Artificial intelligence auto-contouring to aid radiotherapy treatment planning [GID-HTE10015]

External Assessment Group report

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

Early Value Assessment Programme

Produced by Peninsula Technology Assessment Group (PenTAG)
University of Exeter Medical School
South Cloisters
St Luke's Campus
Heavitree Road
Exeter
EX1 2LU

Authors Alan Lovell
Sophie Robinson
Brian O'Toole
Ahmed Abdelsabour
G.J. Melendez-Torres
Edward C.F. Wilson

	Peninsula Technology Assessment Group (PenTAG), University of Exeter Medical School, Exeter
Correspondence to	Sophie Robinson [redacted]
Date completed	

Contains confidential information: Yes

Number of attached appendices: 5

Purpose of the assessment report

The purpose of this External Assessment Group (EAG) report is to review the evidence currently available for included technologies and advise what further evidence should be collected to help inform decisions on whether the technologies should be widely adopted in the NHS. The report may also include additional analysis of the submitted evidence or new clinical and/or economic evidence. NICE has commissioned this work and the report forms part of the papers considered by the Medical Technologies Advisory Committee when it is making decisions about the early value assessment.

Declared interests of the authors

Description of any declared interests with related companies, and the matter under consideration. See NICE's Policy on managing interests for board members and employees.

None.

Acknowledgements

The EAG acknowledges the administrative support of Sue Whiffin and Jenny Lowe (both PenTAG) and Specialist Committee Member (SCM) input from David Bernstein (Royal Marsden Hospital), Teresa Guerrero Urbano (Guy's Hospital), Syed Moinuddin (University College Hospital London), Myur Munshi (Beatson West of Scotland Cancer Centre), Angela Pathmanathan (Royal Marsden Hospital), Jacqueline Poxon (St Bartholomew's Hospital), Carl Rowbottom (Clatterbridge Cancer Centre), Samantha Warren (Newcastle upon Tyne University Hospital).

Responsibility for report

The views expressed in this report are those of the authors and not those of NICE. Any errors are the responsibility of the authors.

Contents

Abbreviations

Term	Definition
AI	Artificial intelligence
API	Application programming interface
CE mark	*Conformité européenne* (European conformity) marking
CEA	Cost-effectiveness analysis
CI	Clinical investigator
CNS	Central nervous system
CRD	Centre for Reviews and Dissemination
CT	Computerised tomography
D	Dosimetrist
DICE	Dice similarity coefficient
DICOM	Digital Imaging and Communications in Medicine
DTAC	Digital Technology Assessment Criteria
EAG	External assessment group
ESTRO	European Society for Therapeutic Radiology and Oncology
EU	European Union
EVA	Early value assessment
GBP	British Pound
GDPR	General Data Protection Regulation
H&N	Head and neck
HD	Hausdorff distance
HTA	Health technology assessment
H&N	Head and neck
ICTRP	International Registry Platform
INAHTA	International Network of Agencies for Health Technology Assessment
MAUDE	Manufacturer and User Facility Device Experience
MDD	Medical devices directive
MDR	Medical device regulation
MeSH	Medical subject headings
MHRA	Medicines & Healthcare products Regulatory Agency
MRI	Magnetic resonance imaging
N/A	Not applicable

NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NLM	National Library of Medicine
NR	Not reported
OAR	Organs at risk
PenTAG	Peninsula Technology Assessment Group
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTV	Planning target volume
RCT	Randomised controlled trial
ReQoL	Recovering Quality of Life quality
RO	Radiation oncologist
RT	Radiotherapy
RTOG	Radiation Therapy Oncology Group
RTT	Radiation therapist
RWE	Real world evidence
SCM	Specialist Committee Member
SIGN	Scottish Intercollegiate Guidelines Network
UK	United Kingdom
UKCA	United Kingdom Conformity Assessed marking
USA	United States of America
USD	United States Dollar
VMAT	Volumetric arc therapy
WHO	World Health Organization

1. EXECUTIVE SUMMARY

Quality and relevance of clinical evidence

The findings of this rapid appraisal suggested that there is strong evidence for the potential clinical usefulness of AI-based auto-contouring, but unclear evidence around cost-effectiveness. The EAG identified a total of 79 reports that were potentially relevant to the present decision problem. Eight full text papers were predominantly prospective in design, 19 were predominantly retrospective, and 52 were conference abstracts; 73 of the reports looked at a single included technology against a relevant comparator. The remaining six reports (all conference abstracts) compared two or more of the included technologies.

Data were extracted for all eight prospective full text papers. For technologies that did not have a prospective full text paper, the highest quality retrospective full text paper was identified and extracted. This included five articles. There were no relevant full text articles for two of the technologies. For these, the EAG extracted data from a selected high quality conference abstract each. Results were therefore extracted from a total of 15 prioritised papers.

All studies had some methodological limitations or misalignment with the NICE decision problem for this appraisal. Studies were often poorly reported, and it was not always clear what the intervention consisted of, nor how they were applied. Samples sizes were often small, and the evidence overwhelmingly focused on the head and neck and the pelvis (predominately the male pelvis, for prostate cancer). The most commonly reported outcome metrics were geometric outcomes, the clinical applicability of which has been questioned by researchers.

All the studies reported either geometric, dosimetric or satisfaction scores which showed that AI-based auto-contouring creates contours, segmentations or plans similar to those created by manual contouring for most organs at risk and clinical target volumes. The majority of auto-contours were either ready to use or usable with only minor edits. However, certain organs at risk were consistently found more difficult to auto-contour, particularly those with a small volume, suggesting that auto-contouring is still best seen as an aide to radiation oncologists in their contouring work, rather than a stand-alone technology that will replace that role.

Manual contouring was the comparator for the majority of studies. In those studies that included an atlas contouring arm, the AI approach consistently created superior contours (e.g., geometrically closer to manual contours).

Quality and relevance of economic evidence

Based on clinical opinion to the EAG and evidence from published literature, AI auto-contouring (including editing and reviewing time) appears to result in time savings when compared to current organs at risk contouring approaches used in clinical practice (albeit there is considerable time variation based on tumour site and the structures typically contoured). Due to the lack of published cost effectiveness evidence and heterogeneity in the pricing and reimbursement strategy for each technology, it was not possible to draw firm conclusions on the cost effectiveness of AI auto-contouring compared to manual or atlas-based segmentation approaches.

Evidence Gap Analysis

More robust evidence is required for the following: 1) AI intervention costs in clinical practice and how these technologies impact on healthcare resource use and/or patient outcomes in a UK context. 2) The impact of local NHS training sets rather than an “off the shelf” approach, from both a clinical and harmonisation/cost-effectiveness point of view. 3) AI auto-contouring effectiveness for body structures beyond head and neck and the pelvis, and identification of those organs at risk that are particularly susceptible to being poor contoured. 4) Relative clinical and cost-effectiveness of auto-contouring using MRI vs CT scans. 5) Direct, head-to-head trials between alternative AI auto-contouring technologies.

.

2. DECISION PROBLEM

Table 1 details the final scope issued by NICE for this EVA, defined per element of assessment.

Table 1: Summary scope of the assessment

Population	People having radiotherapy treatment planning for external beam radiotherapy
Interventions (proposed technologies)	AI auto-contouring technologies for initial treatment planning, namely: • AI-Rad Companion Organs RT (Siemens Healthineers) • ART-Plan (TheraPanacea, Oncology Systems) • AutoContour (Radformation) • DLCExpert (Mirada Medical) • INTContour (Carina Medical) • Limbus Contour (Limbus AI, AMG Medtech) • MIM Contour ProtégéAI (MIM Software) • MRCAT Prostate plus Auto-contouring (Philips) • MVision Segmentation Service (MVision AI Oy, Xiel) • OSAIRIS (Cambridge University Hospitals NHS Foundation Trust) • RayStation (RaySearch Laboratories AB)
Comparators	Contouring methods used in standard care to contour OAR and target volumes including lymph nodes. These include: • manual contouring • atlas-based contouring • model-based segmentation. Comparators may also include ‘no contours or no contouring’ for cases where AI auto-contouring may generate contours for structures not routinely contoured in standard care.
Healthcare setting	Outpatient settings

Outcomes	The outcome measures to consider include: **Accuracy and acceptability** • Clinical acceptability of contours including alignment with national and international guidelines • Accuracy of contours including quantitative measures of DICE coefficient and qualitative measures • Degree of contour edits needed before use in radiotherapy treatment planning • Consistency of contours including interrater reliability • Impact on radiotherapy treatment planning quality assurance including surrogate, qualitative and quantitative measures such as: o Dose prescription changes o Dose volume distributions o Radiation toxicity o Missing targets o Adherence to international guidelines • Usability, user experience and satisfaction **Resource and system impact** • Contouring time including time needed for healthcare professional review and manual edits • Radiotherapy treatment planning time including time saved and difference in time to start of treatment • Number of more complex plans produced including number of structures contoured • Impact on staffing and treatment planning resources, such as changes in skill-mix or healthcare professional grade needed to produce and review contours • Impact of the system on clinical oncology training (including training of all healthcare professionals contributing to radiotherapy treatment planning) • Impact on healthcare professional performance and productivity more broadly, such as efficiency, increase in patient-facing tasks and staff wellbeing.

	Costs will be considered from an NHS and Personal Social Services perspective. Costs for consideration should include: • Costs of AI auto-contouring software including installation, licence fees, maintenance and update costs for additional libraries or features • Costs of any associated technology needed to use AI auto-contouring tools excluding capital costs for equipment that is otherwise used in standard care • Healthcare professional grade and time • Cost of other resource use such as additional appointments or healthcare professional training
Time horizon	The time horizon for estimating the clinical and economic value should be sufficiently long to reflect any differences in costs or outcomes.

Abbreviations: AI = artificial intelligence. DICE = Dice similarity coefficient. NHS = National Health Service. OAR = organs at risk.

3. OVERVIEW OF THE TECHNOLOGY

3.1. Purpose of the medical technology

External beam radiotherapy uses ionising radiation to kill cancer cells in the treated area. It aims to give a high dose of radiation to cancer cells but as low a dose as possible to nearby healthy cells. Contouring is an important part of the radiotherapy treatment planning process. It involves outlining the target volumes and organs at risk (OAR) to guide radiotherapy so that treatment is effective and radiation toxicity is reduced. AI auto-contouring technologies aim to improve contouring efficiency by automatically contouring the organs at risk, with some also contouring the target volumes before radiotherapy. Most of these technologies have been trained using deep learning convolutional neural networks (a type of artificial intelligence learning algorithm) to process images from CT or MRI scans and produce an initial contour. Images and contours are then reviewed by trained healthcare professionals and modified as needed.

During the scoping process, clinical experts advised that they spend a lot of time creating and reviewing manual contours. AI auto-contouring, with healthcare professional review, may be quicker than manual or atlas-based contouring (see section 4 for more details about manual and atlas-based contouring), and hence reduce costs by reducing healthcare professional time needed to do contouring. It may also improve consistency of contouring between people, standardise processes and improve adherence to international guidelines. Clinical experts considered that AI auto-contouring could result in quicker radiotherapy treatment planning pathways and shorter time to treatment for patients. The Royal College of Radiologists clinical oncology census report 2021[1] reports workforce pressure because of staff shortages and continued effects from the COVID-19 pandemic. Increased efficiency from using AI auto-contouring may increase capacity, allow healthcare professionals to focus on patient-facing tasks and reduce waiting lists. During the EVA scoping process, experts advised that potential improvements in consistency may lead to more accurate contours and could reduce unwanted variation and outliers which could reduce toxicity.

3.2. Product properties

This EVA focuses on AI auto-contouring technologies for radiotherapy treatment planning. It includes 11 technologies that:

- use AI-based algorithms to automatically contour organs at risk (OAR) or target volumes as part of initial radiotherapy treatment planning
- are standalone AI auto-contouring software or have AI auto-contouring functionality integrated in treatment planning or radiotherapy platforms
- meet the standards within the digital technology assessment criteria (DTAC), including the criteria to have a CE or UKCA mark where required. Products may also be considered if they are actively working towards required CE or UKCA mark and meet all other standards within the DTAC
- are available for use in the NHS.

