best results for the mandible and spinal cord structures respectively. System 4 performs less well than the other systems for the mandible OAR and systems 1 and 4 demonstrate inferior performance for the spinal cord OAR when compared to systems 2 and 3.

Table A2. Mean 3D DSC for all OARs and all systems

OAR	Mean 3D DSC			
	System 1	System 2	System 3	System 4
Brainstem	0.811 ± 0.063	**0.821 ± 0.052**	0.799 ± 0.051	0.768 ± 0.055
Left Parotid	0.760 ± 0.062	**0.791 ± 0.062**	0.783 ± 0.063	0.787 ± 0.066
Right Parotid	0.744 ± 0.079	**0.787 ± 0.065**	0.775 ± 0.073	0.784 ± 0.076
Mandible	**0.905 ± 0.018**	0.885 ± 0.023	0.886 ± 0.023	0.830 ± 0.030
Spinal Cord	0.703 ± 0.069	0.789 ± 0.072	**0.799 ± 0.071**	0.745 ± 0.065

Table A3 shows results for the APL metric. APL results broadly agree with DSC results in terms of the relative system performance, although a different system was identified as the highest performing for the right parotid structure. It was also observed that the APL did not always identify relatively large differences between structures, highlighting a potential deficiency in the application of this metric. APL values for mandible and spinal cord are, on average, significantly larger than brainstem and parotid values.

Table A3. Mean APL for all OARs and all systems

OAR	Mean APL (mm)			
	System 1	System 2	System 3	System 4
Brainstem	1077 ± 253	**1054 ± 294**	1163 ± 322	1197 ± 312
Left Parotid	917 ± 250	**859 ± 233**	886 ± 211	866 ± 204
Right Parotid	923 ± 266	888 ± 255	**872 ± 255**	899 ± 248
Mandible	**1195 ± 403**	1896 ± 815	1676 ± 804	1688 ± 352
Spinal Cord	1554 ± 699	1346 ± 800	**1301 ± 877**	1377 ± 508

Comparison results using the 2D 95% HD metric (table A4) correlated well with the 3D DSC data, with system 2 being the highest performer for brainstem and parotid structures, and systems 1 and 3 highest performing for mandible and spinal cord respectively. Values for mandible were particularly large for system 4.

Table A4. Mean 2D 95% HD for all OARs and all systems

OAR	Mean 2D 95% HD (mm)			
	System 1	System 2	System 3	System 4
Brainstem	5.30 ± 2.11	**4.39 ± 1.23**	4.84 ± 1.32	5.80 ± 1.73
Left Parotid	10.34 ± 4.70	**8.78 ± 3.93**	9.02 ± 3.44	9.04 ± 4.41
Right Parotid	11.37 ± 5.48	**7.64 ± 3.00**	9.23 ± 4.08	8.80 ± 3.98
Mandible	**5.38 ± 6.93**	7.75 ± 8.72	8.60 ± 9.58	23.91 ± 9.19
Spinal Cord	2.74 ± 0.76	2.73 ± 1.19	**2.59 ± 1.22**	2.87 ± 0.64

Boxplots illustrating the relative spread of results for all metrics, systems and structures are shown in figures A2 to A4.

From figure A2 it can be observed that all systems show a similar spread of values for each OAR for the metric 3D DSC.

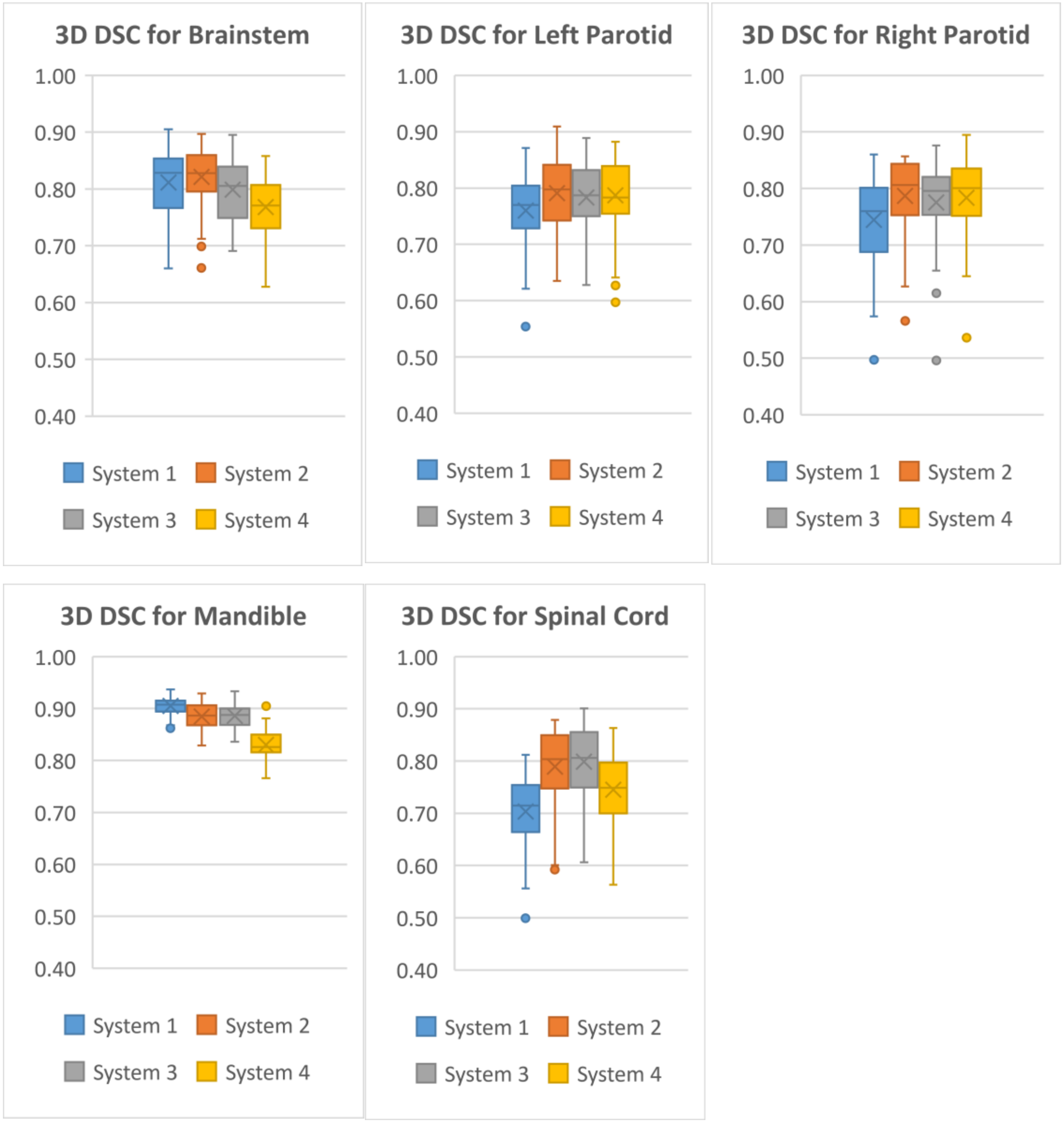

Fig. A2. 3D DSC for H&N structures across all systems

Boxplots for the APL metric (figure A3) are again statistically comparable between systems for the majority of OARs, however, for the mandible, systems 2 and 3 demonstrate a larger spread of results.

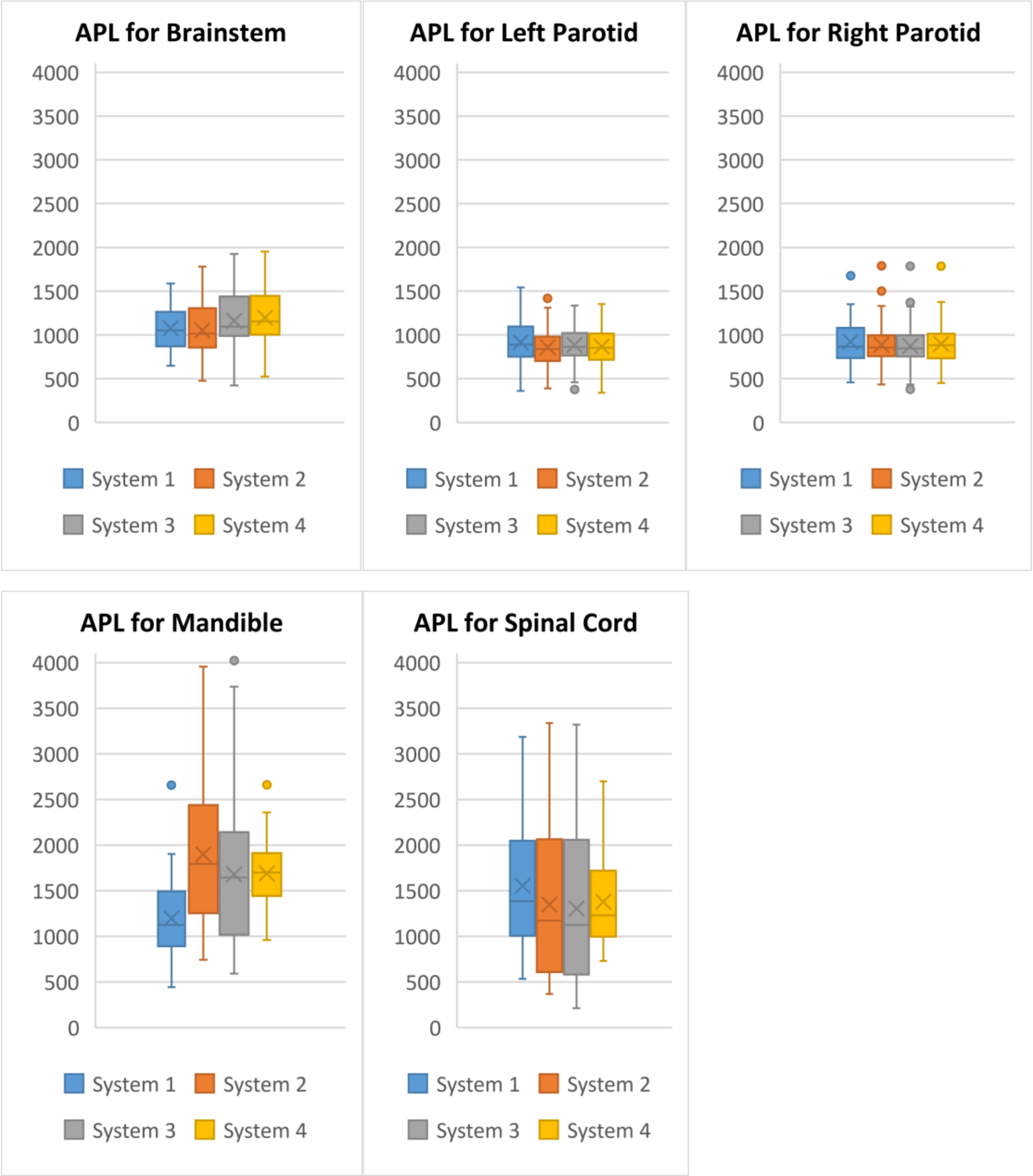

Fig. A3. APL (mm) for H&N structures across all systems

For 2D 95% HD, results are statistically similar with the exception of the system 4 mandible results which are of a statistically significant larger magnitude than other systems.

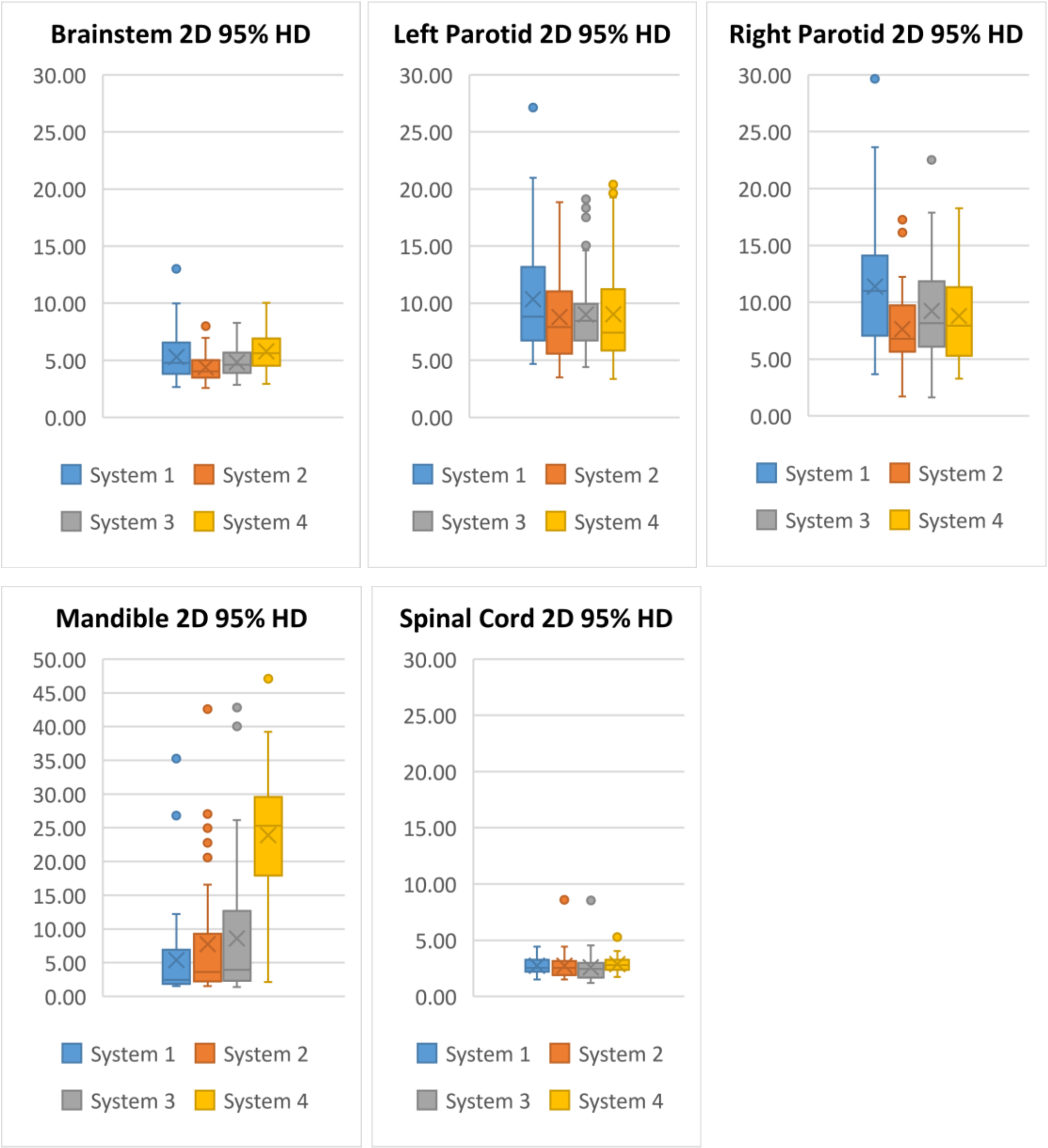

Fig. A4. 2D 95% HD (mm) for H&N structures across all systems

Discussion

The aim of this study was to compare the performance of multiple DL-based auto-segmentation systems for H&N OARs. Other studies have already demonstrated that DL-based systems are superior to atlas-based systems (Choi et al., 2020), but no study has, to date, assessed the variation in quality of different DL-based systems with a common patient cohort.

The results obtained indicate that no single system is significantly superior to other systems for the OARs investigated, and that different systems performed more effectively for different OARs. It is difficult to determine the reasons for variations in system performance for individual OARs due to the nature of DL. The DL algorithm employed and the patient cohort utilised for model training will both affect the resulting contours. This information is not usually readily available to the end user of these algorithms.

Values obtained for 3D DSC compare favourably with values obtained by other researchers when investigating inter-observer variability, indicating that DL systems are now achieving human expert levels of performance. For example, when Stelmes et al. (2021) compared gold standard expert delineations against volumes produced by expert humans in other institutions for mandible and parotids, they observed 3D DSC values of 0.81 and 0.88 respectively. In this study, mean values for parotid and mandible for the highest performing systems were 0.79 and 0.91 respectively.

2D 95% HD for parotids are larger than expected when comparing to the literature. For example Wong et al. (2020) found that the average difference between expert contouring and Deep Learning contouring was 6mm. This discrepancy with existing research requires further investigation.

A further observation from the results is that 3D DSC is high for mandible, indicating good agreement with clinical contours, and yet APL is large, indicating significant change. This can be explained because APL values will usually be larger for larger structures, as contours will need to be edited on a greater number of CT slices. If the APL was divided by the reference contour length to produce a relative APL metric this should in theory be more comparable to other metrics such as 3D DSC, and will provide an indication of the proportion of the contour that requires editing.

It is also important to note that commercial AI auto-segmentation systems are updated frequently, with model improvements and new structures being introduced several times a year. This rate of change is because the systems are still in the relatively early stages of development. This rapid rate of development may therefore mean that the current highest performing commercial system will vary over time, and any comparison study is likely to remain valid for a short period of time.

The clinical OARs used for this study, although of a high quality due to their clinical trial inclusion and associated peer review, will still be prone to human error and inter-observer variability, as shown by Nelms et al. (2012), Stelmes et al. (2021) and van der Veen et al. (2021). There may still therefore be scope for further improvement in the quality of structures used as the gold standard. This could be achieved by introducing multiple observers to inspect the structure, and a further extension to this could be the use of STAPLE contours (Warfield, Zou and Wells, 2004).

3D DSC is commonly used for structure comparison, and although it provides a good indication of structure overlap, it has a low sensitivity for complex boundaries (Sherer at al. 2021). It does, however, allow the results obtained in this study to be considered within the context of already published studies.

APL was developed as a proxy for how much time will be required to adjust contours after auto-segmentation, and not to give an absolute comparison between two structures (Vaassen et al., 2020). One consequence of this is that if this metric is being used to analyse geometric differences between structures, there may be a large difference in the resulting value if the reference and test structures are interchanged. This is because the APL ignores contours that need to be deleted from CT slices. One suggestion for adapting APL to provide a more appropriate metric for structure comparison would be to interchange the reference and test structures, and use the greater of the two values as the ‘difference metric’.

2D 95% HD is a spatial distance-based metric and, although it is sensitive to point positions, it is not able to account for the proportion of a contour that needs to be edited. It may also not be an appropriate metric to use when comparing large bifurcating structures such as the mandible where large differences may be observed on individual CT slices due to the presence of ‘islands’ of contour and minor differences between the superior extent of structures being compared.

A metric that was not utilised in the study, and which may offer advantages over those used, is the Surface DSC (Nikolov et al., 2021). This metric assesses the overlap of two surfaces rather than the overlap of two volumes as in volumetric DSC, and gives an indication of the agreement between the surfaces of two structures. This metric is not yet widely available and was not available locally to use for this study.

Regarding the assessment of differences between structures, although some differences were found to be statistically significant, no further analysis was performed to attempt to determine if differences were clinically significant. Current practice involves manual inspection and modification by a trained operator to ensure all contours are clinically acceptable for each individual patient. If deep-learning derived contours were to be used automatically without modification, further research would be warranted to investigate the potential impact of clinically significant differences between contours depending on their derivation.

One further comment is that it is important to check the anatomical definitions used by the different systems. It may be that definitions used by a particular system conform more closely to local practice, and this should therefore be a further consideration when choosing the most appropriate product. This highlights the importance of the use of a single set of international consensus guidelines for OAR delineation (Mir, R. et al., 2020).

Conclusion

In summary, comparable levels of performance were observed between all four systems. This indicates that DL-based auto-segmentation products are developing at a similar pace in terms of the quality of contours produced.

It may therefore be more appropriate to consider other factors such as financial cost, ease and speed of use, and range of OARs offered, when considering AI auto-segmentation systems for clinical use.

Author contribution statement

Simon Temple contributed to study design, data collection, analysis and interpretation, and drafted the manuscript.

Carl Rowbottom contributed to study design, data collection and critical revision of the manuscript.

Acknowledgements

I would like to thank Mirada Medical Ltd for providing the Contour Insights^{TM} Tool, which was utilised to produce structure comparison metrics.

I would also like to thank Varian Medical Systems Inc., Limbus AI, Siemens Healthcare GmbH and Mirada Medical Ltd for providing access to the AI auto-segmentation systems evaluated in this study.

References

‘Artificial intelligence’ (2019).Wikipedia. Available at: https://en.wikipedia.org/wiki/Artificial_intelligence (Accessed: 13 May 2019).

Bini, S.A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? Journal of Arthroplasty, 33(8), pp.2358–2361. [online]. Available from: https://doi.org/10.1016/j.arth.2018.02.067.

Brunenberg, E.J.L. et al. (2020). External validation of deep learning-based contouring of head and neck organs at risk. Physics and Imaging in Radiation Oncology, 15(June), pp.8–15. [online]. Available from: https://doi.org/10.1016/j.phro.2020.06.006.

Choi, M.S. et al. (2020). Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiotherapy and Oncology, 153, pp.139–145. [online]. Available from: https://doi.org/10.1016/j.radonc.2020.09.045.

Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology (Durham), 26(3), pp.297–302.

Fritscher, K.D., Grünerbl, A. and Schubert, R. (2007). 3D image segmentation using combined shape-intensity prior models. International Journal of Computer Assisted Radiology and Surgery, 1(6), pp.341–350.

Fritscher, K.D. et al. (2014). Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Medical Physics, 41(5), pp.1–11.

Huttenlocher, D.P., Rucklidge, W.J. and Klanderman, G.A. (1992). Comparing images using the Hausdorff distance under translation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992-June(9), pp.654–656.

Lustberg, T. et al. (2018). Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiotherapy and Oncology, 126(2), pp.312–317. [online]. Available from: https://doi.org/10.1016/j.radonc.2017.11.012.

Mir, R. et al. (2020). Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines: GHG OAR consensus contouring guidance. Radiotherapy and Oncology, 150, pp.30–39. [online]. Available from: https://doi.org/10.1016/j.radonc.2020.05.038.

Menze, B.H. et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging, 34(10), pp.1993–2024.

Nelms, B.E. et al. (2012). Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer. International Journal of Radiation Oncology Biology Physics, 82(1), pp.368–378.

Nikolov, S. et al. (2021). Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. Journal of Medical Internet Research, 23(7).

Owadally, W. et al. (2015). PATHOS: A phase II/III trial of risk-stratified, reduced intensity adjuvant treatment in patients undergoing transoral surgery for Human papillomavirus (HPV) positive oropharyngeal cancer. *BMC Cancer*, 15(1), pp.1–10.

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. Neural Networks, 61, pp.85–117. [online]. Available from: http://dx.doi.org/10.1016/j.neunet.2014.09.003.

Sherer, M. V. et al. (2021). Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiotherapy and Oncology*, 160, pp.185–191. [online]. Available from: https://doi.org/10.1016/j.radonc.2021.05.003.

Sims, R. et al. (2009). A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiotherapy and Oncology, 93(3), pp.474–478. [online]. Available from: http://dx.doi.org/10.1016/j.radonc.2009.08.013.

Stelmes, J.J. et al. (2021). Quality assurance of radiotherapy in the ongoing EORTC 1420 "Best of" trial for early stage oropharyngeal, supraglottic and hypopharyngeal carcinoma: results of the benchmark case procedure. Radiation Oncology, 16(1), pp.1–10.

Thomson, D. et al. (2014). Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiation Oncology, 9(1), pp.1–12.

Teguh, D.N. et al. (2011). 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, 81(4), pp.950–957.

Vaassen, F. et al. (2020). Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology, 13(November 2019), pp.1–6. [online]. Available from: https://doi.org/10.1016/j.phro.2019.12.001.

van der Veen, J. and Nuyts, S. (2017). Can intensity-modulated-radiotherapy reduce toxicity in head and neck squamous cell carcinoma? *Cancers*, 9(10).

van der Veen, J. et al. (2021). Interobserver variability in organ at risk delineation in head and neck cancer. Radiation Oncology, 16(1), pp.1–11. [online]. Available from: https://doi.org/10.1186/s13014-020-01677-2.

van Dijk, L. V. et al. (2020). Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. Radiotherapy and Oncology, 142, pp.115–123. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.09.022.

Warfield, S.K., Zou, K.H. and Wells, W.M. (2004). Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging, 23(7), pp.903–921.

4. Study B: Failure rates and Quality Assurance of commercial AI auto-segmentation systems for head and neck cancer

Simon Temple, Carl Rowbottom
The Clatterbridge Cancer Centre NHS Foundation Trust, Liverpool, UK.

Abstract

Purpose/Objective

AI-based commercial software can be used to automatically delineate organs at risk (OAR) on CT scans, with the potential for significant efficiency savings in the radiotherapy treatment planning pathway, and simultaneous reduction of inter- and intra-observer variability.

In order to be able to design a suitable Quality Assurance (QA) program for such systems it is important to have a good understanding of expected failure rates, and the reasons for these failures. There has not yet been research looking at the failure rates of such systems, which has been addressed in this study.

Materials/Methods

Previous research (thesis study A), where 50 anonymised head and neck (H&N) patient data sets with 'gold standard' contours were compared to AI auto-segmented contours, was used to produce expected mean values and standard deviation data for the similarity metric of Dice Similarity Coefficient (DSC) for four commonly used H&N OARs (brainstem, mandible, left and right parotid).

A commercial AI auto-segmentation system was then used to generate OARs on 500 anonymised patient datasets. Auto-segmented contours were compared to existing clinical contours, which had been outlined by an expert human, and a failure rate was set at three standard deviations below the expected mean DSC.

All failures were inspected to assess the reasons for failure and to determine if they were 'true failures' of the auto-segmentation system.

Failures were classified into one of five groups (setup position, anatomical, image artefacts, suboptimal clinical contour and unknown), and failures relating to

suboptimal contouring of the original clinical structure were removed, to produce a ‘true failure’ rate for each OAR.

Final true failure rates were used to inform recommendations for system QA.

Results

True failure rates for the four OARs investigated were 0.4% for brainstem, 2.2% for mandible, 1.4% for left parotid and 0.8% for right parotid.

For brainstem there were a total of 2 true failures from the 500 assessed, with both CT datasets showing a non-standard patient setup position as the likely cause of failure.

For mandible there were a total of 11 true failures, with 8 of these showing non-standard anatomy and 3 showing dental artefacts as likely cause of failure.

For left parotid there were a total of 7 true failures, comprising of 5 failures due to unusual anatomy and 2 for unknown reasons.

For right parotid there were a total of 4 true failures, 1 of these appeared to be caused by an unusual patient setup position, 2 due to unusual patient anatomy and 1 due to dental artefacts.

Conclusion

Where true failures of the auto-segmentation system were identified, there was almost always an evident non-standard element associated with the planning CT dataset, for example unusual setup position, unusual anatomy or the presence of dental artefacts.

It can therefore be hypothesised that these non-standard elements were the reason for the failure, and it can be further suggested that the patient datasets used to train the DL model did not contain sufficient heterogeneity of patient data.

Regardless of the reasons for failure, the true failure rate for AI auto-segmentation systems in the H&N region for the OARs investigated is extremely low (approx. 1%). Due to this very low failure rate, it is advised that QA of resulting auto-segmented OARs should utilise automated methods, because human inspection alone is unlikely to be effective in identifying failures that occur at such low rates.

Introduction

The accurate delineation of organs at risk (OAR) for modern radiotherapy treatment planning is very important (Tong et al., 2018). When performed manually this process can be extremely time-consuming (Cardenas et al., 2019), and is also prone to inter and intra-observer variability (Nelms et al., 2012). The availability of any system which is able to automatically and accurately perform this function is therefore likely to produce substantial quality and efficiency savings.

Early solutions to auto-segmentation used atlas-based methods (Daisne and Blumhofer, 2013), but the use of artificial intelligence (AI) based software for auto-segmentation of OARs has become increasingly common in recent years. Vrtovec et al. (2020) carried out a review of auto-segmentation literature from 2008 to 2020 and demonstrated that a shift from atlas-based methods to AI-based methods started around 2016.

More specifically, deep learning (DL), which is a subset of AI, forms the basis for these new auto-segmentation techniques, and Cardenas et al. (2019) suggested that use of this new technology means we have now entered the fourth generation of auto-segmentation algorithm development. In 2022 Harrison et al. reviewed auto-segmentation techniques used in radiotherapy treatment planning and concluded DL methods have the potential to transform the radiation oncology workflow by increasing efficiency and removing inter and intra-observer variability.

Quality Assurance (QA) of any software system that is used in the radiotherapy treatment planning process is extremely important due to the complexity of systems, and the report from the American Association of Physicists in Medicine (AAPM) Radiation Therapy Committee Task Group 53 (Fraass et al., 1998) recommends that all systems used for treatment planning have an appropriate QA programme.

Systems that utilise DL are often referred to as ‘black box’, because it is not possible for users to understand their internal function and therefore it is not possible to predict their behaviour. Wong et al. (2020) suggest that this means there is a need for robust studies to evaluate performance before such systems are used clinically.

Rudin (2019) describe the concept of '*Explainable Machine Learning*' and suggest that it is often possible to use interpretable black box models, but to date this approach has not been utilised for AI auto-segmentation. Poon et al. (2021) also discuss the use of black box AI in medicine, and suggest that interpretability is a requirement to gain trust and acceptance of AI in medicine from physicians.

Cardenas et al. (2019) stress the importance of auto-segmentation system QA due to the potentially serious consequences of segmentation errors, and Vandewinckele et al. (2020) advise that both case-specific and routine model QA is carried out on such systems.

For all the aforementioned reasons it is therefore important that a robust QA programme is in place when using AI auto-segmentation systems clinically.

The aim of this study was to identify the failure rate of a mature commercial DL-based auto-segmentation system using head and neck (H&N) OARs in order to determine suitable QA requirements. To the best of our knowledge, this is the first study to look at such failure rates.

Methods

In order to be able to define a 'failure' it is important to understand expected behaviour, and for auto-segmentation this expected behaviour can be quantified using similarity metrics.

Previous research (Study A), where 50 anonymised H&N patient data sets with 'gold standard' contours were compared to AI auto-segmented contours, was therefore used to produce expected mean values and standard deviation data for the similarity metric of Dice Similarity Coefficient (DSC) (Dice, 1945) for four commonly used H&N OARs (brainstem, mandible, left and right parotid).

3D DSC is a measure used to indicate the spatial overlap between two delineations, yielding a value of 1 in case of perfect overlap, and a value of 0 if no overlap. It is the most commonly used metric in structure comparison studies. Studies have shown that values obtained for DSC when investigating inter-observer variability are in the region of 0.8 for brainstem, spinal cord and parotid structures and 0.9 for mandible (Stelmes et al., 2021 and Nelms et al., 2012).

The following formula is used to calculate 3D DSC:

$$\left(\frac{2\ X\ TP}{2\ X\ TP + FP + FN} \right)$$

Where the True Positive Volume (TP) is defined as the volume correctly identified as being present in both the reference and test structures.

The False Positive Volume (FP) is defined as the volume incorrectly identified by the test structure as being present in the reference structure.

The False Negative Volume (FN) is defined as the volume incorrectly identified by the test structure as not being present in the reference structure

Figure B1 provides an illustration of these volume.