Technologies for this assessment were identified through NICE topic intelligence, NHS stakeholders and clinical experts, and literature search. The scope of this assessment excludes adaptive radiotherapy systems. Experts advised that these technologies would likely have a different care pathway and should be evaluated separately. The scope also excludes bespoke AI auto-contouring technologies developed in-house by local services using open-source software such as Inner Eye project by Microsoft.

SCMs noted that there is considerable heterogeneity between the technologies being assessed. Some of the included technologies are “stand alone” systems (such as from Limbus and MVision), meaning that they are hardware agnostic. Others, such as MRCAT, use in-built systems, meaning that they’re designed to be used with a specific hardware platform. A number of the technologies focus just on contouring, while some are part of a suite of software (such as DLCExpert) that include other components of the radiotherapy planning pathway.

Other differences include how certain technologies focus on a relatively narrow band of structures (for example MRCAT Prostate plus). Various different datasets and consensus guidelines will have been used to train the algorithms—this is often poorly reported in company literature—and there are also differences in the degree to which a technology can be locally trained with datasets from a users’ institution (INTContour is very customisable, for example). A number of the technologies are well established, while others are more recent additions to the field, and are awaiting their CE Mark (for example AutoContour and OSAIRIS). Finally, most companies are continually updating and releasing auto-contouring models, as well as developing new models, meaning that most of the literature will have compared algorithms which differ from those found in the technologies today. Such heterogeneity makes direct comparison difficult.

Technologies are described in brief in Table 2. Further information can be found in the Final Scope.[2]

Table 2: Description of the technologies

Full technology name and manufacturer	Stand-alone or in-built system	Training guidelines used	Structures contoured	CE mark and DTAC status	Implementation options and limitations
AI-Rad Companion Organs RT (Siemens Healthineers)	Standalone	Trained using RTOG guidelines	Contours over 60 organs at risk on CT scans including abdomen, head and neck, pelvis and thorax. The next version (VA50), rolling out in 2023, will also contour 8 organs at risk using MR images.	CE-marked class IIb medical device under the EU medical devices regulation (MDR) DTAC application is being considered for the software	Designed to be used with treatment planning systems and interactive contouring applications. Is part of a ”family” of companion software for various body regions, including for brain, chest, and prostate
ART-Plan (TheraPanacea, Oncology Systems)	Standalone	Trained using international guidelines such as such as ESTRO and RTOG	Contours over 150 organs at risk and lymph nodes including abdomen, brain, head and neck, thorax and pelvis on CT images and abdomen, brain and male pelvis on MRI	CE-marked class IIb medical device under the EU MDR No information on DTAC status	Provided using installation files transferred to the server that will host the software. If the cloud option is used it is provided using installation files that allow access to the software installed in the cloud server located in that region
AutoContour (Radformation)	Standalone	Trained using consensus guidelines	It contours over 200 structures including organs at risk and lymph node regions in the chest and abdomen, head and neck, and pelvis on CT images and brain on MRI	Currently undergoing regulatory approval with a notified body for CE-marking as a class IIa medical device. DTAC application in process	It has DICOM standalone capability and can also be integrated with Varian Eclipse using the Eclipse Scripting application programming interface (API).

Full technology name and manufacturer	Stand-alone or in-built system	Training guidelines used	Structures contoured	CE mark and DTAC status	Implementation options and limitations
DLCExpert (Mirada Medical)	In-built DLCExpert is deployed on Mirada Medical's Workflow Box platform, which is a software application designed to perform automated workflows.	Trained using consensus guidelines	DLCExpert contours over 160 structures on CT and MRI images, including abdomen, breast, head and neck, prostate and thorax.	CE-marked class I medical device under the EU medical devices directive (MDD). No information on DTAC status	It is designed to be used with existing treatment planning or image processing software.
INTContour (Carina Medical)	Standalone	Not reported In addition to built-in protocols, individual institutions can train their own AI models in house	Delineates organs on CT or MRI images. It contours over 60 target and organs at risk structures from abdomen, head and neck, male pelvis and thorax.	Regulatory approval for use in the UK is expected in 2023. Company plan to apply for a DTAC before 09/2023	INTContour can be accessed using the web-based interface and DICOM tools. Users can create and use customised models. It can also be integrated with Varian Eclipse and RayStation treatment planning systems.
Limbus Contour (Limbus AI, AMG Medtech)	Standalone It is locally hosted and can be installed on any existing hardware without the need for a graphics processing unit	Developed in line with international consensus guidelines	It contours over 200 organs at risk and target volumes including lymph nodes, abdomen, breast, central nervous system, head and neck, lung, pelvis and prostate on CT images, and	CE-marked class I medical device under the EU MDD. DTAC has been completed for a number of customers	It is vendor neutral which means DICOM (digital imaging and communications in medicine) files can be sent to the existing treatment planning system or

Full technology name and manufacturer	Stand-alone or in-built system	Training guidelines used	Structures contoured	CE mark and DTAC status	Implementation options and limitations
	(GPU) or cloud connection.		central nervous system, gynaecologic and brachy structures on MRI		workstation for review and clinical validation.
MIM Contour ProtégéAI (MIM Software)	Standalone	Not reported	Contours organs at risk and sensitive structures from CT or MRI images. It contours head and neck, thorax, lungs and liver, prostate and abdomen structures from CT images and prostate from MRI.	CE-marked class IIa medical device under the EU MDD. Currently applying for DTAC approval	Image data are sent from the hospital picture archiving and communication system (PACS) or local planning system to MIM software for contouring before being saved as DICOM RT structures. Healthcare professionals can manually correct contours before sending to treatment planning systems. MIM Contour ProtégéAI is vendor neutral, and installation can be customised to service needs.
MRCAT Prostate plus Auto-contouring (Philips)	In-built A clinical application integrated in Philips Ingenia system for magnetic resonance	Not reported	Prostate contouring	CE-marked No DTAC application made	MRCAT images conform to DICOM standards and can be exported to treatment planning systems. The company said that the system can replace traditional CT-based workflows with an MRI only

Full technology name and manufacturer	Stand-alone or in-built system	Training guidelines used	Structures contoured	CE mark and DTAC status	Implementation options and limitations
	imaging in radiation therapy (MR-RT).				radiotherapy workflow from imaging and planning to position verification.
MVision Segmentation Service (MVision AI Oy, Xiel)	Standalone	Trained to comply with international guidelines using a peer-reviewed process.	CT or MRI. It contours over 160 structures including organs at risk and target volumes in abdomen and thorax, brain, breast, head and neck, and pelvis.	It is a CE-marked class I medical device under the EU MDD. MVision has received approval of several DTAC submissions to various NHS trusts	Images from the scanner or treatment planning system are exported to MVision. A structure set is created, and contours are added to the original images. These are then sent to the DICOM folder or treatment planning system.
OSAIRIS (Cambridge University Hospitals NHS Foundation Trust)	Standalone	The system was trained with data from the developer's hospital (Cambridge University Hospitals NHS Foundation Trust)	It contours up to 26 head and neck and prostate treatment site structures on CT images	Regulatory approval for use in the UK is in progress No information on DTAC status	OSAIRIS is an open-source standalone AI auto-contouring software. It is a cloud-based workflow acceleration technology, designed for free use and sharing within the NHS. It complies with the NHS Azure Blueprint.
RayStation (RaySearch)	In-built	Trained using multiple reference CT and/or MR image datasets with the patient structures contoured on them.	It contours over 70 structures on CT images including breast and lymph nodes, head and neck, male pelvis, thorax and abdomen.	It is a CE-marked class IIb medical device under the EU MDD No DTAC application made	RayStation is a radiotherapy external beam and brachytherapy planning system with AI auto-contouring functionality included as part of the standard contouring tools.

Abbreviations: AI = artificial intelligence; API = application programming interface; CE = Conformité européenne (European conformity); CT = Computerised tomography; DICOM = Digital Imaging and Communications in Medicine; DTAC = Digital Technology Assessment Criteria; ESTRO = European Society for Therapeutic Radiology and

Oncology; MDD = medical devices directive; MDR = Medical device regulation; MRI = Magnetic resonance imaging; NHS = National Health Service; RTOG = Radiation Therapy Oncology Group.

3.3. Status quo

AI auto-contouring would be used as an alternative to manual or atlas-based contouring or model-based segmentation as part of standard care radiotherapy treatment planning. For some cases, AI auto-contouring may generate contours for structures that are not routinely produced in standard care. In these instances, no contours or no contouring may be an appropriate comparator to consider.

4. CLINICAL CONTEXT

The target population for this assessment is people having radiotherapy treatment planning for external beam radiotherapy.

4.1. Care pathway

Contouring in radiotherapy treatment planning is used to outline the target volume and organs at risk to guide radiotherapy so that treatment is effective and radiation toxicity is reduced. Healthcare professionals most often use manual or atlas-based contouring or model-based segmentation. Manual contouring is the most common contouring method used in standard care. Manual contouring of target regions is usually done by clinical (radiation) oncologists, while contouring of organs at risk may also be done by clinical technologists, dosimetrists, or therapeutic radiographers. There are published guidelines for contouring organs at risk and disease sites from organisations such as European Society for Radiotherapy and Oncology[3] and the Royal College of Radiologists.[4] Atlas-based contouring and model-based segmentation are not as widely used in standard care. Atlas-based contouring is an automated method that contours new images using models based on historical images of similar patient anatomy. Model-based segmentation is also an automated method that contours images using statistical shape models for different organ structures. Contours regardless of contouring method should be reviewed before being used in treatment planning in line with guidance such as the Royal College of Radiologists guidance on radiotherapy, target volume definition and peer review.[4]

It is expected that AI auto-contouring would be used as part of standard care radiotherapy treatment planning. Radiotherapy is usually given in hospital on an outpatient basis. AI auto-contouring would be reviewed and edited as needed by trained healthcare professionals, including clinical oncologists, therapeutic radiographers, clinical technologists and medical physicists. It is recommended that all contours should be reviewed and modified as needed before being used in treatment planning.

4.1.1. Current use of AI auto-contouring to aid radiotherapy treatment planning

Five of the eleven companies covered in this review advised in their company submissions to NICE that their respective AI auto-contouring technologies are in current use within the NHS for radiotherapy treatment planning:

[redacted]

SCMs reported a range of experience in using AI auto-contouring technologies. Some SCMs have used AI auto-contouring for research purposes only; some use this regularly as part of current clinical practice.

4.2. User issues and preferences

SCMs reported enjoying using AI-based auto-contouring technology and thought that it would eventually save time and facilitate standardisation in radiotherapy. Some stated that with more use of AI auto-contouring. expertise and resources would be better employed in critically reviewing contours rather than in laborious manual delineation. Nevertheless, SCMs also described challenges when auto-contouring structures where there was multiple (or no) consensus on definition (for example, different protocols exist for inguinal nodes, or for head neck lymph nodes). They advised that it is therefore unlikely that AI will be acceptable to all users, at least in the first instance.

Companies also offered their view of the technology in their submissions. They anticipated that AI auto-contouring technologies should increase accuracy in organ contouring and improve quality of care, and that use of the technologies may facilitate standardisation of treatment across the NHS. It was also often advised that technologies could lead to a more efficient use of staff time, and reduce the pressures of staff shortages, thus improving waiting times. Some companies described these technologies as an adjunct to standard care and did not anticipate that the patient experience would be affected.

5. SPECIAL CONSIDERATIONS, INCLUDING ISSUES RELATED TO EQUALITY

The following issues were highlighted during the scoping process. No new issues were identified during EAG assessment.

AI models can contain algorithmic bias depending on the population used in training. Populations used in training datasets may not be representative of patient populations in clinical practice which can cause potential age, gender, disability and ethnic bias. Clinical experts advised that there is a potential for gender bias, for example a lack of representation of the female pelvis and male breast cancer in some training datasets. There is also a potential for disability bias, for example not including people with hip replacements in training datasets. Training datasets may also underrepresent children and young people. This may affect the performance of AI auto-contouring for these populations. AI auto-contouring may perform best with certain CT or MRI sequences or with the person being in a specific position such as supine head-first. Training datasets may not include data on atypical positioning which may make AI auto-contouring less accurate for some people with limited mobility. Clinical experts advised that AI auto-contouring may also not work as well for people with atypical anatomy associated for example with previous medical interventions such as surgery.

To mitigate these issues, the potential risk of bias of a specific technology should be considered when deciding if to use that technology in research or clinical settings. This should form part of a local assessment process before purchase and clinical decision-making. Companies should also provide detailed information on training datasets as part of their product information pack, including guidelines used and demographics such as age range, gender ratios and inclusion of disabilities.

Cancer is considered a disability under the Equality Act 2010. Incidence rates in the UK for all cancers combined are highest in people aged 85 to 89 with more than a third of diagnoses each year being in people aged 75 and older. Age and disability are protected characteristics under the Equality Act 2010.

6. POTENTIAL IMPLEMENTATION ISSUES

The NICE adoption and implementation team consulted clinical experts and noted several potential implementation issues. When deciding whether to use AI auto-contouring technologies, radiation oncology services should consider:

- compliance with GDPR, information governance and cybersecurity standards
- staff acceptability of AI auto-contouring including ease of implementation considering workforce skills and workflows
- education and training needed to use the specific technology
- how automated decision-making fits into local protocols
- bandwidth and server requirements
- monitoring performance, risk assessment and quality assurance.

Experts and stakeholders outlined several considerations for using AI auto-contouring technologies in the NHS. AI auto-contouring technologies should:

- conform to national and international guidelines
- come with detailed information on training datasets used, software optimisation and validation
- be DICOM compatible (DICOM stands for Digital Imaging and Communications in Medicine and is the standard for the communication and management of medical imaging information and related data[5])
- be vendor neutral and able to integrate into current workflows easily and automatically
- include all relevant organs at risk and targets including additional structures not manually contoured
- be customisable, such as which structures to include, structure names, colours and fit with local protocols.

7. CLINICAL EVIDENCE SELECTION

7.1. Search strategy

Search strategies were based on those devised during the initial scoping searches by NICE Information Services with some amendments. The search strategies used relevant search terms, comprising a combination of indexed keywords (e.g., Medical Subject Headings, MeSH) and free-text terms appearing in the titles and/or abstracts of database records and were adapted according to the configuration of each database. No date, language or publication status (published, unpublished, in-press, and in-progress) limits were applied. Searches for clinical and cost-effectiveness were combined and carried out in one search strategy.

Databases searched were Medline (including Medline in Process), Embase, Cochrane, INAHTA, CEA Registry and ScharrHUD. Additional trial registries searched were Clinicaltrials.gov (NLM) and ICTRP (WHO). The websites of the individual companies were searched; NICE and SIGN websites were searched for related guidelines, and MAUDE and MHRA were searched for adverse events data. Following deduplication (in Endnote), a total of 933 records of potentially relevant evidence on clinical and/or cost effectiveness were retrieved. The company submission references were also scanned for additional references—from which five new articles were identified.

The search strategies are presented in Appendix A.

7.2. Study selection

The abstracts and titles of references retrieved by the searches were screened for relevance (facilitated by the Rayyan platform). Full paper copies of potentially relevant studies were obtained. The retrieved articles were assessed for inclusion against pre-specified inclusion/exclusion criteria. At each stage of screening, a minimum of 10% of records were independently screened by a second reviewer. Discrepancies were resolved by discussion, with involvement of a third reviewer, where necessary. All duplicate papers were excluded.

This assessment looked across a range of evidence types, including RCTs and real-world evidence, to inform clinical effectiveness.

The following study types were excluded:

- Animal models
- Pre-clinical and biological studies
- Narrative reviews, editorials, opinion pieces
- Meeting abstracts for studies where full-text papers were available. If studies were only available as meeting abstracts, inclusion depended on sufficient information being available to offer meaningful critique.
- Studies not available in the English language.

Eligible studies assessed a scoped intervention in a population of people having radiotherapy treatment planning for external beam radiotherapy.

Studies were included if the comparator did not match the scope or if the outcomes did not match the scope, provided the outcomes appeared reasonable and could offer useful information in the context of the appraisal. The EAG's general approach was one of 'best evidence synthesis', focusing on the most useful and rigorous evidence available over all possible included studies. Because of the large number of included technologies, the EAG focused on prospective studies where they were available. At least one full text article or abstract was identified for detailed assessment for each technology. This was supplemented with additional data from other studies where it was considered appropriate.

Where no prospective studies were available for a given technology, the most relevant retrospective studies were sought. If no retrospective studies were available either, then conference abstracts were reviewed. If retrospective studies were available for a technology with one or more prospective studies, a brief commentary on these were provided.

A PRISMA flow diagram is provided as Appendix B.

Data were extracted from included studies by one reviewer into a bespoke database and a sample of at least 10% was checked by another reviewer. Due to time and resource constraints associated with conducting an EVA, the EAG did not conduct formal risk of bias assessment of the included studies. Informally, studies were prioritised based on a) study design (prospective/retrospective), b) currency and c) sample size. Blinding was also noted, where applicable (depending on the study design and outcomes assessed).

8. CLINICAL EVIDENCE REVIEW

The EAG identified a total of 79 reports that were potentially relevant to the present decision problem. Eight full text papers were predominantly prospective in design, 19 were predominantly retrospective, and 52 were conference abstracts; 73 of the reports looked at a single included technology against a relevant comparator. The remaining six reports (all conference abstracts) compared two or more of the included technologies. Table 3 presents an overview of the evidence landscape.

Data was extracted for all eight prospective full text papers. This covered four technologies: DLCExpert, Limbus Contour, MIM Contour ProtégéAI, and MRCAT Prostate plus Auto-contouring. For technologies that did not have a prospective full text paper, the EAG extracted the highest quality and most relevant–based on currency and sample size—retrospective full text paper (reasons for selection for each of the extracted papers are provided in section 8.5). This included five articles, covering five technologies: AI-Rad Companion Organs RT[*], INTContour, MVision Segmentation Service, OSAIRIS, and RayStation. For the last two technologies, ART-Plan and AutoContour, the EAG extracted data from a selected high quality conference abstract each. Data was therefore extracted from a total of 15 prioritised papers. These are underlined and in bold in Table 3.

Table 4 presents a detailed overview of the study design, characteristics, and limitations of each prioritised study.

[*] The EAG notes that Ginn 2023 had a mixed prospective and retrospective design.

Table 3: Evidence landscape

Technology	Prospective studies (full text)	Retrospective studies (full text)	Conference abstracts
AI-Rad Companion Organs RT		**Ginn 2023**[6*] Hu 2023[7] Marschner 2022[8]	Peng 2022[9] Ginn 2022[10] Maduro Bustos 2022[11]
ART-Plan			**Blanchard 2020**[12] Nachbar 2021[13] Buatti 2022[14] Costea 2021[15] Rivera 2020[16] Gregoire 2020[17]
AutoContour			**Leyva 2022**[18] Bice 2022[19] Marasco 2022[20]
DLCExpert	**Hague 2020**[21] **Van Dijk 2020**[22] **Vaassen 2021**[23]	Walker 2022[24] Vaassen 2022[25] Brouwer 2020[26] Brunenberg 2020[27] Lustberg 2018[28]	Van de Glind 2022[29] Alty 2022[30] Boukerroui 2022[31] Vaassen 2021[32] Gibbons 2021[33] Geng 2021[34] Brunenberg 2020[35] Hague 2020[21]

[*] Note that the Ginn 2023 study has both prospective and retrospective components (see Table 4 for more details)

			South 2020[36] Liu 2020[37] Poortmans 2019[38] Lustberg 2018[28] Aljabar 2018[39] Bakker 2018[40] Gooding 2018[41]
INTContour		**Duan 2022[42]** Chen 2020[43]	
Limbus Contour	**Radici 2022[44]** **Wong 2021[45]** **Wong 2020[46]**	Wong 2021[47] D'Aviero 2022[48] Zabel 2021[49]	Kirkby 2022[50] Kucharczyk 2022[51] Coughlan 2022[52] Wong 2020[53] Wong 2020[54] Wong 2019[55] Fong 2019[56] Wong 2019[57]
MIM Contour ProtégéAI	**Urago 2021[58]**	Lastrucci 2022[59]	Lancellotta 2022[60] Tsai 2022[61] Martinez 2022[62] Kruzer 2020[63] Cole 2020[64] Halley 2020[65]
MRCAT Prostate plus Auto-contouring	**Kuisma 2020[66]**		Maspero 2018[67]
MVision Segmentation Service		**Strolin 2023[68]** Kiljunen 2020[69]	Suresh 2021[70] Heikkila 2020[71]
OSAIRIS		**Oktay 2020[72]**	

RayStation		**<u>Almberg 2022</u>**[73] Rigaud 2021[74]	Liu 2022[75] Sidorski 2021[76]
Multiple tech comparison			Borkvel 2022[77] Rong 2022[78] Liao 2022[79] Yuan 2022[80] Gorgisyan 2022[81] Doolan 2021[82]

Bold and underlined text = extracted study (see Table 4)

Table 4: Study design and characteristics of prioritised clinical effectiveness studies

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
AI-Rad Companion Organs RT (number of prioritised studies = 1)						
Ginn 2023[6]	Part retrospective and part prospective USA	Retrospective 100 cases H&N (x46) Pelvic (x56) Prospective 20 cases H&N (x10) Pelvic (x10)	Auto-contours were generated by Siemens using the stand-alone automatic contouring application (which does not allow any user specific configuration).	Retrospective: original clinical contours used to benchmark the automatic contouring software were generated by several different dosimetrists Prospective: manual contouring was performed by a trained medical dosimetrist. All patients and contouring tasks were assigned in a random order to mitigate bias from having previously seen either the automatic or manual contours	Time-saving metrics (contouring time, including the time required to edit automatic contours) Qualitative assessment (scale scoring of clinical acceptability by physicians on a 4-point scale) Geometric analysis (DICE, HD, Jaccard)	No information on how patients were selected. No comparative element to the qualitative assessment. The timing component of the study only used a small subset of the total dataset.
ART-Plan (number of prioritised studies = 1)						
Blanchard 2020[12]	Abstract (prospective) France	100 cases H&N	Auto-contours were generated by ART-Plan. Two subsets were created:	Manual contouring was performed by five or two experts, depending on the OARs	Qualitative assessment (scale scoring of clinical usability by experts, including intra- and inter-observer	Abstract only means that reporting is brief hence there is uncertainty considerable over

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
			Auto (v.1.0): trained using 6,000 cases per organ. Auto (v.2.0): trained using 21,000 cases per organ.		rating between the two datasets) Geometric analysis (DICE, HD)	exact methods and evaluation.
AutoContour (number of prioritised studies = 1)						
Leyva 2022[18]	Abstract (retrospective) USA	224 structures H&N	Auto-contours were generated by AutoContour. No local training was reported.	Clinically approved organ structures	Geometric analysis (DICE, mean surface distance)	Abstract only means that reporting is brief hence there is uncertainty considerable over exact methods and evaluation.
DLCExpert (number of prioritised studies = 3)						
Hague 2021[21,83]	Prospective UK	38 cases H&N	Auto-contouring models were generated using DLCExpert: 1) CT models (x2) 2) Diagnostic MRI model 3) Planning MRI model 4) MR-Linac model The CT model was trained on 72 local	Manual contouring was performed by a clinician using MR scans	Qualitative assessment (scale scoring of goodness of fit by independent observers on a 7-point scale) Geometric analysis (DICE, distance to agreement)	The MRI model was trained and tested on a small dataset from a single institution. No information on how scans used for testing were selected. No blinding during assessment is reported.

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
			scans and 549 vendor scans. The MRI model was trained on 100 and validated on 28 datasets.			
Van Dijk 2020[22]	Prospective The Netherlands	104 H&N	Auto-contours were generated by DLCExpert, trained on 589 local head and neck cancer patients.	1. Manual contouring was performed by a dedicated team of experts according to international consensus delineation guidelines. Clinically available atlas contours were often used as a basis for the contouring. 2. Atlas contouring was performed by WorkflowBox 1.4, Mirada Medical, designed using a representative set of 30 H&N patients taken from the training set.	Time-saving metrics (time for two RO's—an expert and a beginner—to adjust contours as necessary to make them suitable for clinical use) Qualitative assessment (blinded testing via a Turing test of ability to distinguish auto-contours from human generated contours) Dosimetric analysis (Dose constraints were computed and compared)	The basis for manual contouring was often atlas contours, meaning that the study may be biased towards atlas contouring. The time evaluation, Turing test, and inter-observer evaluation were performed on a subset of the sample.