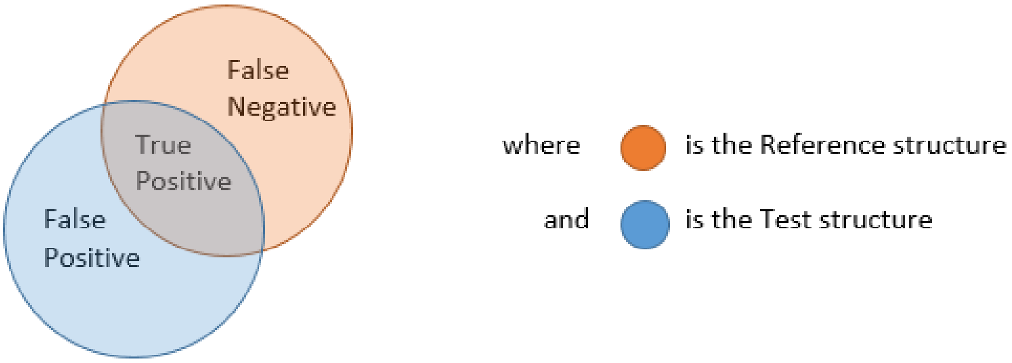
False Negative
True Positive
False Positive
where ● is the Reference structure
and ● is the Test structure

Fig. B1. Illustration of regions used to calculate DSC

A commercial AI auto-segmentation system, Mirada DLC Expert^{TM}, Mirada Medical Systems Ltd, was then used to generate OARs on 500 anonymised patient CT data sets for the four commonly used H&N OARs (brainstem, mandible, left and right parotid). A data set of this size was determined to be necessary due to the absence of any existing research in this area and the need to identify a sufficiently accurate failure rate.

The 500 data sets also contained contours that had been previously used clinically, and the auto-segmented contours were then compared to the clinical contours using the Mirada Contour Insights™ tool to produce a DSC for each patient.

The aim of this research was to identify gross failures, and therefore a three sigma limit was set to determine the failure rate (Pukelsheim, 1994), meaning that 99.7% of results can be assumed to be within this limit. All failures for each OAR were manually inspected by an expert observer, and reasons for failure were categorised as shown in table B1.

Table B1. Categorisation of failures

Category	Description
Setup	Non-standard CT scan patient setup position
Anatomy	Non-standard internal patient anatomy e.g. post-surgery
Artefacts	CT artefacts present in region due to e.g. dental fillings
Clinical	A suboptimal clinical contour has been used for comparison
Unknown	No obvious reason for failure

A ‘true failure’ rate for each OAR was then calculated by removing failures that were due to suboptimal clinical contouring.

Results

Figures B2 to B5 show DSC values for the comparison between AI auto-segmented and manually delineated OARs for 500 patients. The failure level is set at three standard deviations below the mean expected 3D DSC value.

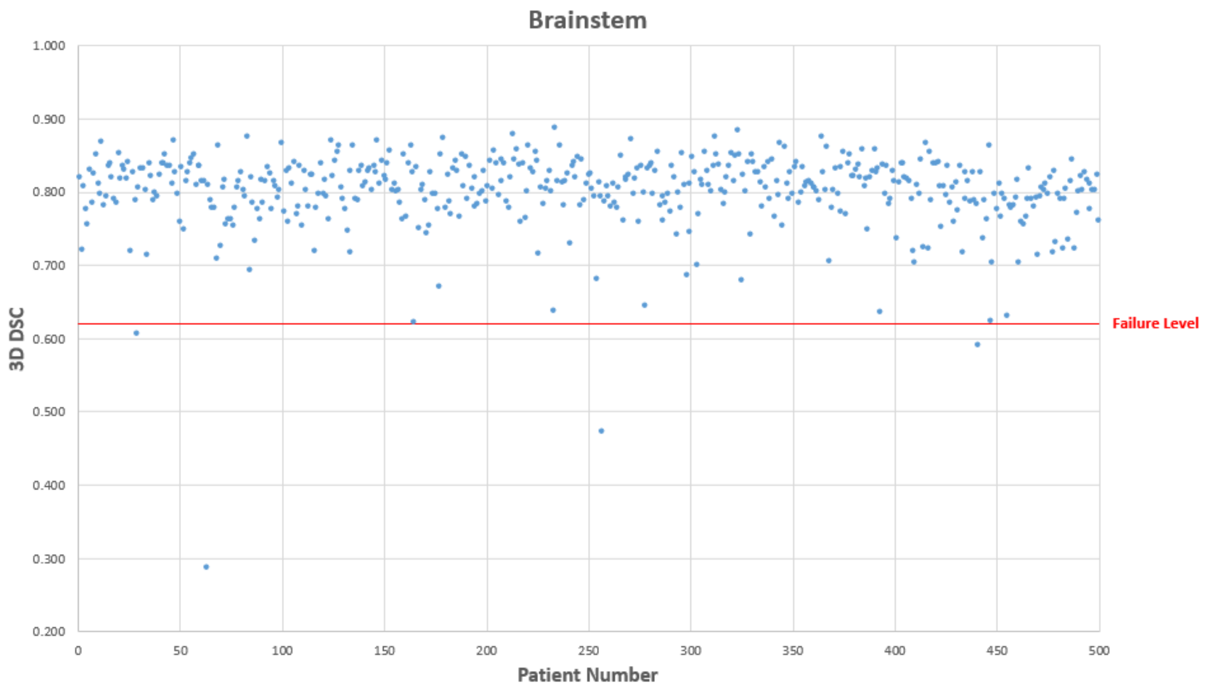

Fig. B2. 3D DSC for the comparison between AI auto-segmented and manually delineated brainstem OAR

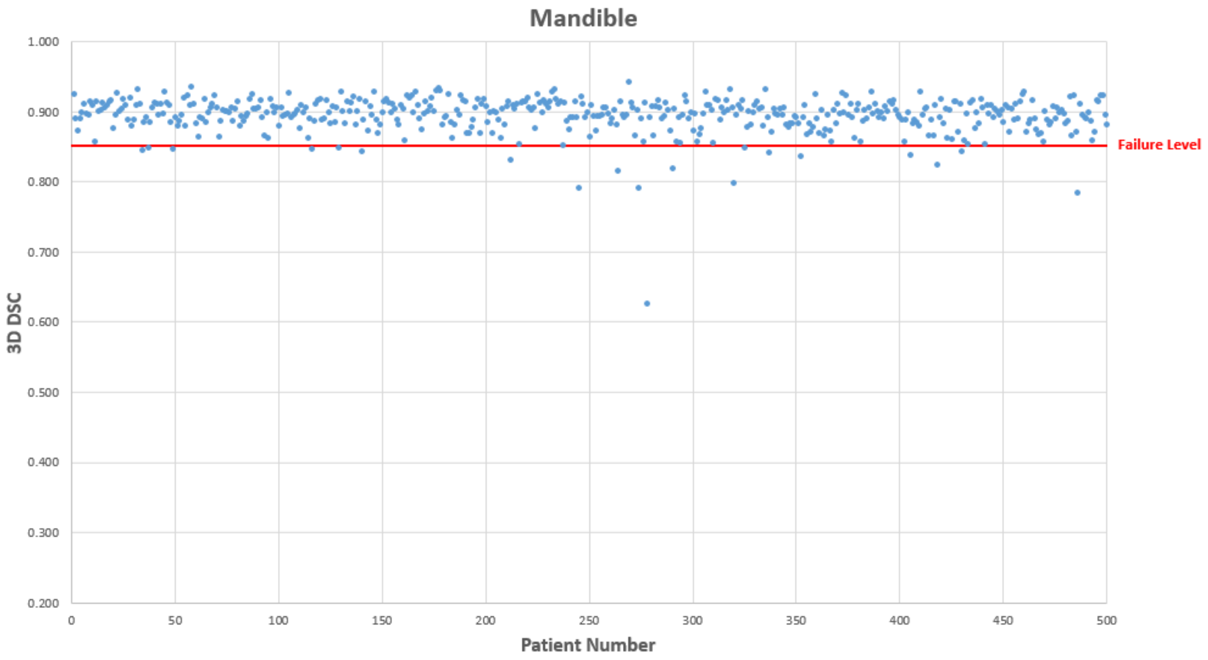

Fig. B3. 3D DSC for the comparison between AI auto-segmented and manually delineated mandible OAR

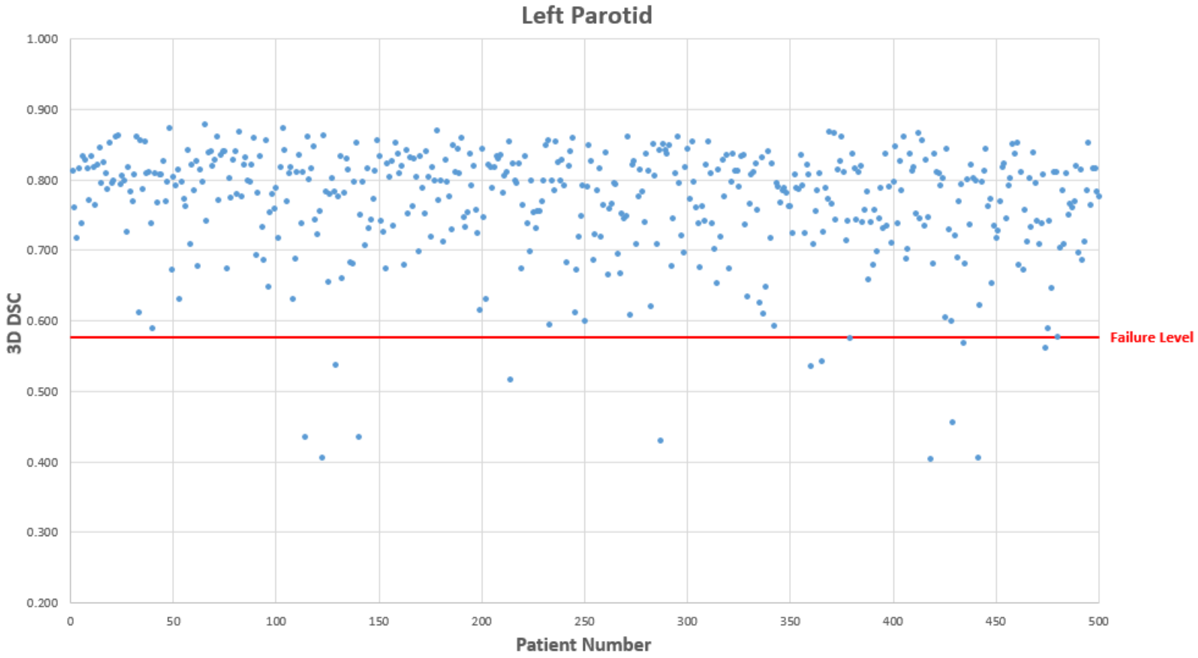

Fig. B4. 3D DSC for the comparison between AI auto-segmented and manually delineated left parotid OAR

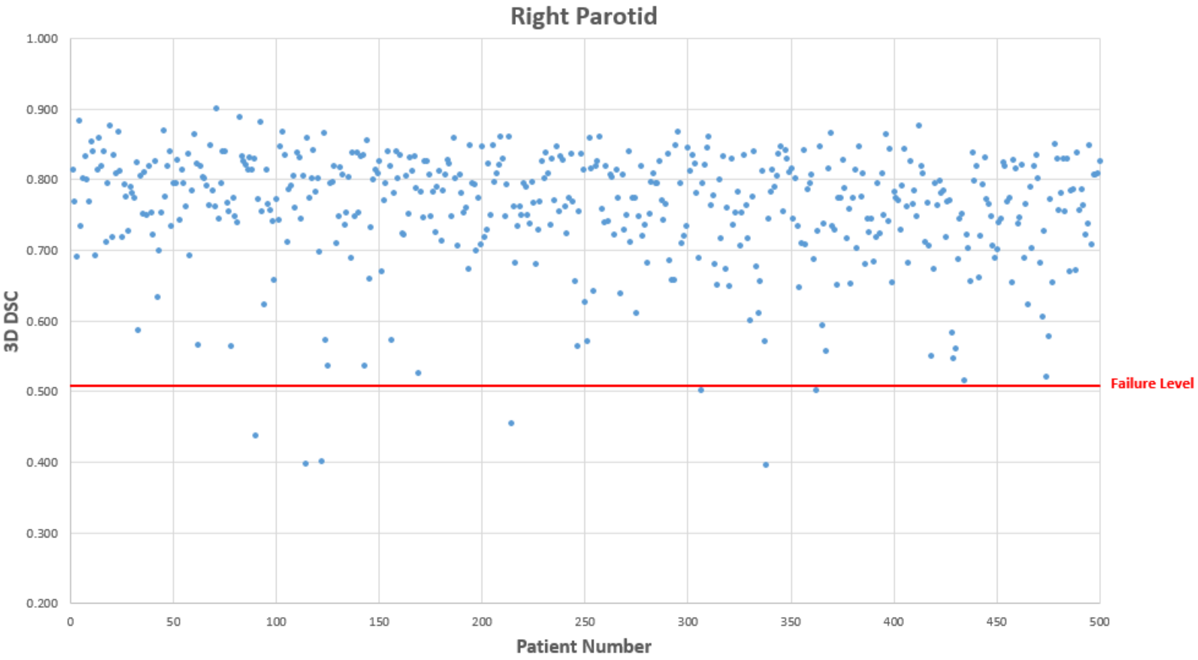

Fig. B5. 3D DSC for the comparison between AI auto-segmented and manually delineated right parotid OAR

Resulting failure rates are shown in table B2. For the brainstem there were 4 total failures, reducing to 2 true failures and a true failure rate of 0.4% following removal of 2 failures caused by suboptimal clinical contouring.

The mandible structure had 20 total failures, reducing to 11 true failures after removal of those with a suboptimal clinical contour. Of these, 8 were due to unusual patient anatomy and 3 appeared to be caused by dental artefacts.

For the left parotid there were 13 total failures, reducing to 7 true failures. Reasons for failure were determined to be caused by unusual patient anatomy for 5 patients, and for the 2 remaining patients the failure reason could not be identified.

For the right parotid 7 total failures became 4 true failures after removal of those with a suboptimal clinical contour. One failure was determined to be caused by a non-standard patient setup position, 2 were due to unusual patient anatomy and 1 was caused by a dental artefact.

The overall failure rate for all OARs was less than 3% and the mean failure rate for the four OARs investigated was found to be 1.2%.

Table B2. AI auto-segmentation failure rates for 500 patients

	Brainstem	Mandible	Lt Parotid	Rt Parotid
DSC Mean	0.811	0.905	0.760	0.744
DSC Standard Deviation	0.063	0.018	0.062	0.079
DSC Failure Level (Mean – 3 x SD)	0.621	0.851	0.576	0.509
Total Failures	4	20	13	7
Failure Reason:				
Setup position	2	0	0	1
Anatomical	0	8	5	2
Dental artefacts	0	3	0	1
Clinical structure suboptimal	2	9	6	3
Unknown	0	0	2	0
True failures (Total – clnical error)	**2**	**11**	**7**	**4**
True failure rate	**0.4%**	**2.2%**	**1.4%**	**0.8%**

An example of a setup failure for the brainstem OAR is shown in figure B6. It can be observed that this patient had an obvious 'roll' in their setup position. When measured, the axial roll was found to be approximately 7°.