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
					Geometric analysis (DICE, HD, absolute dose difference	
Vaassen 2021[23]	Prospective The Netherlands	20 cases Non-small cell lung cancer	Auto-contours were generated by a prototype deep-learning contouring method (DLCExpert), followed by manual adjustment. No algorithm training was reported.	1. Manual contouring (and adjustments to AI and Atlas contours) was performed by one experienced RTT. 2. Atlas contouring was performed by the commercial atlas-based method (Embra-ceCT, Mirada Medical)	Dosimetric analysis (comparison of dose-volume histograms between manual, atlas, auto, and auto + adjusted contours, including estimation of changes in dose distributions on treatment plans) Geometric analysis (DICE, HD)	Small sample size and relatively small number of outcomes measured. All treatment plans were optimized by a "knowledge-based planning model"; it is unclear how this may have impacted the findings.
INTContour (number of prioritised studies = 1)						
Duan 2022[42]	Retrospective USA	23 cases Prostate	Auto-contours were generated by INTContour, trained on 84 local cases.	Manual contouring was performed by a resident RO, and then were reviewed and modified by an RO with 20 years of clinical experience. These were then further reviewed by a	Qualitative assessment (scale scoring of clinical acceptability by an RO on a 5-point scale) Dosimetric analysis (Dose constraints	Small sample size and a small training set. A single RO performed all the manual contouring, which may have led to errors in the reference contours.

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
				third RO with 10 years' experience.	were computed and compared) Geometric analysis (DICE, HD, mean surface distance, inter-observer variability analysis)	
Limbus Contour (number of prioritised studies = 3)						
Radici 2022[44]	Prospective Italy	12 cases H&N (x3) Prostate (x3) Rectum (x3) Breast (x3)	Auto-contours were generated by Limbus Contour (using the same CT scans as manual contouring), reviewed by the competent RO and, if necessary, the contours were modified. No local training of the algorithm was reported.	Manual contouring was performed by four different RO's, each with expertise in the specific clinical setting, following national and international consensus guidelines. Additional imaging was used, if necessary.	Time-saving metrics (contouring time; absolute and relative differences) Dosimetric analysis (differences in dose volume histograms) Geometric analysis (DICE, volume variation, centre of mass shift)	Very small sample sizes. No information on how patients were selected.
Wong 2021[45]	Prospective Canada	~606 RT plans. A selection of which were assessed via scale scoring and geometric analysis:	Limbus Contour auto-segmentation software version 1.0.22 was implemented at two centres. Generated contours underwent manual review and were	Unedited auto-contours were compared to the edited treatment approved contours	Qualitative assessment (scale scoring of the degree of edits required by RTT's/D's and RO's on a 5-point scale)	Relatively limited outcome metrics evaluated. No independent controls (i.e. no manual contours, and hence no

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
		CNS (x27) H&N (x54) Prostate (x93)	edited as needed prior to being used for treatment planning.		Geometric analysis (DICE, HD)	comparative component for the qualitative assessment). Only a relatively small proportion of included plans were formally assessed, which may bias the outcomes.
Wong 2020[46]	Prospective Canada	60 cases CNS (x20) H&N (x20) Prostate (x20)	Auto-contours were generated using the auto-segmentation models in Limbus Contour. There was no local training of the algorithm.	Manual contouring was performed by volunteer ROs using the same scans as used for AI contouring: three ROs for CNS, four for H&N, and three for prostate. ROs contoured according to their training and clinical judgement without viewing pre-existing contours.	Time-saving metrics (contouring time) Geometric analysis (DICE, HD)	Small sample size for each structure. Only certain patients were eligible for inclusion, and patient selection was unclear. Limited number of outcomes metrics.
MIM Contour ProtégéAI (number of prioritised studies = 1)						
Urago 2021[58]	Prospective Japan	51 cases H&N (x30) Prostate (x21)	Auto-contours were generated by MIM Contour ProtégéAI (ver. 0.9). No local training of the algorithm occurred.	1. Manual contouring was performed by three RO's for patients with prostate cancer, and five RO's for patients with H&N cancer.	Time-saving metrics (delineation time) Qualitative assessment (visual	Small sample size for each structure. Uncertainty over how patients were selected.

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
				2. Atlas contouring was performed by the commercial software MIM Maestro (ver. 7.0.3)	evaluation of errors by RO's) Geometric analysis (DICE, HD, mean distance to agreement	
MRCAT Prostate plus Auto-contouring (number of prioritised studies = 1)						
Kuisma 2020[66]	Prospective Finland	65 cases Prostate	Auto-contours were generated by MRCAT after all standard manual delineations had been finished. No local training occurred.	Manual contouring was performed by the clinical investigator (CI) RO's. The CI was blinded to the structures contoured by the RO, and vice versa	Geometric analysis (DICE, HD, absolute volume difference, centre of mass shift)	Limited range of (only geometric) outcome metrics collected. A radiotherapist inspected the auto-contours for outliers (e.g. cases where auto-contouring clearly mis-performed). It is unclear how this may have impacted the findings.
MVision Segmentation Service (number of prioritised studies = 1)						
Strolin 2023[68]	Retrospective Italy	111 cases H&N (x20) Breast (x20) Abdomen (x21)	Auto-contours were generated by MVision and manually adjusted if necessary. The time between manual contouring and	Manual contouring was performed by at least one senior and two in-training ROs, according to institutional protocols.	Time-saving metrics (contouring time, including time for manual	Small sample size for each structure. Uncertainty on how patients were selected.

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
		Thorax (x20) Male pelvis (x20) Female pelvis (x10)	modifying auto-contours was at least six months to remove potential bias. No local training of the algorithm occurred.	An Atlas-based approach was permitted only for lung segmentations	adjustment of auto-contours) Qualitative assessment (scale scoring of level of satisfaction by RO's on a 5-point scale) Geometric analysis (DICE, mean distance agreement)	
OSAIRIS (number of prioritised studies = 1)						
Oktay 2020[72]	Retrospective UK	178 cases Pelvic male (x132) H&N (x46)	Auto-contours were generated by an AI model trained using a subset of the data. For pelvis the algorithm was trained on 345 scans and validated on 42 scans. For head and neck it was trained on 176 scans and validated on 20 scans.	Manual contouring was performed by one expert and later reviewed by another oncologist	Time-saving metrics (contouring time) Geometric analysis (DICE, HD, mean surface-to-surface distance)	Uncertain how patients were selected. Limited number of outcomes metrics.
RayStation (number of prioritised studies = 1)						
Almberg 2022[73]	Retrospective Norway	30 cases Left-sided breast cancer	Deep learning segmentation models were trained by	Manual contouring. The heart, left anterior descending artery, and	Qualitative assessment (scale scoring of level of	Uncertain how patient datasets were selected.

Reference	Study design & country	Sample & structures	Intervention	Comparator(s)	Outcomes	Study limitations
			RaySearch on 170 cases from two centres in Norway. The final model was integrated into RayStation v9B.	thyroid gland were delineated by oncologists (delineation of the heart was based on a previous atlas). The remaining structures were delineated by radiation therapists.	satisfaction by RO's on a 4-point scale) Dosimetric analysis (changes in dose distributions on OARs) Geometric analysis (DICE, HD)	Uncertain generalisability as training from local cases were used. Time savings are discussed but not formally evaluated.

Abbreviations: CI = clinical investigator. CNS = central nervous system. CT = computerised tomography. D = dosimetrist. DICE = Dice similarity coefficient. H&N = head and neck; HD = Hausdorff distance. MRI = magnetic resonance imaging. H&N = head and neck. OAR = organs at risk; RO = radiation oncologist. ROI= regions of interest. RTT = radiation therapist. UK = United Kingdom. USA = United States of America.

8.1. Overview of methodologies of all included studies

All studies described in Table 4 had some methodological limitations or misalignment with the NICE decision problem for this appraisal.

8.2. Study design, intervention and comparator

Of the 15 prioritised papers, eight were prospective, five were retrospective (although one of these had a part prospective and part retrospective design) and two were conference abstracts (one of which was retrospective while the other appeared to be prospective).

Study methods were often poorly reported. The sequence of events, such as when and how patient selection occurred, and when manual and auto-contours were performed, was frequently unclear. It was often challenging to ascertain whether a study was prospective or retrospective in design. Similarly, it was regularly unclear what the intervention consisted of. Software or algorithm version numbers were typically unreported, and it was not always clear whether local training or other user-specific adjustments might have occurred, and if so, what difference this might have made to the “off the shelf” versions of the technology. This impacted on the generalisability of the evidence base. It was also often uncertain whether the intervention was the named technology, the algorithm that powers it, or the local adjustments or scripting algorithm that was performed in the tool.

The most common comparator was manual contouring (in 13 of the 15 prioritised studies)—these typically were used as a reference contour, against which auto-contours were judged. Sometimes manual contouring was performed by a single radiation oncologist, after which the contour may or may not have been reviewed, sometimes by a single of sometimes multiple radiation oncologists. Consequently, there was in some studies ambiguity about the reliability of the “ground truth” contour. SCM commentary also noted that there are tools in widespread use in treatment planning systems to (semi-) automate contouring for certain structures based on CT values, so even “manual contouring” isn’t always entirely manual. There was, for example, at least one instance in the prioritised studies where atlas contours were used as a basis for manual contouring (Van Dijk 2020[22])

The two studies that did not use manual contours as a reference instead used unedited auto-contours as the comparator to edited auto-contours. Four studies used atlas-contouring alongside manual contouring as a comparator; in one of these studies, atlas contouring was only used for lung segmentations.

> **Evidence gap:** There was a paucity of high quality, prospective studies. This may be understandable, as prospective studies are often more time-consuming and hence more expensive to conduct, but they offer greater control over potential confounding variables and are less susceptible to selection, allocation, and recall bias.

> **Generalisability gap:** The description of the intervention was often poor, meaning that it was unclear which version of a technology was being used. Alongside this, there is the question of whether local training sets had been used to train a technology, and if so, how this might affect the generalisability of the findings to other clinics or hospitals using ostensibly the same technology. However, the EAG notes that the question of local training does not apply to all software—some, such as AI-Rad, does not have the functionality to use local training sets.

8.3. Samples and structures

Sample sizes were often small—mostly under 100 cases. Larger sample sizes tended to come from retrospective studies, although in such studies it was often unclear how samples were selected. Some studies reported how patients were selected, for example, as consecutive patients from a certain date in the author's institution. However, it was often unclear how patients were selected and why.

The most commonly assessed structures were head and neck (11 studies) and the pelvis/prostate (10 studies). Other sites included thorax/lungs/breast (3 studies); rectum (1 study), abdomen (1 study), CNS (1 study). Only three of the prioritised studies were from the UK (one prioritised study for DLCExpert, one for OSAIRIS, and one for RayStation).

> **Evidence gap:** First, sample sizes were generally small, but more concerning is that details were often lacking in how the patients, cases or scans were selected for the study. This raises concerns of potential selection bias. Second, the

evidence base is focused on a) head and neck cancer and b) prostate cancer—other parts of the body are less well covered by the evidence.

Generalisability gap: The prioritised literature was global in scope, with only three of the prioritised studies coming from the UK. This raises generalisability concerns, though given the technical nature of most studies, the EAG expects that the evidence will be reasonably generalisable across borders. The rise in guidelines endorsed by regional and international bodies should further improve generalisability[84,85].

8.4. Outcomes reported

Outcome metrics were classified into four subtypes, based on Mackay et al[86]: time-saving, qualitative, dosimetric, and geometric. In line with the wider literature, most of our studies focused on geometric outcomes. Few studies were designed to capture real-world impact or patient outcomes: there was a lack of real-world, prospective trials.

Fourteen of the studies reported geometric analyses. These are the quantitative assessment of similarity between a contour produced by an included technology and a ground truth contour. Metrics include the Dice Similarity Coefficient (DICE), the Hausdorff distance (HD), Jaccard indices, and the mean distance to agreement. These were by far the most reported outcomes, even though research suggests that they may not be clinically meaningful—for example there is often only a weak correlation between geometric and dosimetric outcomes[86-89]. It has been suggested that a more practical assessment procedure should mimic clinical practice as much as possible[87]. As such, the categories of metrics described below (qualitative, dosimetric, and time-saving metrics) may be more meaningful.

Eight studies used qualitative assessments of auto-contours. This was most often via some form of scale scoring, where radiation oncologists or other experts would rank auto-contours on an ordinal scale according to whether they are, for example, “ready to use”, or would require “major” or “minor” adjustments before use, or are “completely unacceptable”. One study used a blinded Turing test, where assessors had to guess whether a contour had been created via auto- or manual contouring.

Seven studies measure time savings from the use of auto-contours. This metric should not just measure the amount of time taken for the creation of an auto- vs a manual contour, but also any time needed to check and edit auto-contours and prepare them for clinical use. Time-saving outcomes are reported both in the clinical section of the report and also in Table 6, on page 68, in the economic evidence section.

The least reported metrics, reported by just five studies, were dosimetric outcomes. These compare radiotherapy dosing plans generated for manual contours as against auto-contours, for example by comparing dose volume histograms and/or by assessing dose constraints.

> **Evidence gap:** Future studies should focus on outcomes beyond geometric analysis. Compared to geometric analyses, satisfaction scores offer a more practical real-world test of the usefulness of plans, while dosimetric analyses are more applicable to potential patient outcomes. Time-saving metrics are required to understand the cost-effectiveness of auto-contouring technology. However, they are often not reported, or only partially reported (e.g., the time required for editing or correcting auto-contours is not reported). As the technology potentially becomes more embedded in systems, then evidence of impact on patients should be prioritised, ideally via randomised, real-world, prospective trials (if ethically possible).

8.5. Results from the evidence base

Short narrative summaries of the evidence base for each technology are provided. For the prioritised extracted studies, see Table 4 for more details on study characteristics, and Table 11 (in the appendices) for a more detailed breakdown of the results. Following this results section is the EAG's interpretation of the clinical evidence (section 9), which summarises the results as follows:

1. That there is strong evidence for the potential usefulness of AI-based auto-contouring.
2. That while most auto-contours were clinically useful, there were always some that either needed major editing or were unusable. The same structures

repeatedly had lower DICE scores or were marked down in qualitative assessment.

3. That while timesaving has been reported by many studies, only reports that include the time taken for manual correction or editing of auto-contours are useful.
4. That the evidence base is focused predominantly on a) head and neck cancer and b) prostate cancer.
5. That it is not possible to say with any certainty that one software is more effective to another.
6. That those studies that included atlas contouring as a comparator were conclusive in showing that AI-based auto-contouring gives superior results to atlas contouring.

DICE is the most reported metric below. While there is no clear consensus on what an "acceptable" DICE score is, Strolin 2023[68] suggests that a score of ≥0.8 is considered agreement between two radiation oncologists. However, expert commentary for this EVA has suggested that this cut-off may not be meaningful as it varies across structures.

8.5.1. AI-Rad Companion Organs RT

The EAG identified three retrospective studies on AI-Rad. Ginn 2023[6] was prioritised for extraction due to its recency, its relatively large sample size, and because it had a partial prospective study design. The study involved 100 patients in its retrospective analysis, and 20 in its prospective component—both parts of the study included a mix of head and neck and pelvis scans. The study reviewed the time savings accrued, concluding that editing auto-contours was faster than manual contouring, with an average time saving of 43.4% or 11.8 minutes per patient. Over 95% of the auto-contours were clinically usable or only needed minor edits to make them so. Problematic structures for the algorithm included the prostate, the oesophagus, and the optic nerves. The authors concluded that the results were promising, but that human review and some editing remains required prior to use.

The other retrospective papers for AI-Rad were Hu 2023[7] and Marschner 2022[8]. Hu looked at a wider range of structures (head and neck, thorax, breast, and the male and female pelvis) and, similar to Ginn, found that auto-contours were clinically acceptable and only required minimal editing. Marschner 2022 focused on the algorithm used in AI-Rad rather than the technology itself and reported high alignment for both geometric and qualitative analysis. The few auto-contours that required manual corrections were mostly for the heart and rectum.

8.5.2. ART-Plan

The EAG identified six abstracts for ART-Plan (no full text studies were identified). Blanchard 2020[12] was prioritised because of its large sample size and its incorporation of a blinded analysis. The study included 100 head and neck cases and looked at versions of ART-Plan that had been trained on either 6000 cases per organ (v.1.0) or 21,000 cases per organ (v.2.0). Both geometric and qualitative analyses showed higher alignment to manual contours for v.2.0 over v.1.0. Auto-contours that were deemed clinically usable (i.e., either “acceptable” or “acceptable after minor corrections”) ranged from 100% for mandibles to 92% for the submandibular gland. The authors concluded that the auto-contours were very close to expert contouring and clinically usable in most cases.

Of the remaining abstracts identified for ART-Plan, Rivera 2020[16] focused on breast cancer, Nachbar 2021[13] on the pelvis, and Buatti 2022[14], Costea 2021[15], and Gregoire 2020[17] on head and neck. Gregoire reported on time savings, finding that two minutes on average were needed to correct the contours after auto-segmentation versus 30 minutes for manual delineation. Three of the abstracts, Costea 2022, Gregoire 2020 and Nachbar 2021, noted in their conclusions that dosimetric analysis is required to complement the geometric analyses presented.

8.5.3. AutoContour

The EAG identified three abstracts for AutoContour (no full text studies were retrieved). Leyva 2022[18] was prioritised for extraction because of its larger sample size. The study included 224 head and neck structures and focused on a geometric analysis of auto-contours vs manual contours. Good agreement was found between contours, with DICE scores of greater than 0.7 for 60% of the sample included. A larger variance in DICE scores was seen for small structures, such as pituitary, chiasm and cochlea. The authors concluded that the tool was efficient in removing inter-user segmentation variability that occurs with manual segmentation, but that while good quantitative agreement was found, further analysis was needed using a randomized qualitative scoring method and a larger segment sample.

One of the other abstracts, Bice 2022[19], looked at the head and neck, thorax, and pelvis regions, while Marasco 2022[20] focused on prostate cancer patients. Bice 2022, unlike Leyva, included a qualitative assessment, which incorporated an element of time-saving analysis. On a scale of 1 (“zero edits required”) to 5 (“entirely unusable output”), with 3 indicating “no time saved by using the software after editing”, the mean score for the auto-contours generated was 2.09. Marasco 2022, focusing on geometric analysis on the delineation of the rectum, reported an average DICE score of 0.8 between manual and auto-contours.

8.5.4. DLCExpert

The EAG identified three prospective studies, four retrospective studies, and 15 conference abstracts for DLCExpert. The three prospective studies were prioritised for extraction: Hague 2020[21] and Van Dijk 2020[22] both focused on the head and neck, while Vaassen 2021[23] selected non-small cell lung cancer patients. Hague 2020 collected qualitative and geometric outcomes that compared manual with auto-contouring using CT scans and MRI scans (MR offers improved soft tissue contrast and may therefore be superior to CT for auto-contouring). Qualitative assessment found that auto-contours were clinically acceptable for diagnostic and planning MRI scans, but not for Magnetic Resonance Linear Accelerator (MR-Linac) scans. Geometric analysis of auto-contours from the MR-Linac scans were correspondingly lower, with particularly low scores for the left and right submandibular glands (mean DICE of 0.1 and 0, respectively). The authors concluded that MR auto-contouring shows promise, with statistically improved performance vs a CT based model, although performance is affected by the method of MR acquisition and further work is needed.

Van Dijk 2020 reported greater time savings and improved qualitative scores for AI over atlas generated contours. Dosimetric analysis also showed a lower mean dose for organs at risk for AI over atlas contours. AI-contours further showed an improved alignment via geometric analysis with manual contouring compared to atlas contours for the thyroid gland, and all upper digestive tract and airway organs (except the oral cavity). The authors concluded that a suitably trained deep learning algorithm outperformed atlas contouring for the majority of head and neck organs at risk, and that deep learning contouring had the potential to replace atlas contouring currently used in their institution.

Vaassen 2021 similarly found that dose differences for the various AI generated treatment plans, when evaluated against the treatment plans generated via manual contouring, were small (on average below 1 Gy/1%) and that the majority of treatment plans fulfilled the planning objectives. The authors noted some instances where organs, contoured by AI, would have received doses above the clinical constraint: this happened on occasion for the heart, the lungs and the spinal cord. From the geometric analysis, the highest DICE scores were for lungs (1.0), while the lowest DICE score was for the oesophagus (0.46). The authors concluded with some specific thoughts about the checking required for individual structures. For example, the authors suggested that for the heart (Heart D_{mean}):

- *“Overlap between heart and PTV: the heart contour should be checked and adjusted.*
- *No overlap between heart and PTV: only a quick check is sufficient.”*

Similar heuristics were set out for the spinal cord, oesophagus and mediastinum.

Among the retrospective studies, Walker 2022 investigated head and neck and prostate structures, finding time savings compared to the existing clinical method (which may have differed slightly depending on the three centres) of 5.9 +/- 3.5 min for prostate contouring, and 16.2 +/- 8.6 min for head and neck structures. The results also showed acceptable geometric alignment.

Brouwer 2020[26] and Brunenberg 2020[27] both focused solely on the head and neck. Brouwer 2020 found that while most contours needed very little editing, some structures occasionally required large adjustments. The authors reported that auto-contouring tended to under-segment the desired contour (i.e., that enlarging of the contour was needed). Brunenberg 2020 similarly reported—using geometric and qualitative analysis—that auto-contours provided a reasonable starting point for delineation, noting that some organs at risk were better contoured than others.

Vaassen 2022[25] focused on the thorax region, comparing auto-contours to manual contours using qualitative real-world analysis. The authors reported that, similar to Brouwer 2020, while most contours needed very little editing, some structures occasionally required large adjustments, and that in such cases the auto-contours tended to need enlarging.

Finally, Lustberg 2018[28] looked at auto-contouring for lung cancer. A prototype of DLCExpert was trained with scans from 450 patients and evaluated on 20 patients (vs manual and atlas contouring). The median time for manual contouring was 20 min. Atlas contouring saved a median time of 7.8 min, while deep learning contouring saved a median time of 10 min. Deep learning also scored better (scale scoring) than atlas contours for most, but not all, structures. The authors concluded that deep learning contouring showed promising results compared to existing solutions.

Because of the large number of full text studies, the EAG offers no commentary on the conference abstracts for DLCExpert.

8.5.5. INTContour

The EAG identified two retrospective studies. Duan 2022[42] was prioritised because of its recency. The study focused on 23 people with prostate cancer. Geometric analysis showed good alignment between auto- and manual contours, with slightly lower scores for the seminal vesicles and penile bulb (DICE scores of 0.72 and 0.53 respectively). A qualitative analysis showed that 95.7% of the auto-contours were either “perfect” or “acceptable”. And dosimetric analysis similarly showed good alignment with the manual approach, with no statistically significant differences between the two for organs at risk, except for the bladder (where the auto-contour had generated a lower dose). The authors concluded that, using the investigated model, the implementation of an automated prostate treatment planning process was clinically feasible.

The second retrospective study, Chen 2020[43], focused on head and neck images, comparing deep learning contours with atlas and manual contouring. Geometric and dosimetric analyses both found that deep learning contours were more closely matched with that of manual contours when compared with atlas contours.

8.5.6. Limbus Contour

The EAG identified three prospective studies, three retrospective studies, and eight conference abstracts. The three prospective studies were prioritised for extraction: Radici 2022[44], Wong 2021[45] and Wong 2020[46]. Radici 2022 investigated 12 cases: three head and neck, three prostate, three rectum and three breast. Time savings were identified for head and neck contours (-65%), breast (-46%), prostate (-18%), and rectum contours (-17%). Good geometric alignment was found between auto- and manual contours, although the penile bulb scored slightly lower (DICE score of 0.39). Dose distributions were also similar, except for the bowel, the reason for which was discussed. The authors concluded that auto-contouring was able to save time, simplify the workflow, and reduce interobserver variability, and that its implementation improved the radiation therapy workflow in their department.