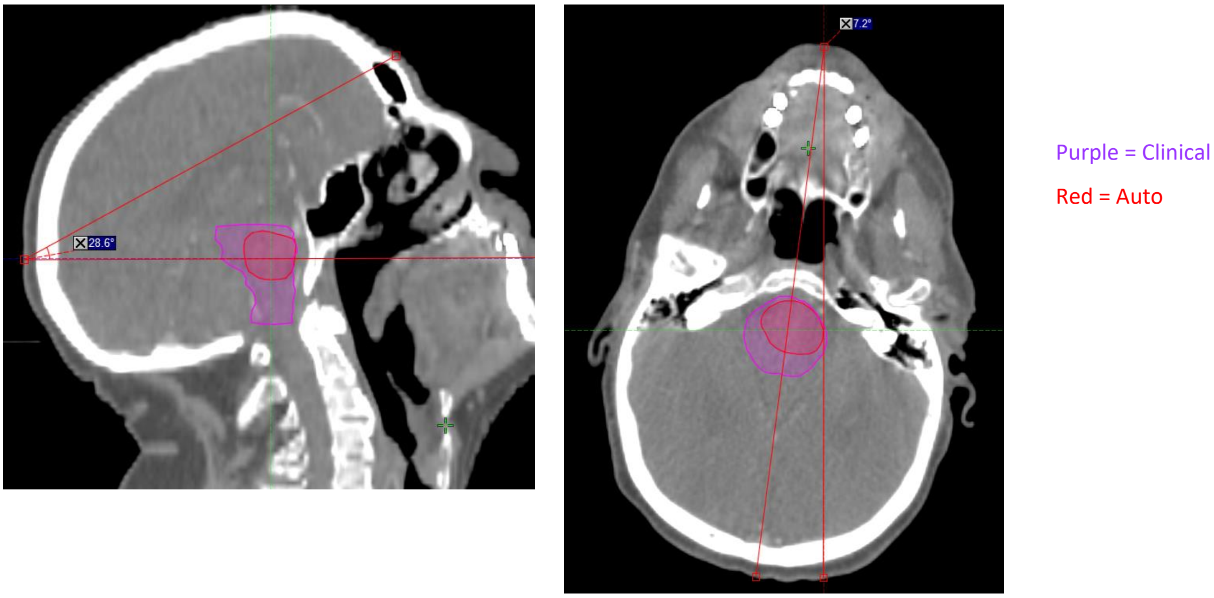

Fig. B6. Sagittal and axial CT images to illustrate example of a gross failure for auto-segmentation of brainstem

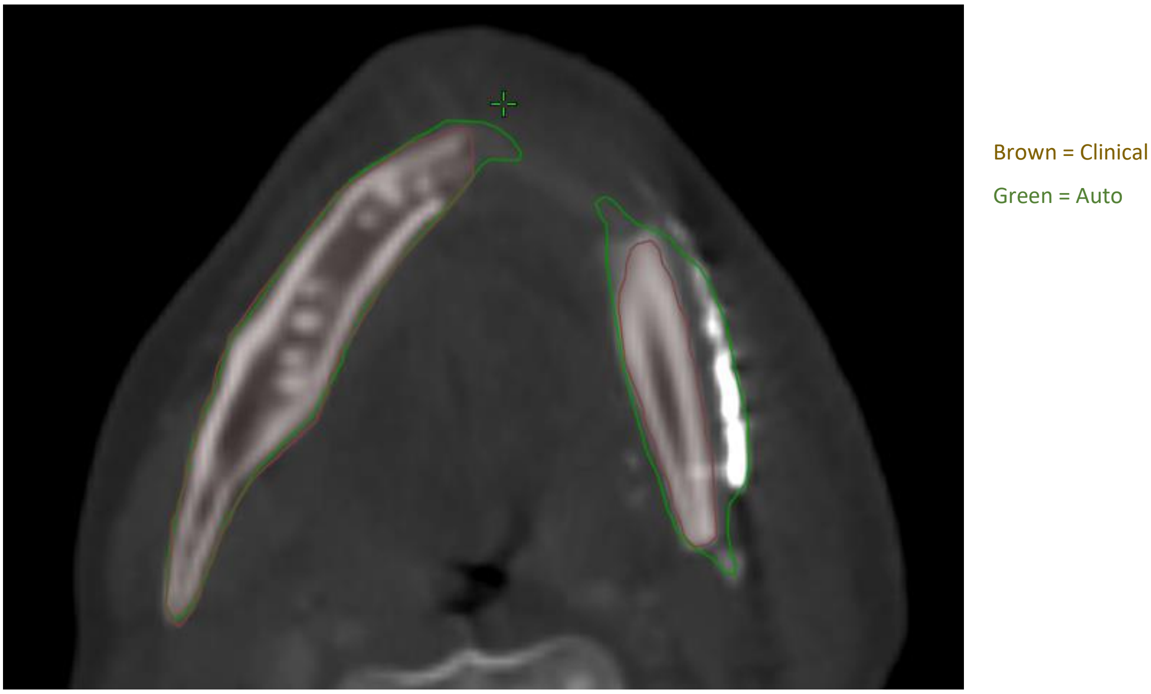

Fig. B7. Axial CT image to illustrate example of a gross failure for auto-segmentation of mandible

An example of anatomical failure for the mandible OAR is shown in figure B7. It can be observed that the auto-segmented mandible contour includes part of a surgical metal plate.

An example of failure for the right parotid OAR is shown in figure B8. It can be observed that the inferior extent of the auto-segmented parotid contour stops at the level where dental CT artefacts appear.

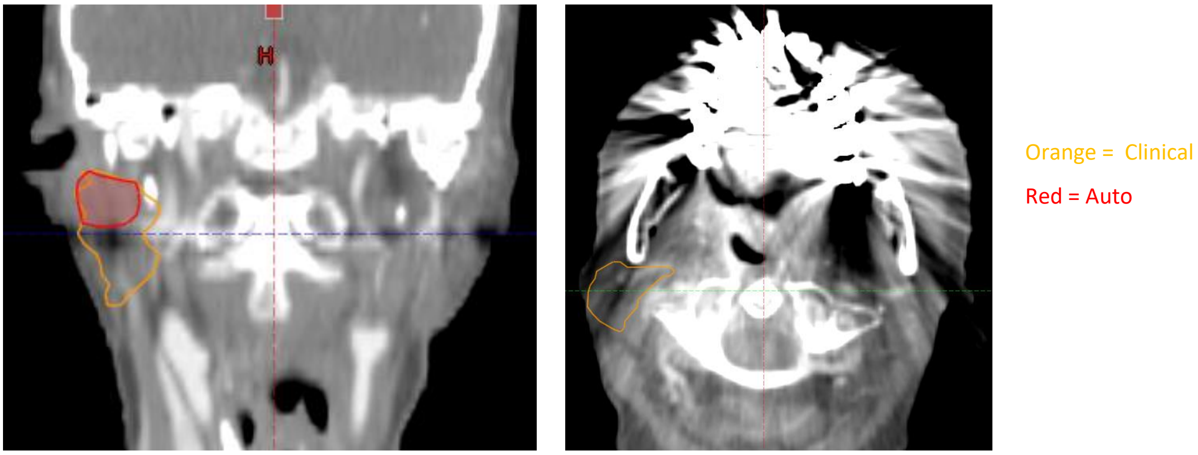

Fig. B8. Coronal and axial CT images to illustrate example of a gross failure for auto-segmentation of rt parotid

Discussion

The aim of this study was to assess gross failure rates of a commercial AI auto-segmentation system (Mirada DLC Expert^{TM}).

The study has shown that true gross failures of such systems are rare, being less than 3% for all OARs investigated and less than 1% for some OARs e.g. brainstem, with a mean failure rate of 1.2% for all OARs.

In addition, one of the reasons for failure, the presence of dental artefacts, is both easy to identify as a possible issue during review, and is less likely to apply to modern CT scanners, which utilise metal artefact reduction algorithms (Kovacs et al., 2018).

In terms of the other possible reasons for failure, which were patient setup position and non-standard anatomy, it can be hypothesised that such failures are caused by the absence of sufficient numbers of these ‘unusual’ case types in the model training data set.

This sort of dataset bias in AI is a well-known issue outside radiotherapy (Kusters et al., 2020) and is entirely predictable, given the relative frequency of such cases in the clinic. In addition, due to the observed wide anatomical variation in patients with non-standard anatomy, it is unlikely that it would be possible to include sufficient numbers of this patient type in a DL model training data set for the training to be effective due to the large patient dataset numbers typically required to train DL models (Fan et al., 2019).

In cases where a patient is in a non-standard position, for example they are rotated, or their neck is in an unusual state of flexion, it may be more realistic to suggest that this situation can be incorporated into a model. It is therefore suggested that this is something commercial model system developers should investigate in future iterations of this technology. It may also be beneficial for developers to build in post-processing checks to identify anomalies such as non-standard patient rotations or shape metrics of resulting OARs.

These findings highlight the need for manufacturers to ‘open up’ the black box nature of the DL software being produced. Gebru et al. (2021) have proposed that datasheets containing comprehensive information about the dataset used to produce a DL model should be provided, and such information could in theory be used by healthcare professionals to determine the limits of clinical use of such DL models.

Regarding variations in patient setup position, it can be hypothesised that if models have been trained using data from institutions that utilise particular forms of patient immobilisation, there may be an increased risk of failure when applied to CT scans that use alternative patient immobilisation approaches. This may be an effect particular to the H&N region, but further research is required to determine this.

Options regarding QA checks to identify patient datasets which are prone to these errors are to incorporate a check on ROI density. This could be used to identify the presence of artefacts or metalwork in the dataset and additionally could be used to identify unusual density values in segmented OARs. The latter is likely to be an OAR-specific check due to the varying electron density of different structures depending on whether they comprise bone, soft tissues etc.

A further QA check that could be incorporated is to flag datasets where the patient is set up with an unusual 'roll' or 'tilt'. Results have shown that such patients are more susceptible to AI auto-segmentation errors.

Regardless of the reasons for failure, due to the extremely low true failure rates identified, the ideal approach for the QA of auto-segmented OARs would be to use automated methods, rather than manual human expert inspection. This is partly because attribute inspection errors by humans will always exist (Burke, 2001), and with errors occurring at such low rates it is a distinct possibility that they may be missed by a human who does not often encounter such errors, or that multiple checks would be required to provide sufficient levels of safety (Papadakis et al., 1988), which would add an increased QA burden for a process with very small failure rates.

It is also very inefficient to use human manual inspection to identify errors that occur at a rate in the region of 1%, and therefore if it were possible to remove such a requirement, this would introduce significant further efficiency gains.

A further consideration when utilising auto-segmentation systems clinically is the risk of human complacency. For example, if a person is aware an automated system is in use and that resulting contours are normally of a sufficiently high quality, there is a risk that they will not adequately inspect all contours that are produced (Merritt et al., 2019).

To date there has been limited research into the use of automated QA to identify organ delineation errors. In 2015 Chen at al. looked at automating contour error detection using geometric attribute distribution models. Their methods produced promising results for error detection and the authors suggested that the proposed strategy could serve as a tool to support manual peer review. In 2018 Hui et al. investigated automated QA using a parametric statistical approach that looked at volumetric features. The methods used detected 37% of minor and 85% of major errors, and when combined with expert review were shown to increase error detection sensitivity.

A further option for automated QA of these systems may be to use a second, independent, AI auto-segmentation system. Resulting contours could be automatically compared using similarity metrics, and differences that exceeded set threshold values could be flagged as requiring further manual inspection. Disadvantages of such a process include the possibility that a second system using the same technology and possibly a similar homogeneity of datasets may be prone to the same error types, and also that this is likely be a financially costly solution, due to the need to purchase two commercial auto-segmentation systems.

It is important to note that this research is looking at gross failures, rather than more subtle quality issues, which may still be clinically significant. Recent research has shown that the quality of contours that are being produced by some AI auto-segmentation systems are not yet at the level where they do not sometimes require further manual editing (Robert et al., 2021 and Rhee at al., 2019). It should, however, be noted that human observer variation can also be of a clinically significant magnitude (Peng at al., 2018), yet such differences are often ignored when a human is involved. For example van der Veen at al. (2021) found mandible inter-observer variability to have a median value of 0.9 which is a similar order of magnitude to the results obtained in this study, and suggests that the quality of some AI auto-segmented OARs may already have reached human expert levels. This is supported by a study carried out by Wong et al. (2020) which concluded that the accuracy of AI auto-segmented contours is now at a comparable level to that of expert inter-observer variability.

Future research to assess the clinical significance of minor contouring failures would therefore be beneficial to determine the true importance of the often-perceived requirement for further human manual inspection of contours produced by modern commercial AI auto-segmentation systems.

A further point of interest is that suboptimal clinical contours made up 42 to 50% of initial total failures. This failure rate is a similar order of magnitude to the true failure rate of the AI system, and raises questions around clinical significance of these failures which could be investigated in future research.

Conclusion

To conclude, this study has demonstrated that gross failure rates for the H&N OARs tested, using a modern commercial AI auto-segmentation system, are extremely low, being in the region of 1% for some contours, and it is therefore advised that some method of automated QA is utilised as part of the clinical workflow.

It is also advised that manufacturers provide data relating to the datasets they use to produce AI models to assist users with identifying possible dataset bias, and further that manufacturers attempt to reduce this bias in future models using careful patient dataset selection criteria.

Recent research has shown that differences between gold standard and AI auto-segmented contours are at a comparable level to inter and intra-observer variability differences. It is therefore suggested that as resulting auto-segmented contour quality improves with future iterations of this technology, it may be possible to remove the need for manual human expert inspection in the near future. This approach would require a sufficiently accurate automated independent QA system to be included as part of the workflow.

To support this theory, further studies are required to assess the clinical significance of minor errors in auto-segmented structures.

Author contribution statement

Simon Temple contributed to study design, data collection, analysis and interpretation, and drafted the manuscript.
Carl Rowbottom contributed to study design, data collection and critical revision of the manuscript.

Acknowledgements

I would like to thank Mirada Medical Ltd for providing Mirada DLC Expert™ and the Contour Insights™ tool, which was utilised to produce structure comparison metrics.

References

Burke, R. (2001). Inspection planning for mission-critical quality. IEEE International Engineering Management Conference, pp.329–334.

Cardenas, C.E. et al. (2019). Advances in Auto-Segmentation. Seminars in Radiation Oncology, 29(3), pp.185–197. [online]. Available from: https://doi.org/10.1016/j.semradonc.2019.02.001.

Chen, H.C. et al. (2015). Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy. Medical Physics, 42(2), pp.1048–1059.

Daisne, J.F. and Blumhofer, A. (2013). Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation. Radiation Oncology, 8(1), pp.1–11.

Fan, J. et al. (2019). Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. Medical Physics, 46(1), pp.370–381.

Fraass, B. et al. (1998). American association of physicists in medicine radiation therapy committee task group 53: Quality assurance for clinical radiotherapy treatment planning. Medical Physics, 25(10), pp.1773–1829.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), pp.86–92.

Harrison, K. et al. (2022). Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clinical Oncology, 34(2), pp.74–88. [online]. Available from: https://doi.org/10.1016/j.clon.2021.12.003.

Hui, C.B. et al. (2018). Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. Medical Physics, 45(5), pp.2089–2096.

Kovacs, D.G. et al. (2018). Metal artefact reduction for accurate tumour delineation in radiotherapy. Radiotherapy and Oncology, 126(3), pp.479–486. [online]. Available from: https://doi.org/10.1016/j.radonc.2017.09.029.

Kusters, R. et al. (2020). Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. Frontiers in Big Data, 3(November), pp.1–7.

Merritt, S.M. et al. (2019). Automation-induced complacency potential: Development and validation of a new scale. Frontiers in Psychology, 10(FEB), pp.1–13.

Nelms, B.E. et al. (2012). Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer. International Journal of Radiation Oncology Biology Physics, 82(1), pp.368–378.

Papadakis, E.P. et al. (1988). Inspection decision theory: Deming inspection criterion and time-adjusted rate-of-return compared. Engineering Costs and Production Economics, 13(2), pp.111–124.

Peng, Y. lin et al. (2018). Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. Oral Oncology, 82(April), pp.1–7. [online]. Available from: https://doi.org/10.1016/j.oraloncology.2018.04.025.

Poon, A.I.F. and Sung, J.J.Y. (2021). Opening the black box of AI-Medicine. Journal of Gastroenterology and Hepatology (Australia), 36(3), pp.581–584.

Pukelsheim, F. (1994). The Three Sigma Rule. The American Statistician, 48(2), pp.88–91.

Rhee, D.J. et al. (2019). Automatic detection of contouring errors using convolutional neural networks. Medical Physics, 46(11), pp.5086–5097.

Robert, C. et al. (2021). Clinical implementation of deep-learning based auto-contouring tools–Experience of three French radiotherapy centers. Cancer/Radiotherapie, 25(6–7), pp.607–616.

Rudin, C. and Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2), pp.1–10.

Tong, N. et al. (2018). Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. Medical Physics, 45(10), pp.4558–4567.

Vandewinckele, L. et al. (2020). Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiotherapy and Oncology, 153, pp.55–66. [online]. Available from: https://doi.org/10.1016/j.radonc.2020.09.008.

Vrtovec, T. et al. (2020). Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Medical Physics, 47(9), pp.e929–e950.

Wong, J. et al. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiotherapy and Oncology, 144, pp.152–158. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.10.019.

5. Study C: Patient views on the implementation of artificial intelligence in radiotherapy

Simon Temple, Carl Rowbottom
The Clatterbridge Cancer Centre NHS Foundation Trust, Liverpool, UK.

Abstract

Purpose/Objective

To date there has been limited research looking at patient views on the implementation of artificial intelligence (AI) in radiotherapy.

The aim of this study is to adapt and utilise a validated patient questionnaire to develop an understanding of current patient views on the use of AI in radiotherapy.

Materials/Methods

An existing questionnaire, developed by Ongena et al. (2020) to assess understanding of patients' views on the implementation of AI in radiology, was adapted to the field of radiotherapy. The questionnaire grouped individual questions into five different factors, representing five underlying latent variables.

The questionnaire was distributed to over 100 cancer patients receiving radiotherapy treatment at the Clatterbridge Cancer Centre between November 2021 and March 2022.

Completed questionnaires were analysed to assess patient levels of positivity or negativity towards AI. Correlation between the five factors was assessed, along with correlation of factors with demographic variables.

Results

In total, 95 patients participated. Overall, there was a moderately negative patient view towards the use of AI in radiotherapy. Certain factors drew a more negative response than others, for example patients indicated that they would prefer a good knowledge of the procedures involved in their radiotherapy planning and treatment, and that they also desire significant personal interaction with healthcare professionals during the course of their treatment.

In terms of correlations between factors, there was a strong relationship between almost all factors, with patients who were negative towards AI for one factor displaying the same opinion for all other factors. No significant correlation was found between the demographics of age and gender and the strength of views towards the use of AI in radiotherapy.

Conclusion

This study has found that there are clear patient concerns around the use of AI in radiotherapy.

As the use of AI in this field increases in future years, it will therefore be extremely important to educate and involve patients in the future direction of this technology.

Introduction

The use of Artificial Intelligence (AI) based software is becoming increasingly common in all fields due to rapid advances in technology and the increased computing power that is now available. Healthcare is no exception to this phenomenon, and there is the potential for AI software to introduce significant quality and efficiency savings in a wide variety of specialist healthcare areas (Aung, Wong and Ting, 2021, Esteva et al., 2019, and Yu, Beam and Kohane, 2018).

When implementing any new technology in healthcare it is extremely important to consider patients views, and it has been suggested that these views should determine the limits of use of such technology (Haan et al., 2019).

The use of AI-based software is more well known in some healthcare fields than others. For example, the field of radiology has been the focus of media attention in recent years in relation to the use of such software to assess diagnostic images, and in 2020 McKinney et al. published the results of a study which concluded that AI-based software could outperform human experts (radiologists) in screening mammography to identify breast cancer.

Limited research has been carried out looking at healthcare professional views on the use of AI. For example, in 2021 Huisman et al. carried out an international survey on AI in radiology to capture the views of radiologists in terms of fear of replacement, knowledge and attitude towards AI, and a recent publication by Petragallo et al. (2022) investigated barriers and facilitators to the adoption of automated treatment planning tools.

There are fewer publications that investigate patient views on the use of AI in healthcare, for example Lennartz et al. (2021) published a survey studying patient perspectives on the use of AI in various aspects of the medical workflow, and Yakar et al. (2022) investigated Dutch patient views on AI in medicine.

There is also some limited research on patient views in the field of radiology, for example York et al. investigated patient perceptions of AI in skeletal radiography in 2020. This study concluded that there was significantly higher confidence in clinician interpretation over AI-assisted interpretation. Another radiology-based study was carried out in 2020 by Ongena at al. This study developed and validated a questionnaire relating to the implementation of AI in radiology.

In the field of radiotherapy there has been less research on patient views of AI, and a literature search carried out by the author in May 2022 could not find a single publication relating to patient views on the use of AI in radiotherapy. The absence of research into this area highlights a significant gap, which has been addressed in this study.

Questionnaires exist for measurement of patient acceptance of Consumer Health Information Technology (CHIT) using the technology acceptance model (TAM) developed by Davis et al. in 1989. Such questionnaires assume that patients are active users of the technology, and are therefore not directly applicable to the situation where patients are being subjected to results of the use of AI-based technology. This definition more accurately reflects the situation with the use of AI-based software in both radiology and radiotherapy.

Ongena et al. (2020) identified this shortcoming and the need for a new method to measure patient acceptance of such technology, and produced and validated a more appropriate questionnaire to ascertain patient views on the use of AI in radiology. The questionnaire was developed using results of a previous study (Haan et al., 2019) which used semi-structured qualitative interviews to identify key domains of patients' perspective. Multiple five point Likert-type questions were then developed, and cognitive interviews were used to refine the questionnaire.

Exploratory factor analysis (EFA) was used to assess internal consistency of questions, and the result of this analysis generated five factors relating to the perspective of the patient (table C1).

Table C1. Factors representing underlying latent variables

Factor Number	Underlying Variables
1	Distrust and accountability
2	Procedural knowledge
3	Personal interaction
4	Efficiency
5	Being informed

This study adapts the work carried out by Ongena et al. (2020) to the field of radiotherapy. Radiology and radiotherapy can be considered sufficiently similar fields of medicine such that this approach is appropriate.

Methods

This prospective study was carried out at the Clatterbridge Cancer Centre, which is a specialist cancer centre serving a population of approximately 2.4 million across the Cheshire and Merseyside region of the north west of England. The main hospital site is located in the city of Liverpool.

Questionnaire Construction

Questions from the questionnaire produced by Ongena et al. (2020) were utilised and, if applicable, adapted to the field of radiotherapy.

For example, a question from the original radiology study was '*I find it important to be able to ask questions personally about the results of a scan'*. This question was changed to read '*I find it important to be able to ask questions personally about the planning and delivery of my radiotherapy treatment*'.

Some questions did not require any adaptation, for example '*I would never blindly trust a computer*'. Other questions were specific to the field of radiology and required adjustment, for example '*I wonder how it is possible that a computer can give me the results of a scan*' was changed to '*I wonder how it is possible that a computer could draw my bodily organs on a CT scan*'.

The five factors identified by Ongena et al., (2020) were retained for this questionnaire. They were not presented to patients due to the risk of introducing bias, but were used in the subsequent analysis and findings.

Use of five point Likert-type scales was retained and, in addition, three demographic questions were included, to capture information relating to gender, age, and digital device ownership and use. The resulting questions are shown in table C2.

Questionnaire approval

Prior to circulation of the questionnaire, patient and carer voice representatives of the Clatterbridge Cancer Centre Patient Experience and Inclusion Group (PEIG) reviewed its content and approved it for distribution to patients. Suggestions made by staff and patient representatives during the consultation process were incorporated into the final version of the questionnaire.

Data Collection Procedure

The questionnaire was distributed to over 100 cancer patients who were receiving radiotherapy treatment at the Clatterbridge Cancer Centre between November 2021 and March 2022.

All patients were approached in the hospital waiting room and all were informed that participation was optional. Patients who completed the questionnaire gave verbal informed consent and were assured that completion of the questionnaire would in no way affect the quality of the treatment they would receive.

Data Analysis

Ongena et al. (2020) recoded the results of some questions to ensure a higher score always indicated a negative view of AI. Identical recoding was utilised for this study to ensure consistency with the EFA previously carried out.

As suggested by Boone (2012), a median was used to quantify the central tendency for individual question responses due to the ordinal nature of the Likert-type questions. In addition the interquartile range (IQR) was specified to provide an indication of the spread of responses. For the five overarching factors, mean and standard deviation were used for statistical analysis due to this being Likert-scale data. In addition, Pearson's test was used to identify correlations between factors, as is appropriate for Likert-scale data interpretation.

IBM SPSS Statistics (Version 28.0.1.1) was used for all statistical analysis.

Results

Patient Sample

In total 95 completed questionnaires were returned. The age of respondents ranged from 32 to 87 years (mean = 68.05, SD = 9.86), 72.3% were male and 80% owned a smartphone.

Patients' views on AI in radiotherapy

Table C2 shows results of responses to individual questions and mean scores for each of the five factors. A high score indicates negativity towards AI.

For factor 1, *distrust and accountability*, the average score was 3.24.
This indicates that patients are moderately negative concerning their trust in AI for use in radiotherapy. However, one particularly low scoring response (median score of 2) suggested that patients do not believe the use of AI would make medical professionals lazy.

The average score for factor 2, *procedural knowledge*, was 4.39, which indicates that patients have a strong desire to understand the steps involved in their radiotherapy treatment and to be able to talk to a person about it.

For factor 3, *personal interaction*, the average score was 4.36. This indicates that patients strongly prefer some level of personal interaction. Responses indicated that patients do not want to be treated as a number and need to be able to ask a human questions about the treatment they are receiving.

Table C2. Descriptive figures of 37 attitudinal items for the five questionnaire factors

Attitudinal items	Median[IQR]*	Mean	Standard deviation
Factor 1 - Distrust & Accountability of AI in radiotherapy:			
Overall		3.24	0.66
A computer can never compete against the experience of a specialist doctor (Oncologist)	4[1]		
Through human experience, an Oncologist will always be more effective than a computer	4[1]		
Humans have a better overview than computers of what happens in my body	4[1]		
It worries me when computers perform tasks on CT scans without the involvement of humans	3[2]		
I wonder how it is possible that a computer could draw my bodily organs on a CT scan	3[1]		
Artificial intelligence makes medical professionals lazy	2[2]		
I do not think radiotherapy is ready for implementing artificial intelligence in the creation of treatment plans	3[2]		
I think replacement of doctors by artificial intelligence will happen in the far future	3[2]		
I would never blindly trust a computer	4[1]		
Artificial intelligence should only be implemented to check human judgment	4[1]		
I find it worrisome that a computer does not take feelings into account	4[2]		
It is unclear to me how computers will be used in radiotherapy plan preparation	3[2]		
Even if computers are better at certain tasks, I still prefer an Oncologist	4[2]		
When artificial intelligence is used, my personal data may fall into the wrong hands	3[1]		
Artificial intelligence may prevent errors**	2[1]		
Factor 2 - Procedural Knowledge of AI in radiotherapy:			
Overall		4.39	0.55
I find it important to have a good understanding of my radiotherapy treatment	5[1]		
I find it important to be able to ask questions personally about the planning and delivery of my radiotherapy treatment	5[1]		
I find it important to talk with someone about my radiotherapy treatment	5[1]		
I find it important that a CT scan provides as much information about my body as possible	5[1]		
I find it important for my treatment to commence as soon as possible	5[1]		
I find it important to ask questions about the accuracy of my treatment	5[1]		
I find it important to be well informed about how my radiotherapy treatment plan is made	4[1]		
I find it important to understand how Oncologists work before I receive my treatment	4[1]		
Factor 3 - Personal Interaction with AI in radiotherapy:			
Overall		4.36	0.49
When discussing the detail of my treatment humans are indispensable	5[1]		
Getting the detail of my radiotherapy treatment plan involves personal contact	4[1]		
As a patient, I want to be treated as a person, not as a number	5[1]		
When a computer gives results, I would miss the explanation	4[2]		

Table C2 (continued)

Factor/Question	**Median[IQR]***	**Mean**	**Standard deviation**
I find it important to ask questions about my treatment	5[1]		
Even when computers are used to produce treatment plans, humans should always remain responsible	5[1]		
Humans and artificial intelligence can complement each other	4[1]		
Factor 4 - Efficiency of AI in radiotherapy:			
Overall		3.36	0.39
As far as I am concerned, artificial intelligence can replace medical professionals in the production of my treatment plan**	4[2]		
The sooner I receive my treatment, even when this is due to computer involvement, the more I am at ease	4[1]		
Because of the use of artificial intelligence, fewer doctors will be required in the future**	3[1]		
Producing radiotherapy treatment plans using artificial intelligence will reduce waiting times**	3[1]		
In my opinion, humans make more errors than computers**	3[1]		
Factor 5 - Being Informed of AI in radiotherapy:			
Overall		3.82	0.67
If a computer gave me results, I would not feel emotional support	4[2]		
A computer should only look at body parts that have been selected by my doctor	3[2]		
When a computer can predict that I will get a disease in the future, I would like to know, no matter what	4[1]		

* Items are measured on a five point Likert scale. For all factors a higher score indicates being more negative towards AI in radiotherapy.

** Items marked were recoded such that all questions measure in the same direction.

For factor 4, *efficiency of AI in radiotherapy*, the questionnaire produced an average score of 3.36. This score was moderately negative towards AI but does not give a strong patient preference in either direction.

For the 37 individual Likert-type questions, the IQR of 27 of the questions was 1, and the IQR of the remaining 10 questions was 2, indicating that there was a small spread of responses for the majority of questions, and therefore a relatively high level of consistency between answers from respondents.

Questions with a low median score of 2, indicating a more positive view towards AI, were '*Artificial intelligence makes medical professionals lazy*' and '*Artificial intelligence may prevent errors'.*

Table C3 shows the results of a correlation between factor analysis. A significant association was found between the majority of factors. For example factor 2 (procedural knowledge) was found to be strongly associated with all of the other four factors.

Table C3. Correlation between factors

		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1	**Pearson Correlation**	-	0.389**	0.524**	0.095	0.510**
	Sig. (2-tailed)		<0.001	<0.001	0.362	<0.001
	N	95	93	95	95	95
Factor 2	**Pearson Correlation**	0.389**	-	0.807**	0.318**	0.464**
	Sig. (2-tailed)	<0.001		<0.001	0.002	<0.001
	N	93	93	93	93	93
Factor 3	**Pearson Correlation**	0.524**	0.807**	-	0.245*	0.514**
	Sig. (2-tailed)	<0.001	<0.001		0.017	<0.001
	N	95	93	95	95	95
Factor 4	**Pearson Correlation**	0.095	0.318**	0.245*	-	0.131
	Sig. (2-tailed)	0.362	0.002	0.017		0.207
	N	95	93	95	95	95
Factor 5	**Pearson Correlation**	0.510**	0.464**	0.514**	0.131	-
	Sig. (2-tailed)	<0.001	<0.001	<0.001	0.207	
	N	95	93	95	95	95

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

The factor with the weakest association to others was factor 4 (efficiency), but even this factor had a strong association with factor 2 and a weak association with factor 3 (personal interaction).

Table C4 shows correlation between the 5 factors and other variables. A weak relationship was observed between gender and factors 1 (distrust and accountability) and 3 (personal interaction), with females being less trusting of AI and more likely to desire human interaction (although a large male gender bias was present). No relationship was found between age and any of the factors.

Table C4. Correlation between factors and other variables

		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Gender	**Pearson Correlation**	0.225*	0.077	0.203*	0.037	0.107
	Sig. (2-tailed)	0.028	0.464	0.048	0.723	0.303
	N	95	93	95	95	95
Age	**Pearson Correlation**	0.079	-0.025	-0.039	-0.021	-0.028
	Sig. (2-tailed)	0.449	0.809	0.710	0.843	0.791
	N	95	93	95	95	95
Do you have a smartphone?	**Pearson Correlation**	0.060	-0.030	-0.044	-0.067	-0.066
	Sig. (2-tailed)	0.566	0.773	0.669	0.517	0.525
	N	95	93	95	95	95

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Discussion

It is inevitable that the use of AI-based technology in healthcare will rapidly increase in the coming years. To date, however, there has been a noticeable lack of research and discussion around the subject of patient perception of such a change.

The research carried out by Ongena et al. (2020) considered this perception for the field of radiology, and this study has expanded the research to include radiotherapy. Findings were very similar, with average values for the five factors for both radiotherapy and radiology shown in table C5.