Wong 2021 reported looking at 601 plans, from a subset of which assessment surveys were reported for the central nervous system (n = 27), head and neck (n = 54), and prostate (n = 93). They reported that satisfaction, as measured on a five-point scale (from 1 [“poor”] to 5 [“high”]), was scored at an average of 4.1 for the CNS, 4.4 for head and neck, and 4.6 for the prostate. The optic chiasm had the poorest geometric scores and needed the most editing. The authors concluded that the high user satisfaction suggested that the auto-contours served as appropriate starting points for patient specific edits.

Wong 2020 included 60 cases, twenty for each of central nervous system, head and neck, and prostate. Similar to Radici 2022, Wong also reported time savings: the mean auto and manual contouring times were, respectively: 0.4 vs 7.7 min for CNS; 0.6 vs 26.6 min for head and neck; 0.4 vs 21.3 min for prostate. Geometric analyses showed that deep learning contours approximated the expert Inter-observer-variability seen for organs at risk, although deep learning contours for clinical target volumes were less accurate. The authors concluded that auto-contours would likely serve as a usable starting template for patient specific adjustments.

Among the retrospective studies, Wong 2021[47] focused on the training and validation of lung auto-contours, reporting good geometric alignment to clinical contours for most organs, but slightly poorer alignment for the brachial plexus (DICE score 0.52). D’Aviero 2022[48] looked at the head and neck. Geometric analysis showed that alignment with manual contours was good, particularly for the brain (DICE score 1), left and right eye globes and the mandible (DICE score 0.98). The structures that required greater editing were the optic chiasm, optic nerves, and cochleae. Finally, Zabel 2021[90] focused on the bladder and rectum, and found that more editing of atlas contours was needed than deep learning contours. Time savings were also reported. Mean durations for initial contour generation were 10.9 min, 1.4 min, and 1.2 min for manual, deep learning, and atlas contours, respectively. However, because initial deep learning contours were more similar to manual contours, the mean durations of the editing steps for manual, deep learning, and atlas contours were 4.1 min, 4.7 min, and 10.2 min, respectively, leading to overall time savings for deep learning auto-contouring.

Because of the large number of full text studies, the EAG offers no commentary on the conference abstracts for Limbus Contour.

8.5.7. MIM Contour ProtégéAI

The EAG identified one prospective study and one retrospective study for MIM Contour. The prospective study was prioritised for extraction. Urago looked at 51 cases, 30 from the head and neck and 21 from the prostate, and compared AI-contours with manual and atlas contours. The researchers reported that AI-based delineations were more consistent with the manual ones than the atlas contours were. There were no significant differences between manual and AI-contours except for some small delineations such as the optic chiasm and optic nerve. For prostate patients the processing time to create delineations was approximately 3 min per case for atlas and approximately 5 min per case for AI, while for patients with head and neck cancer the processing times were both approximately 6 min. The authors concluded that the effectiveness of the AI-based model can be expected to improve efficiency and to significantly shorten delineation time.

The retrospective study by Lastrucci 2022[59] focused on prostate cancer and compared AI against both atlas and manual contours. The geometric analysis reported that AI performed consistently better than atlas contouring: for example, mean DICE scores for the prostate were AI 0.78 vs atlas 0.64, and for the rectum AI 0.86 vs atlas 0.58.

Because of the full text studies, and large number of abstracts, the EAG offers no commentary on the conference abstracts for MIM Contour ProtégéAI.

8.5.8. MRCAT Prostate plus Auto-contouring

The EAG identified one prospective study and one conference abstract. The prospective study by Kuisma 2020[66] looked at 65 prostate cases and compared them via geometric analysis to manual contouring. The researchers reported that DICE scores showed high alignment for delineating prostate (0.84), bladder (0.92), and rectum (0.86), although scores were lower for seminal vesicles (0.56) and penile bulb (0.69). The authors concluded that the auto-contours showed good agreement and repeatability compared with manual contours, although manual review and adjustment of some structures in individual cases remained important.

The abstract by Maspero 2018[67] investigated whether MRCAT, designed for patients with prostate cancer, might also be suitable for patients with rectal cancer. The geometric results showed good alignment and the dosimetric results suggested that dose distributions were accurate—and therefore that MRCAT appears feasible for use in a clinical radiotherapy workflow for patients with rectal cancer.

8.5.9. MVision Segmentation Service

The EAG identified two retrospective studies and two abstracts. Of the two retrospective studies, Strolin 2023[68], who looked at 111 cases of head and neck, breast, abdomen, thorax, male pelvis, and female pelvis cases, was prioritised because of its recency and large sample size. The authors reported that median DICE scores, when comparing manually adjusted auto-contours vs unedited auto-contours, were higher than 0.8 for all the organs except for the oesophagus and glottis. Qualitative analysis showed that radiation oncologists scored 44% of unedited auto-contours as 4 (“well done”) and 43% as 5 (“very well done”). The median time for manual delineation, deep learning-based segmentation, and subsequent manual corrections were 25, 2.3 and 10 minutes, respectively. The authors concluded that the tool offered a high level of user satisfaction, saved time, and improved consistency among radiation oncologists.

The second retrospective study, by Kiljunen 2020[69], focused on 30 prostate cancer patients across six clinics. The study reported that mean time saved by using auto-segmentation as against manual contouring was 12 minutes for the whole data set (-46%). In terms of geometric outcomes, mean DICE scores, when comparing manual with auto-contours (across all six clinics), were 0.82 for prostate, 0.72 for seminal vesicles, 0.93 for bladder, 0.84 for rectum, 0.69 for femoral heads and 0.51 for penile bulb. The authors concluded that using auto-contouring saves time and improves consistency.

The two abstracts looked at prostate cancer (Suresh 2021[70]) and breast cancer (Heikkila 2020[71]). Suresh reported that geometric and dosimetric outcomes showed good alignment between manual and auto-contours. Heikkila investigated AI contours against both manual and four atlas-based segmentation software. They reported that the AI method resulted in equal or better contours as compared to atlas-based methods.

8.5.10. OSAIRIS

The EAG identified one study (Oktay 2020[72]), a retrospective analysis of 132 pelvic male and 46 head and neck cases that compared auto-contours to manual contouring. Geometric analysis revealed that auto-contouring achieved levels of clinical accuracy within the bounds of expert interobserver variability for 13 of 15 structures (the left and right submandibular glands were the only two structures outside the bounds). Auto-contouring also saved time. Manual segmentation of nine organs at risk took 86.75 min/scan for an expert reader and 73.25 min/scan for a radiation oncologist. Whereas the correction time of auto-contours was 4.98 min/scan for head and neck scans and 3.40 min/scan for prostate scans. The authors concluded that, with the availability of open-source libraries and reliable performance, the tool creates significant opportunities for the transformation of radiation treatment planning.

8.5.11. RayStation

The EAG identified two retrospective studies and two abstracts. Of the two retrospective studies, Almberg 2022[73] was prioritised because of its currency. The study focused on breast cancer, and the algorithm was trained on 170 patients and evaluated on a further 30. The authors reported that "no" or only "minor corrections" were required for 14% and 71% of the clinical target volumes and 72% and 26% of the organs at risk, respectively; the most frequent corrections were made for the cranial and caudal parts of the structures. Geometric analysis revealed that auto-contour variation was generally less than observed manual inter-observer variation. Dosimetric analysis also found consistency between auto- and manual contours—while some statistically significant differences were found, no differences in organs at risk dosage were considered clinically relevant. Finally, although not formally measured, the authors estimated that the use of AI reduced total delineation time from roughly 1 hour to 15 minutes per patient.

The second retrospective study, by Rigaud 2021[74], focussed on cervical cancer, with the algorithm in RayStation 9B trained on 255 scans and tested on 61 validation, 62 internal test (at a centre in the US) and 30 external test scans (at a centre in France). The authors found similar performance between the two institutional data sets and reasonable dosimetric accuracies. Auto-contouring challenges were found when there was an absence of clear contrast between organs (e.g. between the cervix and bladder) or other difficulties with the scan, though the authors noted that in these cases segmentation failures could be easily identified visually and quickly corrected manually.

Both of the abstracts focused on the male pelvis. Liu 2022[75] reported that geometric analysis showed that the agreement between auto-segmented structures and manually segmented structures was similar to previously reported values of interobserver variability. DICE scores between auto- and manual contours were 0.95, 0.85, 0.82, 0.92 and 0.91 for the bladder, prostate, rectum, left femur, and right femur, respectively. The second abstract, Sidorski 2021[76], reported an average DICE score of 0.85 for organs at risk (bladder, rectum and femoral heads) and 0.7 for prostate.

8.5.12. Multiple technology comparison

The EAG retrieved six abstracts that compared two or more of the included technologies: Borkvel 2022[77], Rong 2022[78], Liao 2022[79], Yuan 2022[80], Gorgisyan 2022[81], and Doolan 2021[82]. Because they are abstracts, reported details are sparse.

Borkvel 2022[77] compared MVision, ART-Plan, DLCExpert, RayStation and AutoContour, for prostate and head and neck cancer. For prostate, they reported that, on average, manual contours took 43.1 min and that the range of time reduction provided by AI tools was between 1.1 and 67.4%, with a median value of 50.7%.

(The EAG noted the abstract did not report the mean reduction, which is the statistic required for decision making.) Larger time savings were seen for tools with higher average DICE scores and for a tool that included a local training data set (from the North Estonia Medical Centre). For head and neck, manual contouring took an average 2.81 hours. AI tools were found to reduce this time by nearly 50% (though the EAG noted that the authors do not report whether this is a median or mean figure). The authors also noted in their conclusion the importance of training the algorithm, and that including local data in the algorithm training set can improve outcomes.

Rong 2022[78] compared MIM, Limbus Contour, DLCExpert, INTContour and AutoContour, for head and neck cancer. They reported that lower DICE scores (compared to manual contours) were found for the optical chiasm, oral cavity, optical nerves, and the cochlea. The authors also reported that some AI platforms showed better consistency compared to the corresponding manual contours, especially for those soft tissue organs that are difficult to identify on CT images—the authors do not report which platforms.

Liao 2022[79] compared MIM, Limbus, DLCExpert, INTContour and AutoContour, also for head and neck cancer. The authors report that AI generated contours were accurate for high contrast and relatively large organs, such as: mandible, brain, parotids, eyes, and sub mandibular glands, with DICE scores of around 0.8-0.9 from all modules. However, for low contrast regions, more complex and/or smaller structures, such as the brachial plexus and chiasm, DICE scores decrease to <0.5. The score also varied among modules for long structures like spinal cord and oesophagus. The study further incorporated a qualitative assessment of the AI contours by two dosimetrists and two physicists using a 4-point scoring scale, in which the authors report that Limbus was the top performer.

Yuan 2022[80] looked at MIM, Limbus Contour, DLCExpert, INTContour and AutoContour for three anatomical sites: thorax, abdomen, and pelvis. Among the 25 organs investigated, 10 had DICE scores of >0.9, including lung, liver, kidney, femoral head, bladder, and heart. Eight averaged DICE scores of 0.7 to 0.89, including spinal cord, rectum, and stomach. The remaining organs, including gallbladder, bronchus, duodenum, seminal vesicle, penile bulb, and brachial plexus, reported average DICE scores of <0.7. The authors concluded that AI contouring results in large variations in accuracy among organs of interest, indicating that quality assurance of these tools is necessary before clinical implementation.

Gorgisyan 2022[81] investigated MVision and RayStation for the male pelvis. The study found that both AI models demonstrated good performance for the bladder and rectum, but that clinic specific training data might be necessary to achieve segmentation results in accordance with the clinical specific standard for some anatomical structures, such as the femoral heads in the case of the authors' institution. The authors also noted that manual delineation took on average 13 minutes compared to 0.5 minutes (RaySearch) and 1.4 minutes (MVision), although this did not include manual correction.