Table C5. Mean scores for radiotherapy and radiology questionnaires

Factor	Radiology Questionnaire (Ongena et al.) Mean score ± 1 SD	Radiotherapy Questionnaire Mean score ± 1 SD
1. Distrust & Accountability	3.28 ± 0.58	3.24 ± 066
2. Procedural Knowledge	4.47 ± 0.67	4.39 ± 0.55
3. Personal Interaction	4.38 ± 0.48	4.36 ± 0.49
4. Efficiency	2.89 ± 0.61	3.36 ± 0.39
5. Being Informed	3.31 ±0.70	3.82 ± 0.67

This similarity indicates that patient views regarding the use of AI in radiology can be closely equated with patient views regarding the use of AI in radiotherapy. The only factor with a potentially significant difference between scores for the two fields was ‘efficiency’. Upon inspection, there were only minor changes to some of the original questions for this factor and scores for radiotherapy are, on average, slightly higher for all efficiency-related questions. The author would therefore suggest that differences might be explained either by variations in the patient populations being surveyed or, more simply, by the expected statistical variation of resulting values, considering the sample sizes used and the standard deviation of this data (see table C5).

Regardless of any minor differences between studies, both indicate that there is an overall negative view of the use of AI in healthcare, and a clear need to improve patient knowledge and acceptance of this important technology. This finding is supported by other studies (Tran, Riveros, and Ravaud, 2019, and Yakar et al., 2022).

A suggested pre-requisite to this patient education is that healthcare staff have sufficient knowledge of the subject, particularly those who are patient-facing and may need to provide explanations to patients.

Unfortunately, recent studies have also shown a strong need to educate healthcare staff about AI, due to current poor levels of knowledge, for example, Brouwer et al. (2020) surveyed 213 medical physicists on the use of AI in their institutions and found that there was limited knowledge of ethics, legislation and data sharing among the responders. Shelmerdine, Rosendahl, and Arthurs (2022) also carried out a study looking at health care professionals' opinions on AI in paediatric radiology and identified that there is a currently a lack of education of professionals in this area. This lack of knowledge is a concern and highlights the importance of providing staff education and advice in the form of guidelines and training around the implementation and quality assurance of AI-based applications.

With regard to the education of both staff and patients, some commercial providers of AI software have already recognised this issue and have employed staff whose role is solely to address it. For example, one company has a 'Patient Communications Editor' whose role is to provide information to both staff and patients directly. The aim of this is two-fold:

- To equip patient-facing staff such as clinicians with basic AI knowledge and to provide them with the vocabulary to explain the main concepts of AI to patients in an accessible way.

- To explain the basic concepts of AI to patients directly in order to alleviate concerns and make patients feel more comfortable about the use of AI in their own healthcare.

Regardless of whether such education is provided by commercial providers or healthcare staff themselves, this study has highlighted the importance of demonstrating to patients that the use of AI in their own healthcare is safe, effective and ethical.

In terms of radiotherapy treatment planning, and more specifically AI auto-segmentation, certain individual questions can be more easily equated to this subject area. For example, the question ‘*Through human experience an Oncologist will always be more effective than a computer’, can arguably be used to assess patient views on the effectiveness of AI auto-segmentation software. The resulting median score for this question was 4, indicating that patients agree with the statement, and do not believe computers can be sufficiently effective.*

Existing research already indicates that current AI auto-segmentation performance levels rival that of the inter-observer variability of a Radiation Oncologist (Wong et al., 2020). As AI technology advances, it is not unreasonable to expect AI system performance may exceed human performance in future years. This supports the identified need to educate staff and patients around current and expected future levels of performance of this technology.

For the question ‘*I wonder how it is possible that a computer could draw my bodily organs on a CT scan’*, which directly relates to the subject of auto-segmentation, the median answer was 3, indicating that patients neither agree nor disagree with this statement. This suggested that patients have already accepted a certain level of performance of computers in radiotherapy.

Conclusion

In summary, there is currently very limited knowledge of patient views of AI in radiotherapy, but the results of this study demonstrate that patients have significant concerns around the use of this technology in radiotherapy.

There is therefore a need for both further research on patient views, and also for the provision of appropriate staff and patient education in this rapidly growing area.

Author contribution statement

Simon Temple contributed to study design, data collection, analysis and interpretation, and drafted the manuscript.
Carl Rowbottom contributed to study design and critical revision of the manuscript.

Acknowledgements

I would like to thank Janette Simpson and Jennifer Cadwallader for their assistance distributing questionnaires to patients.

References

Aung, Y.Y.M., Wong, D.C.S. and Ting, D.S.W. (2021). The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare. British Medical Bulletin, 139(1), pp.4–15.

Boone, H.N. and Boone, D.A. (2012). Analyzing Likert data. Journal of Extension, 50(2).

Brouwer, C.L. et al. (2020). Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. Physics and Imaging in Radiation Oncology, 16(July), pp.144–148.

Davis, F.D.., Bagozzi, R.P.. and Warshaw, P.R.. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Management Science, 35(8), pp.982–1003. [online]. Available from: https://www.jstor.org/stable/2632151.

Esteva, A. et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), pp.24–29. [online]. Available from: http://dx.doi.org/10.1038/s41591-018-0316-z.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), pp.86–92.

Haan, M. et al. (2019). A Qualitative Study to Understand Patient Perspective on the Use of Artificial Intelligence in Radiology. Journal of the American College of Radiology, 16(10), pp.1416–1419.

Huisman, M. et al. (2021). An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. European Radiology, 31(9), pp.7058–7066.

Huisman, M. et al. (2021). An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. European Radiology, 31(11), pp.8797–8806.

Jutzi, T.B. et al. (2020). Artificial Intelligence in Skin Cancer Diagnostics: The Patients' Perspective. Frontiers in Medicine, 7(June).

Kusters, R. et al. (2020). Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. Frontiers in Big Data, 3(November), pp.1–7.

Lennartz, S. et al. (2021). Use and control of artificial intelligence in patients across the medical workflow: Single-center questionnaire study of patient perspectives. Journal of Medical Internet Research, 23(2), pp.1–10.

Markus, A.F., Kors, J.A. and Rijnbeek, P.R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, 113(December 2020), p.103655. [online]. Available from: https://doi.org/10.1016/j.jbi.2020.103655.

McKinney, S.M. et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), pp.89–94. [online]. Available from: http://dx.doi.org/10.1038/s41586-019-1799-6.

Ongena, Y.P. et al. (2020). Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. European Radiology, 30(2), pp.1033–1040.

Petragallo, R. et al. (2022). Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: A survey study of medical dosimetrists. Journal of Applied Clinical Medical Physics, (February), pp.1–10.

Shelmerdine, S.C., Rosendahl, K. and Arthurs, O.J. (2022). Artificial intelligence in paediatric radiology: international survey of health care professionals' opinions. Pediatric Radiology, 52(1), pp.30–41.

Tran, V.-T., Riveros, C. and Ravaud, P. (2019). Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. npj Digital Medicine, 2(1), pp.1–8. [online]. Available from: http://dx.doi.org/10.1038/s41746-019-0132-y.

Wong, J. et al. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiotherapy and Oncology, 144, pp.152–158. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.10.019.

Yakar, D. et al. (2022). Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine. Value in Health, 25(3), pp.374–381. [online]. Available from: https://doi.org/10.1016/j.jval.2021.09.004.

York, T., Jenney, H. and Jones, G. (2020). Clinician and computer: A study on patient perceptions of artificial intelligence in skeletal radiography. BMJ Health and Care Informatics, 27(3), pp.1–7.

Yu, K.H., Beam, A.L. and Kohane, I.S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), pp.719–731. [online]. Available from: http://dx.doi.org/10.1038/s41551-018-0305-z.

6. Critical Appraisal of Thesis

6.1. Introduction

This thesis has focussed on the use of AI-based software for auto-segmentation of OARs in H&N radiotherapy treatment planning, incorporating a study on patient views on AI in radiotherapy.

As previously discussed, the use of this technology has the potential to provide significant efficiency savings and quality improvements (Kosmin et al., 2019). Although the anatomical site of H&N has been used for this research, findings are expected to equally apply to OARs in other anatomical sites (Cardenas et al., 2019).

This chapter will discuss how each study contributes to knowledge in the field and clinical practice. It will also discuss limitations of the research methods used, and provide suggestions for the future direction of research relating to the use of AI-based software in radiotherapy treatment planning and associated patient views.

6.2. Study A

This study evaluated the performance of four commercial AI auto-segmentation systems for five commonly used head and neck OARS using a common patient cohort.

6.2.1. Study A: Contribution to knowledge and clinical practice

To date there has been no published research comparing multiple commercial AI auto-segmentation systems using a common patient cohort. This statement applies not just to the H&N anatomical site, but to all anatomical sites.

Findings of this research therefore add new knowledge in terms of the understanding of levels of performance variation among commercial systems. The availability of this knowledge may be beneficial for radiotherapy centres looking to purchase and implement an AI auto-segmentation system.

Due to the similar levels of performance identified, other factors such as financial cost, range of structures outlined and ease of system use may become relatively more important when evaluating systems.

Further to this, having a good understanding of the quality of resulting auto-segmented contours is beneficial for clinical practice, as it allows professionals involved in the planning process to develop sufficient levels of trust in a system, which is an important aspect of their clinical use (Harrison et al., 2022).

6.2.2. Study A: Strengths and weaknesses of study

In terms of strengths, this study utilised the majority of commercial systems which were available at the time of the work, and used a relatively large number of patient datasets (50) to obtain good statistical power, as recommended by Vrtovec et al. (2020).

Another strength of the study was the use of multiple comparison metrics, namely 3D DSC, 2D 95% HD and APL. Vrtovec et al. (2020) recommended that DSC should always be accompanied by at least one distance metric, preferably HD or its 95-percentile version. In addition the APL metric was used, which is considered to be a more meaningful metric in terms of clinical time savings (Vaassen et al., 2020).

In terms of weaknesses, there were some of aspects of this study which could arguably have been improved with the availability of more time and resource. Although the gold standard contours were peer reviewed, there is still likely to be a level of inter and intra-observer availability present (Cardenas et al., 2019 and Peng et al., 2018). In order to reduce this further, use of a STAPLE algorithm (Warfield, Zou and Wells, 2004) or a STEPS algorithm (Cardoso et al., 2013) to produce ground truth contours from multiple independent contours based on the work of multiple human experts would be preferable. This methodology is likely to give a superior set of gold standard contours which are more robust to outliers (Deeley et al., 2011), but would require significant resource from multiple expert observers.

Regarding comparison metrics utilised for the study, ideally the surface DSC (Nikolov et al., 2018) would also have been employed. This metric was not available to the author at the time of the study but there is evidence to suggest that it may be superior to the volumetric DSC (Vaassen et al., 2020) in being a better indicator of contouring time-savings and also for providing an additional quantifiable surrogate for assessment of quality.

The APL metric was developed as a proxy for contour adjustment time savings (Vaassen et al., 2020) and does not necessarily provide information around absolute geometrical differences between contours, because it does not take into account contours that need to be deleted from entire CT slices. This should be considered when using the metric for contour comparison. A modified version of this metric could be used in addition to the standard APL metric by interchanging reference and test structures and using the greater of the two. This revised metric would provide a more robust indication of absolute geometric differences between contours.

Some of the most commonly used OARs for H&N radiotherapy planning were evaluated in this study, but some key OARs were also missing, for example pharyngeal constrictor muscles, oral cavity and glottic larynx. Reasons for the absence of such OARs from the study were due to the absence of the OAR from either the gold standard reference dataset or from one or more of the AI systems. This situation is likely to change as systems develop with time and new OARs are added to system libraries.

The majority of commercially available systems at the time of the research were evaluated. A number of new commercial products have since been released, and in addition new models have been produced for at least some of the systems evaluated. This highlights the rapid change that is continually occurring in the area of AI auto-segmentation, and whilst not a criticism of the study design, it does mean that results of such research may become outdated in a relatively short period of time. This should be taken into account when using the existing literature to acquire knowledge regarding performance of such systems.

6.2.3. Study A: Future direction of research

Some limitations of this study have been identified in relation to available resources. Further knowledge could therefore be obtained by carrying out a modified version of the study with the following adaptations:

- Multiple observers to independently generate expert contours.
- Use of a STAPLE or STEPS algorithm to produce higher quality ground truth reference datasets.
- Utilise a larger number of reference datasets to increase statistical power.
- Addition of the surface DSC comparison metric and a modified APL metric.
- Addition of further commonly used H&N OARs.
- Addition of newly available commercial systems and updated models from existing systems.

In addition to OARs, there is increasing research looking at the auto-segmentation of target volumes, both nodal and primary. For example Cardenas et al. (2021) developed a DL model to auto-segment H&N lymph and retropharyngeal nodes which produced clinically acceptable results based on a qualitative review by physicians. Future research into AI auto-segmentation for H&N could therefore incorporate evaluation of target volumes.

Whilst this research has provided useful information regarding the accuracy of AI auto-segmented structures, the clinical effect of using unmodified structures is largely unknown. A future direction of research could be an investigation into the relationship between auto-segmentation and dosimetric aspects of the resulting treatment plan, as recommended by Valentini et al. (2014). Such research would provide insight into implications of the use of unmodified automatically created structures, and could be used to inform on clinical practice. Patient Reported Outcomes (PRO) (Caissie et al., 2018) could be used to monitor such an intervention.

It should also be noted that most modern inverse treatment planning approaches use some form of normal tissue objective (NTO) (Ndrayani et al., 2022) to reduce the dose to surrounding OARs, and therefore planning systems are not relying solely on OAR contours for this function.

A key finding of this research was that systems appeared to perform at similar levels for a given OAR. It could therefore be argued that other aspects associated with commercial AI auto-segmentation systems may become relatively more important, for example speed of performance and ease of use.

Further research could therefore include an assessment of system efficiency and incorporate technical detail of system function, for example whether local servers are required or whether a system is cloud based. Such information is likely to be of important value for users when considering system options. A suggestion around this is for professional bodies to produce a checklist for evaluation, which could be used to help drive commercial AI auto-segmentation products in directions that are useful for end users.

A further area for future research is regarding the use of MR for auto-segmentation. This study has focussed on the use of CT, but radiotherapy treatment planning often requires the use of a combination of CT and MR due to the higher quality soft tissue contrast visible on MR (Winter et al., 2018). There are currently a limited number of published studies on this subject area (Močnik et al., 2018, Hague et al., 2021) and it is likely to be an important area for future research.

6.3. Study B

This study investigated failure rates of a commercial AI auto-segmentation system using a large patient cohort for commonly used H&N OARs.

6.3.1. Study B: Contribution to knowledge and clinical practice

Prior to this study a literature review did not identify any existing research that quantified failure rates of commercial AI auto-segmentation systems. Results from the study therefore contribute important new information to the existing knowledgebase.

Failure rates were found to be in the region of 1% on average for the OARs investigated, and reasons for failure appeared to be associated with the existence of a non-standard element in either the CT scan, the patient setup position or the patient anatomy. This highlights the importance of end-users having greater information from commercial providers on the characteristics of the datasets used for model generation (Gebru et al., 2021).

The availability of this knowledge of failure rates and failure reason has potentially valuable safety and efficiency implications for clinical practice. If users understand when such systems are more likely to fail then they may be more able to target QA methods to identify failures.

Further to this, knowledge of failure data can also be used to determine appropriate QA for AI auto-segmentation systems. This is again likely to introduce further quality and efficiency savings. Due to the low failure rates identified, manual human inspection for every individual case will be a very inefficient process (Papadakis et al., 1988) and the development of routine automatic patient-specific QA is strongly advised.

6.3.2. Study B: Strengths and weaknesses of study

In terms of study strengths, a large number of patient datasets (500) were used in order to provide high statistical power. Use of such a large patient cohort meant that a sufficient number of unusual patient cases were included, and had a smaller number of datasets been used in the study it is very possibly that no failures would have been identified. This large cohort also gives confidence in the accuracy of resulting failure rates and in the identification of different possible failure modes.

The study also provides new information on QA requirements for AI auto-segmentation. For example, due to the identified reasons for failure it can be surmised that use of a secondary AI to check a primary AI includes the risk that both systems may be similarly biased, leading to a failure of the QA process.

In terms of study weaknesses, the resulting reason for failure was attributed to the quality of the clinical reference contour for a relatively large number of failures (42% to 50% depending on OAR). This provides important information in itself i.e. that human failure rates may be comparable to AI auto-segmentation system failure rates, nevertheless, it is also a weakness of the study in that a higher quality set of reference contours would yield more robust results, but this would require a significant amount of resource for such a large dataset. It is also important to note that all contours had been checked as part of standard QA processes and used to develop patient treatments.

A further weakness of the study was the use of a single comparison metric. Ideally multiple comparison metrics should be utilised for such studies (Vrtovec et al., 2020) and therefore the addition of a distance metric such as HD is recommended. The absence of such a metric was a reflection of the available time resource when using a large 500 patient dataset.

When failures were identified in the study, contours were manually inspected by a single expert human observer. It could be argued that this is a further study weakness, which could be addressed by using multiple human expert observers.

Other resource-related weaknesses were the limited number of OARs investigated and the use of a single commercial DL system.

6.3.3. Study B: Future direction of research

Aside from addressing some of the identified study weaknesses, which were present due to resource limitations, such as improved gold standard contours, multiple comparison metrics, wider range of OARs and use of multiple commercial systems, study results can be used to suggest a number of future directions for related research.

One finding was that DL systems appeared to encounter issues with ‘unusual’ patient cases, for example a non-standard patient set-up position or post-surgery patient anatomy. This suggests that a level of dataset bias (Kusters et al., 2020) may be present in existing DL models and it is therefore advised that further research is carried out in this area. If such research is carried out in collaboration with manufacturers then this may lead to improved results in future model iterations. This information can also be used to direct commissioning efforts by ensuring that such ‘unusual’ cases are included in any local commissioning datasets. A further extension of this would be to develop a database of unusual cases that could be used for testing when implementing new AI systems.

Another future direction for research is in the use of data augmentation (Khalifa, Loey and Mirjalili, 2022). This is a technique that can be used to address limitations in deep learning training datasets by manipulating existing data in a variety of ways to produce larger training datasets and improve model performance.

One discussion point arising from the study is around the ‘black box’ nature of DL systems (Poon and Sung, 2021), and paucity of information from commercial vendors on the construction of patient models. If manufacturers were encouraged by end-users and professional bodies to provide standardised datasheets containing information about datasets that were used to produce DL models (Gebru et al., 2021) then this information could be used to inform future research, local testing and development of quality assurance programmes. This approach is being encouraged in other areas of AI development (Rudin and Radin, 2019).

Further possibilities for future research are in the development of consensus for automated QA methods in order to avoid 100% inspection, which is a significant waste of resource due to the extremely small failure rates identified.

It is not yet clear what an ideal QA programme in this area would be. One possibility is the use of a second, independent AI auto-segmentation system to identify failures.

Claessens et al. (2022) recently investigated such a possibility for prostate radiotherapy, and it is recommended that similar research be extended to other anatomical sites such as H&N. There are implications for QA approaches if using a secondary AI system to check a primary AI system, for example it should be ensured that models, algorithms and datasets used by the secondary algorithm are completely independent of the primary. Based on the findings of this research it would also be important to include unusual cases and sufficient data augmentation in both primary and secondary systems.

As identified in the review of existing literature (Chapter 2), recent research around automatic QA of contours makes use of either data abstraction or use of a secondary AI system. It may be that a robust QA system requires a combination of both methods in order to minimise weaknesses of each approach. This is a further suggested direction of research.

It is important to note that this study and several other recent studies (Wong et al., 2020) have found that DSC values for AI auto-segmented OARs are comparable with reported levels of inter and intra-observer variability. This indicates that modern systems may have reached human levels of performance, and if they have, it could be argued that there is in fact no need to manually inspect every auto-segmented contour, as is currently recommended (Vandewinckele et al., 2020), because 100% independent inspection of human generated OARs is not standard practice in radiotherapy treatment planning. Further research into this possibility is suggested, to include an evaluation of the clinical significance of minor and major auto-segmentation failures.

6.4. Study C

This study adapted and utilised a standardised patient questionnaire to develop an understanding of patient views of AI in radiotherapy.

6.4.1. Study C: Contribution to knowledge and clinical practice

A literature review carried out in April 2022 could not find any existing research into patient views on the use of AI in radiotherapy. Results from this study therefore add important new information to the current, extremely sparse, knowledgebase, and support findings of other non-radiotherapy studies that patients have real concerns around the use of AI in healthcare (Lennartz et al., 2021).

Regarding use of this knowledge in clinical practice, it is important that any patient-facing staff are aware of patient concerns, and also that they have sufficient levels of knowledge and understanding to reassure patients on this subject. Research to date does not indicate that this is currently the case (Brouwer et al., 2020) and therefore both staff and patient education is an important future step. A study by Elsner et al. (2017) highlighted the importance of the radiation therapist role in reducing patient anxiety by sharing information with patients.

6.4.2. Study C: Strengths and weaknesses of study

In terms of strengths of this study, a previously developed and validated questionnaire has been adapted. The questionnaire was originally developed for radiology, but due to the similarity between fields only minor modifications to the questionnaire were required. This provides high confidence in the questions and factors that have been used.

A further strength of the study is that 95 completed questionnaires were received, which provides good statistical power.

In terms of study weaknesses, this questionnaire was adapted from a questionnaire which had been validated for use on the subject of radiology, and therefore it could be argued that some re-validation work would be beneficial, and that questions could be further tailored to be radiotherapy-specific. As described earlier, the fields or radiology and radiotherapy are similar and it is therefore suggested that this is a minor weakness.

Upon analysis of the data it was discovered that 72.3% of respondents were male. This gender bias can be attributed to the fact that the majority of questionnaires were distributed to patients undergoing treatment at a satellite centre where a greater proportion of patients were male due to the high numbers of prostate radiotherapy treatments occurring at this site. It is recommended that for future studies an attempt be made to distribute questionnaires to a more balanced patient group in terms of gender.

Other weaknesses include the fact that this questionnaire only captured views from a small percentage of patients treated at the centre during the collection period, and that this was a single centre study, mainly carried out at a single site. Further to this, only one instance of the questionnaire was used, so no longitudinal data was produced to provide information on any changes in patient attitudes over time.

6.4.3. Study C: Future direction of research

The author believes this is the first study researching patient views on AI in radiotherapy, and it is therefore advised that further such research is carried out in this area to strengthen knowledge. Such research could be extended to the use of patient interviews and focus groups (Schulte-Vieting et al., 2021).

It is also advised that more research is carried out to understand staff views and knowledge levels on this subject, another area where a clear gap in the existing literature has been identified.

A further suggestion for future research is a longitudinal study to investigate patient attitude changes over time. Such a study could make use of quality improvement interventions such as staff and/or patient education sessions, with repeated use of the questionnaire at a later time point to test effectiveness of these interventions.

Studies could also be extended to cover multiple sites, or more ambitiously to be national studies to test whether patient attitudes differ depending on factors such as geographical location, level of deprivation etc.

6.5. References

Brouwer, C.L. et al. (2020). Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. Physics and Imaging in Radiation Oncology, 16(July), pp.144–148.

Caissie, A. et al. (2018). Improving patient outcomes and radiotherapy systems: A pan-Canadian approach to patient-reported outcome use. *Medical Physics*, 45(10), pp.e841–e844.

Cardenas, C.E. et al. (2019). Advances in Auto-Segmentation. Seminars in Radiation Oncology, 29(3), pp.185–197. [online]. Available from: https://doi.org/10.1016/j.semradonc.2019.02.001.

Cardenas, C.E. et al. (2021). Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach. International Journal of Radiation Oncology Biology Physics, 109(3), pp.801–812. [online]. Available from: https://doi.org/10.1016/j.ijrobp.2020.10.005.

Cardoso, J. M. et al. (2013). STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical Image Analysis, 17(6), pp.671–684. [online]. Available from: http://dx.doi.org/10.1016/j.media.2013.02.006.

Claessens, M. et al. (2022). Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm. Physics in Medicine & Biology, 67(11), p.1150

Deeley, M.A. et al. (2011). Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: A multi-expert study. *Physics in Medicine and Biology*, 56(14), pp.4557–4577.

Elsner, K. et al. (2017). Reduced patient anxiety as a result of radiation therapist-led psychosocial support: a systematic review. Journal of Medical Radiation Sciences, 64(3), pp.220–231.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), pp.86–92.

Hague, C. et al. (2021). An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. Radiotherapy and Oncology, 158, pp.112–117. [online]. Available from: https://doi.org/10.1016/j.radonc.2021.02.018.

Harrison, K. et al. (2022). Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clinical Oncology, 34(2), pp.74–88. [online]. Available from: https://doi.org/10.1016/j.clon.2021.12.003.

Khalifa, N.E., Loey, M. and Mirjalili, S. (2022). A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, 55(3), pp.2351–2377. [online]. Available from: https://doi.org/10.1007/s10462-021-10066-4.

Kosmin, M. et al. (2019). Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiotherapy and Oncology, 135, pp.130–140. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.03.004.

Lennartz, S. et al. (2021). Use and control of artificial intelligence in patients across the medical workflow: Single-center questionnaire study of patient perspectives. Journal of Medical Internet Research, 23(2), pp.1–10.

Močnik, D. et al. (2018). Segmentation of parotid glands from registered CT and MR images. Physica Medica, 52(January), pp.33–41.

Nikolov, S. et al. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv, pp.1–31. [online]. Available from: http://arxiv.org/abs/1809.04430.

Ndrayani, L.I. et al. (2022). Normal tissue objective (NTO) tool in Eclipse treatment planning system for dose distribution optimization. *Polish Journal of Medical Physics and Engineering*, 28(June), pp.99–106.

Papadakis, E.P. et al. (1988). Inspection decision theory: Deming inspection criterion and time-adjusted rate-of-return compared. Engineering Costs and Production Economics, 13(2), pp.111–124.

Peng, Y. lin et al. (2018). Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. Oral Oncology, 82(April), pp.1–7. [online]. Available from: https://doi.org/10.1016/j.oraloncology.2018.04.025.

Rudin, C. and Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, 1(2), pp.1–10.

Schulte-Vieting, T. et al. (2021). Developing a question prompt list for the oncology setting: A scoping review. *Patient Education and Counseling*, 105(7), pp.1689–1702. [online]. Available from: https://doi.org/10.1016/j.pec.2021.10.006.

Vaassen, F. et al. (2020). Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology, 13(November 2019), pp.1–6. [online]. Available from: https://doi.org/10.1016/j.phro.2019.12.001.

Valentini, V. et al. (2014). Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiotherapy and Oncology, 112(3), pp.317–320. [online]. Available from: http://dx.doi.org/10.1016/j.radonc.2014.09.014.

Vandewinckele, L. et al. (2020). Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiotherapy and Oncology*, 153, pp.55–66. [online]. Available from: https://doi.org/10.1016/j.radonc.2020.09.008.

Vrtovec, T. et al. (2020). Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Medical Physics, 47(9), pp.e929–e950.

Warfield, S.K., Zou, K.H. and Wells, W.M. (2004). Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging, 23(7), pp.903–921.

Winter, R.M. et al. (2018). Assessment of image quality of a radiotherapy-specific hardware solution for PET/MRI in head and neck cancer patients. Radiotherapy and Oncology, 128(3), pp.485–491.

Wong, J. et al. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiotherapy and Oncology*, 144, pp.152–158. [online]. Available from: https://doi.org/10.1016/j.radonc.2019.10.019.

7. Conclusions

The use of AI-based software in radiotherapy is a rapidly developing field, with continually increasing numbers of commercial providers and OARs contoured. Despite this, the research carried out in this thesis has maintained validity even within this developing environment.

The work presented in this thesis has identified comparable levels of performance between multiple commercial AI auto-segmentation systems for the H&N OARs investigated. These findings highlight the importance of considering other factors when procuring such a system, for example financial cost, range of contours offered and ease of use.

Use of a very large patient cohort has determined that gross failure rates of modern AI auto-segmentation systems are extremely low. Automated QA of such systems is therefore advisable due to these low failure rates, but there are still deficiencies in current models that require addressing by manufacturers, for example the inability to deal with more unusual patient anatomy or setup position.

The 'black box' nature of current AI systems is an issue, and the provision of datasheets by manufacturers accompanying AI models is therefore recommended, to include details of the training datasets and any augmentation used for model development in order to inform on possible dataset bias. This information can be used to develop future methods of testing and QA for such systems.

Regarding patient views on the use of AI software in their own radiotherapy treatment planning, there are clear concerns, and patients currently have a moderately negative view towards AI. The work carried out in this thesis has identified a need for further education of both staff and patients in this subject area to allow individuals to make a more informed decision, and to be more receptive towards use of this technology in their healthcare.

In conclusion, the thesis has demonstrated that commercial implementations of AI auto-segmentation in radiotherapy have reached the level of human performance, with high levels of consistency and very low failure rates. Although technical improvements and development in the AI software will continue, greater focus on education of clinical staff and patients in the role of AI in radiotherapy is needed.

APPENDIX A List of Modules

AMBS – A Units		
Unit title	**Credits**	**Assignment wordcount**
A1: Professionalism and professional development in the healthcare environment	30	Assignment 1 – 2500 words Group presentation (10%) Assignment 2 – 3000 words
A2: Theoretical foundations of leadership	20	Assignment 1 – 3000 words Assignment 2 – 3000 words
A3: Personal and professional development to enhance performance	30	Assignment 1 – 1500 words Assignment 2 – 4000 words
A4: Leadership and quality improvement in the clinical and scientific environment	20	Assignment 1 – 3000 words Assignment 2 – 3000 words
A5: Research and innovation in health and social care	20	Group presentation (25%) Assignment – 4000 words
Medical Physics – B Units		
B1: Medical Equipment Management	10	2000 word assignment
B2: Clinical and Scientific Computing	10	2000 word assignment
B3: Dosimetry	10	1500 word assignment
B4: Optimisation in Radiotherapy and Imaging	10	Group presentation 1500 word assignment
B6: Medical statistics in medical physics	10	3000 word assignment
B8: Health technology assessment	10	3000 word assignment
B9: Clinical applications of medical imaging technologies in radiotherapy physics	20	Group presentation 2000 word assignment
B10a: Advanced Radiobiology	10	2000 assignment
B10f: Radiation Protection Advice	10	Virtual experiment + 2000 word assignment
B10m: Advanced Computing	10	Presentation + 1500 word assignment
Generic B Units		
B5: Contemporary issues in healthcare science	20	4000 word assignment
B7: Teaching Learning Assessment	20	20 minute group presentation

APPENDIX B Patient Questionnaire

The Clatterbridge Cancer Centre NHS
NHS Foundation Trust

Patient views on the use of Artificial Intelligence in radiotherapy treatment planning

Information for participants:

Introduction

Thank you for agreeing to take part in this survey.

Completion of the survey is confidential and entirely voluntary, and if you prefer not to participate the care that you receive will **in no way** be affected.

Survey results will be used as part of a scientific research study designed to improve our understanding of patient views on the use of AI in healthcare, and in particular radiotherapy.

Artificial Intelligence (AI)

Computer software that makes use of AI has become increasingly common in recent years due to the much greater processing power of modern computers.

This newfound computing power allows software to process data in a similar way to the human brain, and effectively 'learn'.

This in turn allows the software to perform tasks that have previously been impossible for computers, often with equal or even higher quality results than those produced by a human being.

In the future, it is possible that a number of tasks relating to the production of radiotherapy treatment plans will be performed using artificial intelligence.

The radiotherapy treatment planning process

Treatment plans for radiotherapy are produced over a number of stages, summarised below:

- CT scan
- Outlining (drawing) of organs at risk on CT scan by Dosimetrist
- Outlining (drawing) of tumour on CT scan by Clinical Oncologist
- For complex or rare cancers - Peer Review of tumour volume by other Clinical Oncologists prior to treatment
- Plan creation by Dosimetrist or Physicist
- Plan check by a different Dosimetrist or Physicist
- Plan approval by Clinical Oncologist

Examples of the possible use of AI software in the above stages are in the 'outlining' of critical bodily organs on CT scans, and in the creation of the treatment plan, which controls how radiation is precisely delivered into the body during your radiotherapy treatment, whilst at the same time, avoiding and therefore minimising damage to the surrounding healthy tissue.