Doolan 2021[82] compared MVision and DLCExpert. They focused on breast, head and neck, lung, and prostate. The authors reported that both commercial AI contouring solutions generated contours with consistently high DICE and low HD scores, offering good quality structures across all anatomical sites. Also, it was noted that the time to correct the AI contours was less than the time required to contour the structures manually, for both technologies.

9. INTERPRETATION OF THE CLINICAL EVIDENCE

The rapid review of the evidence for eleven AI-based auto-contouring technologies has identified some key themes in terms of the clinical evidence.

First, there is strong evidence for the potential usefulness of AI-based auto-contouring in healthcare systems. All the studies reported either geometric, dosimetric or satisfaction scores which showed that AI-based auto-contouring creates contours, segmentations or plans similar to those created by manual contouring for most organs at risk and clinical target volumes. The majority of auto-contours were either ready to use or usable with only minor edits.

Second, while most auto-contours were clinically useful, there were always some that either needed major editing or were unusable. The same structures repeatedly had lower DICE scores or were marked down in qualitative assessment. Mostly, they were smaller structures, including the pituitary gland, the optic chiasm and optic nerves, the cochlea, the submandibular glands, the oesophagus, the seminal vesicles, and the penile bulb (SCMs note that these difficulties with small volumes may arise because of CT slice thickness, or could simply be a function of how DICE is calculated*). This suggests that auto-contours should be used—at least at the present time—as starting points for clinical contouring; that all contours need to be evaluated and edited as necessary before clinical use. While there is anticipation that auto-contouring accuracy will continue to improve, the EAG are interpreting the evidence as it currently stands.

This leads onto the third point, which is that while timesaving has been reported by many studies, only reports that include the time taken for manual correction or editing of auto-contours are useful. An SCM noted that, “unless perfect, which is unlikely, there will always be a requirement for an expert to review and modify”. Therefore, some of the larger claims of time-savings need to be considered carefully, particularly if they are a crude comparison of time to perform a manual contour vs time for AI-

* See, for example, the AAPM (American Association of Physicists in Medicine) report on image assessment in radiotherapy at: https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.12256

based auto-contouring, with no assessment of the entire workflow, including quality assurance.

Fourth, the evidence base is focused predominantly on a) head and neck cancer and b) prostate cancer. While the included studies have not exclusively investigated these cancers and the relevant structures (some studies evaluated auto-contouring in, for example, the CNS and thorax), other parts of the body are less covered by the evidence-base.

Fifth, because of the multiple metrics reported, the lack of clinical relevance of many of the metrics used, the limited generalisability due to local software training, and the risk of bias in the trials, the EAG cannot say with any certainty that one software is more effective to another. The EAG notes that three technologies have a larger evidence-base than others (measured simply in terms of numbers of studies identified): these are DLCExpert and Limbus Contour, followed by MIM Contour ProtégéAI.

Sixth, those studies that included atlas contouring as a comparator were conclusive in showing that AI-based auto-contouring gives superior results to atlas contouring. This was seen in satisfaction scores, geometric and dosimetric outcomes. And because AI-based contours were often initially closer to the (reference standard) manual contours, they took less editing time too.

10. ADVERSE EVENTS AND TECHNOLOGY CONSIDERATIONS

None of the prioritised studies reported adverse effects. This is most likely because all the AI-based auto-contours required approval by a qualified clinician before use; clinicians were on hand to manually delineate structures in cases where automated tools failed to produce a suitable contour.

Nevertheless, adverse events in the field of auto-contouring are possible. AI-contouring may occasionally misidentify organs, leading to potential serious errors; for example, by mistaking the bladder for one of the kidneys (company submission: INTContour). There is also the risk that as AI contouring tools become more successful, there becomes an increased reliance on these tools, resulting in decreased human review of AI-generated contours. Even with quality assurance processes, there remains a risk of poor-quality contouring if AI segmentation is inaccurate or not based on guidelines, which could misguide staff members during segmentation. It will be necessary to prevent the deskilling of the workforce and ensure that clinicians maintain their skills to mitigate potential safety issues.

MRCAT relies on images generated by Magnetic Resonance Imaging (MRI) (company submission; MRCAT). MRI is often considered superior to Computed Tomography (CT), in terms of soft tissues detailing, and therefore using MRI offers the potential for improved contouring—at least for some structures[91]. However, it is important to note that MRI is contraindicated in patients with certain implants due to the high magnetic field involved[92].

There are also software considerations. The EAG searched MAUDE (U.S. Food and Drug Administration) and MHRA (UK Government alerts), for the last five years, and identified two safety alerts. The first, from MAUDE in July 2021[93], described a software implementation error concerning dose tracking in RayStation; the second, from MHRA in February 2023[94], also concerned dose tracking in RayStation. In both cases the event happened in the wider treatment planning system—rather than in the auto-contouring algorithm itself. The company stated that this will be corrected in a new release of the technology in June 2023.

Finally, processing data locally or sending it to the cloud of a host company after pseudonymization, as done with software like MVision (company submission; MVision) or building an NHS cloud as OSAIRIS (company submission; OSAIRIS), entails data protection risks that should be considered[95]. The EAG notes that companies who acknowledge this risk emphasise that patient-identifiable data would not be compromised.

> **Evidence gap**: There is currently very little reporting of adverse effects because the evidence base relates primarily to technological over patient outcomes. However, if auto-contouring becomes embedded into systems then, as noted above, real-world, prospective trials with patient outcomes should be performed, at which time evidence of harms should be collected. The EAG noted that there are numerous reports in the evidence base of individual auto-contours being clinically unusable or would have led to a dangerously high dose to an organ at risk. There is, therefore, real potential for patient harm if quality assurance procedures are not followed.

11. ECONOMIC EVIDENCE

11.1. Published economic evidence

Based on the NICE scope, the EAG did not identify any published cost effectiveness evidence within the literature i.e., cost utility analyses, cost effectiveness analyses, cost minimisation analyses or cost consequences analyses comparing the listed AI-auto contouring technologies to either manual or atlas-based contouring were not available.

> **Evidence gap**: There is currently a lack of published cost effectiveness evidence comparing AI auto-contouring interventions (as outlined in the NICE scope) to manual or atlas-based auto contouring in people having radiotherapy treatment planning for external beam radiotherapy. There is a need for robust evidence generation regarding AI intervention costs in clinical practice and how these technologies impact on healthcare resource use and/or patient outcomes.

11.2. Economic evaluation

Conventional health economic evaluation techniques with patient-focused outcomes were not considered appropriate for analysis. A cost utility analysis was not feasible due to the lack of quality of life/patient reported outcome data within the published literature for the interventions of interest. Similarly, a cost effectiveness analysis was not considered appropriate due to the lack of quantifiable health outcomes reported such as life years gained. The EAG hypothesised that patient outcomes were unlikely to be affected by use of AI auto-contouring, the benefit being primarily a reduction in time taken in treatment planning. A case could be made for improved patient outcomes if AI auto-contouring were shown to be superior to manual, for example through better targeted radiation leading to a reduction in dose and less damage to surrounding healthy tissues, but this is purely conjectural. A cost minimisation analysis was considered, but due to the heterogeneity in AI intervention cost data provided to the EAG, it was not feasible to identify the lowest cost AI intervention (see Section 6211.2.1.1).

A simple cost consequences analysis was selected as the most appropriate evaluation method given the extent and quality of data available. A cost consequences analysis is a form of economic evaluation that compares the costs and consequences of different healthcare interventions/programmes. It provides a comprehensive assessment of the costs incurred and the outcomes achieved, without aggregating them into a single metric such as cost effectiveness or cost utility ratios.

Based on the heterogeneity/variability of the evidence available with respect to published clinical outcomes (consequences) and technology costs (see section 11.2.1 for more detail), a robust cost consequences analysis reporting incremental results (AI interventions vs manual contouring and atlas-based auto contouring), was not possible. The EAG has therefore opted to take a summative approach which lists the costs associated with each AI technology, and where available, reports information on a key consequence identified in the literature. In this instance resource use or time associated with the use of AI auto-contouring technology (vs manual or atlas) was considered as the primary consequence, as it was the health economic outcome most consistently reported in the literature.

11.2.1. Costs

11.2.1.1. Intervention costs

The costs associated with each AI technology were provided by the manufacturers (See Table 5: AI intervention costs). Intervention costs consisted of software costs (including license and subscription costs), hardware costs (one-off installation costs), data storage costs and maintenance costs.

The EAG noted considerable heterogeneity and uncertainty surrounding the reporting of intervention costs, including the following.

- Costs for OSAIRIS, DLCExpert and ART-Plan were only reported costs on a per patient basis i.e., there was no information on the cost of the software, hardware of maintenance costs for these interventions.

- Per patient costs ranged from £4 to £50. However, the cost per patient reported by manufacturers was dependent on hospital size/patient throughput and number of scans. Note that cost per patient was not reported for all technologies.

- [redacted]

- Data storage costs were dependent on whether cloud storage would be used or if data were to be stored on a hardware device. The applicability of these costs therefore depend on each individual hospital's data storage set up.

- Some intervention costs were noted to be exclusive of VAT, whilst the VAT status of others was not clear.

Ultimately the EAG considered there to be significant variability in the pricing and reimbursement strategies for each technology. The lack of complete costing information from all manufacturers and variability in annual treated patient numbers/CT scans introduced further uncertainty. It is challenging to therefore identify the true cost for each intervention.

> **Evidence gap**: The cost per technology appears to be dependent on multiple factors including utilisation rates in clinical practice. For a meaningful assessment of technology costs, it would be helpful to have hospital case study evidence for all interventions.

Table 5: AI intervention costs

Technology	Software cost	Hardware cost	Maintenance cost	Cost per patient
INTContour	[redacted]	[redacted]	[redacted]	[redacted]
AI-Rad	[redacted]	[redacted]	[redacted]	[redacted]
ART-Plan	[redacted]	[redacted]	[redacted]	[redacted]
AutoContour	[redacted]	[redacted]	[redacted]	[redacted]
MRCAT	[redacted] The software can be added to existing MR for [redacted]	The MR scanners range from £750K to £950K depending on configuration requirement Note: The full RT imaging capability may be more than this if LAP laser	Customer support agreement would be [redacted] per annum for this equipment. Costs (ex VAT)	No limit on number of patients

		bridges are required in addition		
MVision	[redacted]	[redacted]		[redacted]
OSAIRIS	NR	NR	NR	£4/ patient
Raystation	[redacted]	[redacted]	[redacted]	[redacted]
DLCExpert	[redacted]	[redacted]	[redacted]	[redacted]
Limbus AI	[redacted]	[redacted]	[redacted]	[redacted]
MIM	[redacted]	[redacted]	[redacted]	[redacted]

Abbreviations: AI artificial intelligence, GPU graphic processing unit, LAP laser ablation of the prostate, MR magnetic resonance, NHS National Health Service, NR not reported, RT radiotherapy, USD United States dollars, VAT value added tax,

11.2.2. Consequences

Time taken to delineate organs at risk and review and edit images associated with AI auto-contouring (vs manual or atlas) was considered an outcome of economic interest on the basis that this could be quantified and therefore potentially be used to inform an economic analysis. As previously mentioned in Section 8, seven prioritised studies were identified that measured time from the use of AI auto-contours, these included Ginn et al. (2023)[6], Van Dijk et al. (2020)[22], Radici et al. (2022)[44], Wong et al. (2020)[46], Strolin et al. (2023)[68], Oktay et al. (2020)[72], and Urago et al. (2021)[58]. Results are reported in Table 6. The EAG noted the following concerns surrounding the reporting time as an outcome measure/consequence within the published literature.

- There was considerable heterogeneity between studies in terms of study design, number of patients included and site of tumour/AI-auto-contouring use, thus making it challenging to compare interventions and meaningfully interpret reported times.
- There was variability between studies in relation the time outcome itself i.e., some studies only reported time taken to delineate organs at risk whilst others reported delineation time and the editing time required by a clinician.
- Time estimates were not available for all AI auto-contouring technologies i.e., ART-Plan, AutoContour, INTContour and MRCAT.

On balance, the majority of studies which assessed time as an outcome measure, reported time savings associated with the use of AI auto-contouring compared to manual contouring. However, the EAG noted one study by Urago et al. (2021)[58] reported that an atlas-based model resulted in less processing time to create delineations compared to the AI-based model for the patients with prostate cancer (see Table 6).