Please turn over to begin the questionnaire

Questionnaire

All responses will be treated anonymously, and if you require any help completing the questionnaire please discuss with your treatment radiographers.

Q1. What is you gender? Male □ Female □ Non-binary/other □ Prefer not to say □

Q2. What is your age in years? ……….

Q3. Do you own and regularly use a smartphone? Yes □ No □

Below are several statements. Please check the box that most closely represents your opinion.

		Strongly disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
Q4.	A computer can never compete against the experience of a specialist doctor (Oncologist)	□	□	□	□	□
Q5.	Through human experience, an Oncologist will always be more effective than a computer	□	□	□	□	□
Q6.	Humans have a better overview than computers of what happens in my body	□	□	□	□	□
Q7.	It worries me when computers perform tasks on CT scans without the involvement of humans	□	□	□	□	□
Q8.	I wonder how it is possible that a computer could draw my bodily organs on a CT scan	□	□	□	□	□
Q9.	Artificial intelligence makes medical professionals lazy	□	□	□	□	□
Q10.	I do not think radiotherapy is ready for implementing artificial intelligence in the creation of treatment plans	□	□	□	□	□
Q11.	I think replacement of doctors by artificial intelligence will happen in the far future	□	□	□	□	□
Q12.	I would never blindly trust a computer	□	□	□	□	□
Q13.	Artificial intelligence should only be implemented to check human judgment	□	□	□	□	□
Q14.	I find it worrisome that a computer does not take feelings into account	□	□	□	□	□
Q15.	It is unclear to me how computers will be used in radiotherapy plan preparation	□	□	□	□	□

		Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Q16.	Even if computers are better at certain tasks, I still prefer an Oncologist	□	□	□	□	□
Q17.	When artificial intelligence is used, my personal data may fall into the wrong hands	□	□	□	□	□
Q18.	Artificial intelligence may prevent errors	□	□	□	□	□
Q19.	I find it important to have a good understanding of my radiotherapy treatment	□	□	□	□	□
Q20.	I find it important to be able to ask questions personally about the planning and delivery of my radiotherapy treatment	□	□	□	□	□
Q21.	I find it important to talk with someone about my radiotherapy treatment	□	□	□	□	□
Q22.	I find it important that a CT scan provides as much information about my body as possible	□	□	□	□	□
Q23.	I find it important for my treatment to commence as soon as possible	□	□	□	□	□
Q24.	I find it important to ask questions about the accuracy of my treatment	□	□	□	□	□
Q25.	I find it important to be well informed about how my radiotherapy treatment plan is made	□	□	□	□	□
Q26.	I find it important to understand how Oncologists work before I receive my treatment	□	□	□	□	□
Q27.	When discussing the detail of my treatment humans are indispensable	□	□	□	□	□
Q28.	Getting the detail of my radiotherapy treatment plan involves personal contact	□	□	□	□	□
Q29.	As a patient, I want to be treated as a person, not as a number	□	□	□	□	□
Q30.	When a computer gives results, I would miss the explanation	□	□	□	□	□
Q31.	I find it important to ask questions about my treatment	□	□	□	□	□
Q32.	Even when computers are used to produce treatment plans, humans should always remain responsible	□	□	□	□	□

		Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Q33.	Humans and artificial intelligence can complement each other	□	□	□	□	□
Q34.	As far as I am concerned, artificial intelligence can replace medical professionals in the production of my treatment plan	□	□	□	□	□
Q35.	The sooner I receive my treatment, even when this is due to computer involvement, the more I am at ease	□	□	□	□	□
Q36.	Because of the use of artificial intelligence, fewer doctors will be required in the future	□	□	□	□	□
Q37.	Producing radiotherapy treatment plans using artificial intelligence will reduce waiting times	□	□	□	□	□
Q38.	In my opinion, humans make more errors than computers	□	□	□	□	□
Q39.	If a computer gave me results, I would not feel emotional support	□	□	□	□	□
Q40.	A computer should only look at body parts that have been selected by my doctor	□	□	□	□	□
Q41.	When a computer can predict that I will get a disease in the future, I would like to know, no matter what	□	□	□	□	□

Anything else to say?

If you have any other comments relating to the use of AI in radiotherapy or healthcare in general please write them in the box below:

Thank you very much for completing the survey. Your opinions are extremely valuable to us.

APPENDIX C Innovation Proposal

The Clatterbridge Cancer Centre
NHS Foundation Trust

Business Case Template for Investments

Agenda Item:		**Date:**	28 March 2022
Subject / Title:	Purchase of AI auto-segmentation system		
Author:	Simon Temple		
For:	Radiation Services Division		

1. Executive Summary

This proposal relates to the purchase of commercially developed AI-based software to automatically segment organs at risk and target volumes on CT scans for radiotherapy treatment planning.

At present all structures are manually delineated on multiple CT slices for every individual patient by either clinical oncologists, physicists or treatment planning staff. This can be an extremely time consuming process (Fritscher et al., 2014). Studies have identified significant inter and intra-observer variability in the manual delineation of structures on CT scans (Peng et al., 2018) which may have implications for clinical outcomes (Sherer et al., 2021). Using commercially available AI software for auto-contouring reduces this observer bias.

An initial scoping project to assess system performance on a free trial basis has previously been completed.

Based on indicative quotes the total cost of investment for year 1 is £56,650, with a recurrent cost of £40,750 in subsequent years assuming the system is used for 2000 patients annually. This is an additional cost per patient of around £22 over five years.

This software is not currently standard of care for radiotherapy planning and does not attract any additional funding from NHS England tariff. Therefore possible sources of funding for the development are local investment, cost savings, capital contingency or charitable funding.

Option 2, the purchase of an AI auto-segmentation system is recommended, and the division is asked to support the case to implement such a system.

2. Case for Investment

Business need:
The workforce, comprising clinical oncologists, physicists and treatment planning staff, currently spend a significant amount of time manually delineating organs at risk (OAR) and target volumes on CT scans (Fritscher et al., 2014).

Due to the current high workloads and the available staff resource, particularly clinical oncologists, delays at this stage of the radiotherapy planning pathway often cause delays to radiotherapy treatment start dates, which can lead to missed targets and may also have a negative effect on patient outcomes due to an increased risk of local recurrence (Chen et al., 2008).
Studies have also identified significant inter and intra-observer variability in the manual delineation of structures on CT scans (Peng et al., 2018) which may have implications for clinical outcomes (Sherer et al., 2021).

Further to this, increases in complexity of modern radiotherapy techniques such as VMAT, and increases in frequency of adaptive planning, have seen associated increases in required planning resources. In addition, recent studies have shown that it may be clinically beneficial to outline further OARs than have historically been required for radiotherapy treatment planning (Nutting et al., 2020).

It is anticipated that this trend of increased plan complexity requiring increased planning resource will continue, therefore it is essential that further efficiency savings are introduced into planning processes. Without this there is a risk that workloads will lead to high staff stress levels and potential burnout, and to patient treatment delays.

Link to trust objectives for the Strategic Plan 2021 - 2025

By introducing such software the Trust will be ensuring the following objectives are achieved:

- Be Outstanding – Offer a high quality patient experience by implementing an efficient planning pathway.
- Be Collaborative – Driving better clinical outcomes for our patients by reducing inter and intra-observer variability of contours and ensuring timely access to radiotherapy treatment for patients.
- Be a great place to work – Reduce staff stress levels by increasing efficiency of planning process.
- Be Research Leaders – Early adoption of new cutting edge technology will allow participation in associated national and international research.
- Be Digital – Implementation of ground-breaking AI-based digital technology to improve patient outcomes.
- Be innovative – Introduction of new innovative technologies to support staff and improve patient care.

References

Chen, Z. et al. (2008). The relationship between waiting time for radiotherapy and clinical outcomes: A systematic review of the literature. Radiotherapy and Oncology, 87(1), pp.3–16.

Fritscher, K.D. et al. (2014). Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Medical Physics, 41(5), pp.1–11.

Nutting, C. et al. (2020). Results of a randomized phase III study of dysphagia-optimized intensity modulated radiotherapy (Do-IMRT) versus standard IMRT (S-IMRT) in head and neck cancer. Journal of Clinical Oncology, 38(15_suppl), p.6508. [online]. Available from: https://doi.org/10.1200/JCO.2020.38.15_suppl.6508.

Peng, Y. lin et al. (2018). Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. Oral Oncology, 82(April), pp.1–7. [online]. Available from: https://doi.org/10.1016/j.oraloncology.2018.04.025.

Sherer, M. V. et al. (2021). Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiotherapy and Oncology, 160, pp.185–191. [online]. Available from: https://doi.org/10.1016/j.radonc.2021.05.003.

3. Options Identified

Two options have been identified:

Option 1 – No change to current system

Advantages:

- No funding required
- No infrastructure required for servers

Disadvantages:

- Risk of increasing workload over time causing staff burnout, in particular for clinical oncologists.
- Risk of increase in patient start delays and missed waiting-time targets, with the possible associated negative effect on patient outcomes.
- No improvement to current levels of inter and intra-observer variability and therefore no improvement in associated patient outcomes.

Option 2 (Preferred) – Purchase of AI Auto-segmentation system

Advantages:

- Reduction of current high staff workload leading to reduced staff stress levels and potential transfer to patient contact activities.
- Reduction in patient delays and missed waiting-time targets leading to improved patient satisfaction and possible improvements to clinical outcomes.
- Reduction of current levels of inter and intra-observer variability and potential improvement in associated patient outcomes.
- Increased availability of staff resource to work on other service improvement projects.

Disadvantages:

- Funding required with no additional tariff for the activity.
- Temporary increase in staff resource required for clinical implementation.
- Associated server infrastructure required (power, network, server room rack space).

The preferred option is Option 2 - Purchase of AI Auto-segmentation system

4. Financial Profile

There is no additional funding stream for this product as it is not yet a standard of care, and consequently there is no associated tariff from NHS England. The current planning tariff is bundled and includes provision for the CT scan and the plan creation task. It is anticipated that use of such a system would eventually feed through to reference costs.

Funding for the development could be via local investment, cost savings, capital contingency or charitable funding.

Initial 12 month indicative cost is £56,650, which comprises:

£40,750	12 x month subscription for 6 x CT OAR AI Models assuming 2000 patients annually
£14,500	Hardware, Supply, Installation & configuration of software
£1,250	Software training
£150	Hardware delivery costs

Revenue Costs of Investment	£	Recurrent Costs (£)	Non-Recurrent Costs (£)
Pay	0	0	0
Non-Pay	42,150	40,750	1,400
Total Revenue Costs of Investment	**42,150**		

Capital Costs of Investment	£
Pay	0
Non-Pay	14,500
Total Capital Costs of Investment	**14,500**

Over a 5-year period the cost per patient, based on use on 2000 patients per year, would be around £22. N.B. Server hardware replacement will be required after 5 years.

For implementation, the Radiation Services Project/Change Management procedure (2 QS5), part of our ISO accredited Quality Management System will be followed. The project will be managed through the Physics Project Committee, with monthly reports to the Divisional level Radiation Services Project Committee.

No recruitment is required.

5. Recommendation

- **It is recommended the Trust chooses Option 2.**

APPENDIX D Poster discussion for ESTRO 2022

NHS
The Clatterbridge Cancer Centre
NHS Foundation Trust

A comparison of multiple deep learning-based auto-segmentation systems for head and neck cancer

ESTRO 2022

Simon Temple, Carl Rowbottom
Physics Department
The Clatterbridge Cancer Centre NHS Foundation Trust

Purpose

Commercial software which utilises deep learning (DL) can be used to automatically delineate organs art risk (OAR) with the potential for significant efficiency savings in the radiotherapy treatment planning pathway and simultaneous reduction of inter- and intra-observer variability.

Vendors of commercial systems often claim superiority of their own system in comparison to competitor systems. To date there has been limited research comparing multiple systems using multiple comparison metrics and a common patient cohort. This has been addressed in this study.

Methods

Four different DL-based auto-segmentation systems, which had been independently developed for commercial use, were used to create five commonly used head and neck (H&N) OARs (brainstem, spinal cord, mandible, left and right parotid), for 30 H&N patient datasets. All systems were running their latest available software version at the time of study (June 2021 – Sep 2021).

Auto-segmented contours were compared to 'gold standard' clinical contours, created by Oncologists at our centre using similarity metrics of 3D Dice Similarity Coefficient (DSC) and Added Path Length (APL). All data used originated from patients entered into the PATHOS[1] clinical trial. The associated trial protocol includes clear anatomical guidelines for OAR delineation and trial entry involved pre-trial OAR outlining Quality Assurance.

Results

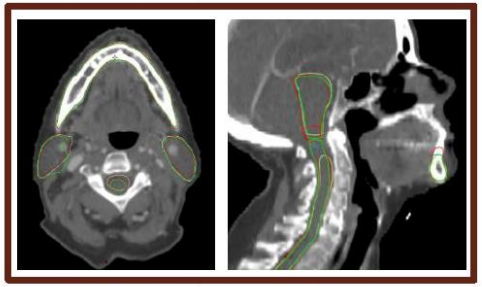

Figure 1. Axial and sagittal CT images showing resulting contours for all systems

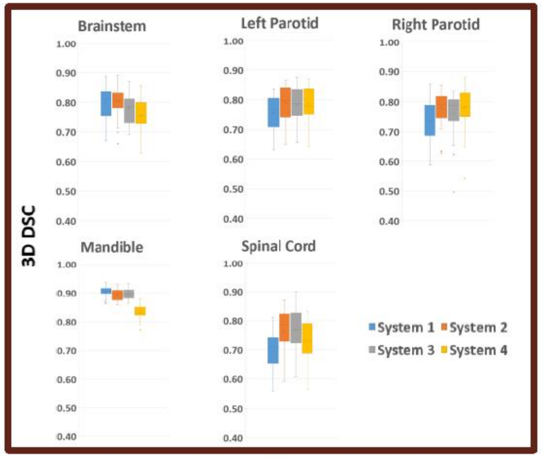

Figure 2. 3D DSC boxplots for all OARs and systems evaluated

Figure 1 shows an example of resulting contours for all systems, displayed on axial and sagittal CT images.

Figures 2 and 3 show 3D DSC and APL boxplot results for all OARs and all systems. Resulting values correlate well with other recent published studies.

Overall performance differences between the four systems were found to be statistically insignificant for both 3D DSC and APL metrics, with different systems performing best for different OARs.

Conclusions

Comparable levels of performance were observed between all four systems. This indicates that deep learning-based auto-segmentation products are developing at a similar pace in terms of the quality of contours produced.

It is therefore likely to be more beneficial to consider other factors such as financial cost and range of contours offered when considering the evaluation of such a system for clinical use.

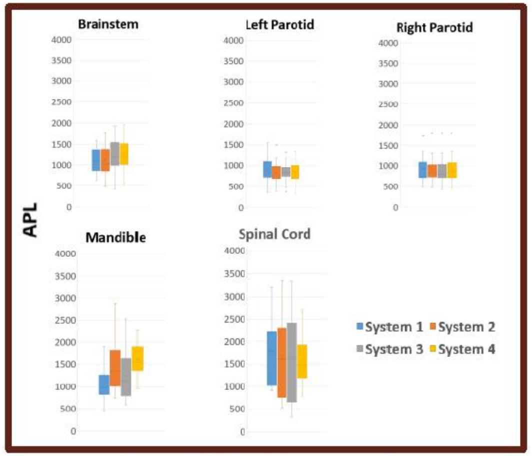

Figure 3. APL boxplots for all OARs and systems evaluated

Acknowledgments

The authors would like to thank the following system vendors for their assistance:

Varian Medical Systems Inc., Palo Alto, CA, USA

Limbus AI, Regina, Saskatchewan, Canada

Siemens Healthcare GmbH, Erlangen, Germany

Mirada Medical Ltd, Oxford, UK

References

1. Owadally, W. et al. (2015). PATHOS: A phase II/III trial of risk-stratified, reduced intensity adjuvant treatment in patients undergoing transoral surgery for Human papillomavirus (HPV) positive oropharyngeal cancer. BMC Cancer, 15(1), pp.1–10.