Due to the variability of reported data, SCMs were asked to provide additional commentary that could help fill data gaps and potentially be used to inform an economic evaluation. The full list of EAG 'cost effectiveness' questions and SCM responses are provided in Table 12, Appendix E. The majority of SCMs confirmed that AI auto-contouring is likely to result in a reduction in clinician time compared to standard approaches used (manual and atlas-based contouring), though one clinician noted this may not be the case when AI auto-contouring is used in patients who have unusual anatomy/post-surgical changes in anatomy. When asked how much time clinicians spent editing AI auto-contours, responses ranged from 10 minutes to 30 minutes. Several clinicians noted that editing time is highly variable and depends on the different structures and tumour sites which are contoured in clinical practice.

Table 6: Studies from the prioritised literature which contained a 'time saved' component

Study	Name of Tech/Manufacturer	Comparator	Results
Ginn et al. (2023)[6]	AI Rad-Companion organs RT (Siemens Healthineers)	Manual contouring	Edited AI saved 11.8 mins per patient
Van Dijk et al. (2020)[22]	DLCExpert (Mirada Medical)	Manual contouring/Atlas	**Atlas contours:** Average expert delineation time: 36 ± 7 Average beginner delineation time: 59 ± 14 minutes **Deep learning contours:** Average expert delineation time: 34 ± 6 Average beginner delineation time: 54 ± 8 minutes
Radici et al (2022)[44]	Limbus Contour (Limbus AI, AMG Medtech)	Manual contouring	**Time savings absolute and relative for:** Head and neck contours (80 min, -65%) Breast contours (7 min, -46%) Prostate contours (4 min, -18%) Rectum contours (3 min, -17%).
Wong et al. (2020)[46]	Limbus Contour (Limbus AI, AMG Medtech)	Manual contouring	**Mean auto and manual contouring times:** 0.4 vs 7.7 min for CNS 0.6 vs 26.6 min for head and neck 0.4 vs 21.3 min for prostate.
Strolin et al. (2023)[68]	MVision Segmentation Service (MVision AI, oy Xiel)	Manual contouring	**The median (range) time (mins) for:** Manual delineation 25.0 (8.0-115.0) Deep learning-based segmentation, and subsequent manual corrections were, 2.3 (1.2-8) and 10.0 minutes (0.3-46.3), respectively

Oktay et al. (2020)[72]	OSAIRIS	Manual contouring	Manual segmentation of nine organs at risk took 86.75 min/scan for expert reader and 73.25 min/scan for radiation oncologist. With AI to assist them in reviewing and editing it took: 4.98 (95% CI, 4.44-5.52) min/scan for head and neck 3.40 (95% CI, 1.60-5.20) min/scan for prostate The autogenerated contours represented a 93% reduction in time
Urago et al. (2021)[58]	MIM	Atlas contouring	**Delineation time per case:** Prostate cancer atlas: approximately 3 min Prostate cancer AI-based approximately 5 min Head and neck cancer using both atlas and AI-based: approximately 6 min (range, 3–8 min).

Abbreviations: AI artificial intelligence, CI confidence interval, CNS central nervous system, SIB simultaneous integrated boost, VMAT volumetric arc therapy

12. INTERPRETATION OF THE ECONOMIC EVIDENCE

No cost effectiveness evidence was found for any of the 11 AI auto-contouring interventions listed in the NICE scope. There remains uncertainty surrounding the cost effectiveness of these AI auto-contouring interventions compared to manual contouring, atlas-based auto contouring and model-based segmentation in people having radiotherapy treatment planning for external beam radiotherapy. The EAG initially aimed to utilise manufacturer cost data and published outcome data (on time taken to delineate organs at risk and review and edit images associated with AI auto-contouring vs manual or atlas) in order to create a bespoke cost consequences analysis that could help to inform healthcare decision making. However, due to the variability in the pricing and reimbursement strategies for each technology and limitations surrounding the economic outcome of interest within the published literature, a robust analysis was not considered feasible. A summative approach, collating manufacturer AI technology costs, SCM opinion and published data on time to delineate organs at risk and review and edit images was therefore adopted to provide a comprehensive overview of the information available to the EAG.

Overall, the key costs and outcomes that are of relevance to the decision problem are intervention costs (licence fees and upgrade costs, service contracts, data storage and capital equipment, costs involved in commissioning—i.e. the testing and evaluation of systems prior to clinical use—and costs involved in preparing data for training models should this be required), differences in time taken to contour and plan a course of treatment, and user (i.e. clinician) satisfaction. It may be possible to hypothesise improved patient outcomes from more carefully directed radiation (via increased dose delivered to the cancer and reduced damage to surrounding tissue), however the evidence base to support or refute this is still emerging. The evidence base focuses on head and neck and prostate cancers and overall intervention costs ranged from £4 to £50 per planning session. Evidence on reduced time for treatment planning was heterogeneous, including by site of contouring, and thus unclear in the magnitude of effect: a critical element is the time required for any manual editing, which was not reported consistently across studies. The majority opinion of SCMs was that AI-based auto-contouring was likely to reduce treatment planning time, but manual editing may negate some (or all) of this. Estimates of time saved from the literature ranged widely from approximately 3 to 80 minutes, depending largely on tumour site.

The EAG have produced a simple, interactive, cost offset calculator to illustrate the relationship between time savings and staffing costs (see Appendix F for more details)

13. EVIDENCE GAP ANALYSIS AND FUTURE RESEARCH

13.1. Evidence gap analysis

A summary of evidence gaps, pertaining to outcomes, study design and structures covered, is summarised in Table 7. The table was populated based on full text evidence. Therefore, ART-Plan and AutoContour have been marked up as 'N/A', as the EAG only identified conference abstracts for these technologies. A narrative assessment of evidence gaps in other methodological areas besides outcomes is presented within the clinical section of the report. In terms of gaps in the cost effectiveness evidence the EAG noted that there is a need for robust evidence generation regarding AI intervention costs in clinical practice and how these technologies impact on healthcare resource use and/or patient outcomes.

Table 7: Evidence Gap Analysis (based on full-text evidence only)

	AI-Rad	**ART-Plan**	**Auto contour**	**DLC Expert**	**INT Contour**	**Limbus Contour**	**MIM Contour**	**MRCAT Prostate**	**MVision**	**OSAIRIS**	**Ray Station**
Outcomes											
Time-saving metrics	✓	N/A	N/A	✓		✓	✓		✓	✓	
Qualitative assessment	✓	N/A	N/A	✓	✓	✓	✓		✓		✓
Dosimetric analyses		N/A	N/A	✓	✓	✓				✓	✓
Geometric analyses	✓	N/A	N/A	✓	✓	✓	✓	✓	✓		✓
Adverse effects		N/A	N/A								
Study design											
Prospective (full text)	**RED**[*]	**RED**	**RED**	**GREEN** 3 studies	**RED**	**GREEN** 3 studies	**AMBER** 1 study	**AMBER** 1 study	**RED**	**RED**	**RED**
Retrospective (full text)	**GREEN** 3 studies	**RED**	**RED**	**GREEN** 4 studies	**GREEN** 2 studies	**GREEN** 3 studies	**AMBER** 1 study	**RED**	**GREEN** 2 studies	**AMBER** 1 study	**GREEN** 2 studies
Structures covered											
Central nervous system		N/A	N/A			✓					
Head and neck	✓	N/A	N/A	✓	✓	✓	✓		✓	✓	
Pelvis	✓	N/A	N/A	✓	✓	✓	✓	✓	✓	✓	✓
Thorax/abdomen	✓	N/A	N/A	✓		✓			✓		✓

Colour coding: Green = ≥2 studies; Amber = 1 study; Red = no studies

[*] Note that Ginn 2023, as previously described, has both prospective and retrospective study components.

There are a number of evidence gaps in respect of the clinical evidence base as it pertains to the decision problem. Key gaps included:

Population gaps

- The prioritised literature was global in scope, with only three of the prioritised studies coming from the UK. This raises generalisability concerns, though given the technical nature of the decision question, the EAG expects that the evidence will be reasonably generalisable across borders.
- Sample sizes were mostly small, but more concerningly is that there were often no details provided on how the patients, cases or scans were selected for the study. This raises concerns of potential selection bias.
- The evidence base is focused on a) head and neck cancer and b) prostate cancer—other structures of the body are less well covered by the evidence.
- There is uncertainty about the availability of training sets for certain demographics, such as children or people with disabilities. This should be addressed as a potential equity issue.

Intervention gaps

- The description of the intervention was often poor—it was often unclear which version of a technology was being used.
- There is the question of whether local training sets had been used to train a technology, and if so, how this might affect the generalisability of the findings to other clinics using the same technology.
- There is no full-text evidence available for ART-Plan and AutoContour.

Comparator gaps

- Manual contouring was the most common comparator. There is relatively less evidence of AI-based auto-contouring vs atlas contouring.
- There is no full text evidence comparing any of the included technologies against each other.

Outcome gaps

- The most common outcomes across the included studies were one or geometric metrics (such as DICE or HD). However, there is a lack of an agreed, high-quality metric which can be universally applied, which makes comparison of geometric performance across studies difficult. Furthermore, satisfaction scores offer a more useful real-world test of the useful of plans, while dosimetric analyses are more applicable to potential patient outcomes. (Although Borkvel 2022[77] did find that larger time savings were seen for tools with higher average DICE scores.)
- Time-saving metrics were often not reported, or only partially reported. Without reporting of the time spent to quality assure and edit auto-contours, timings are of limited use.
- As the technology potentially becomes more embedded in the NHS, evidence of impact on patients will be required, ideally via randomised, real-world, prospective trials. While process efficiency is important, there is nevertheless potential to improve patient outcomes. For example, accurate organ at risk delineation should lead to a reduction in treatment toxicity. Also, cluster trials may be able to identify improvements in throughput, leading to reduced treatment waiting times, and therefore improved outcomes for patients treated more swiftly.
- Outcomes differ sharply between different structures. The evidence-base suggests that larger structures are generally well delineated by AI-based auto-contour software, but that smaller and/or elongated structures are more difficult to delineate. More evidence is needed on which structures can best be handled by auto-contouring, and where extra caution may be required.

Other considerations

- There was a paucity of high quality, prospective studies. This may be understandable, as prospective studies are often time-consuming and hence expensive to conduct, but they offer more control over potential confounding variables and are less susceptible to selection, allocation, and recall bias.

- Most contouring is done with CT scans. However, some recent research has investigated the use of MRI, because of its perceived superiority to CT in terms of soft tissues detailing. While some articles have compared MRI to CT scans for auto-contouring (such as Hague 2020[21]), the evidence base remains limited. SCM commentary also noted that MRI datasets have only limited applicability for dose calculation, which is required in radiotherapy treatment planning. CT datasets are more routinely used, so although MR images may have superior soft tissue detail, it is often not appropriate for creating the radiotherapy treatment plan alone. Therefore, CT datasets must also be used. Further, it was noted that some structures may be better visualised using MRI rather than CT (or potentially vice versa), and that it may be that a software works better with one modality than another for specific structures.

13.2. Integration into the NHS

In addition to the formal evidence gaps identified above, there are wider issues to consider with regards to the integration of AI-based auto-contouring technology in the NHS.

Company submissions report that there is already some use of the scoped auto-contouring technologies within certain NHS trusts. If further adopted, wider use of the technology would involve upscaling across more trusts. Potential challenges include ensuring sufficient appropriate staff resource and training to deliver such interventions—including promoting critical reflection on current clinical practice. There is also the question of deciding for which structures auto-contouring technology should be used—as noted above, the evidence base is focused mostly on the head and neck and pelvis. Beyond these two anatomical sites, the evidence base is weaker.

Related to this is the fact that SCMs reported that contouring is already challenging for structures where there is no shared consensus on definitions and process. For example, different protocols exist for inguinal nodes and for head and neck lymph nodes. It is therefore very unlikely that AI-based systems, however good, will be acceptable to all users within or between hospitals. This raises the need to ensure adequate quality assurance and safety protocols. SCMs notify us that validation of protocols is currently done via a mixture of metrics (such as DICE) and visual review by experts/trained staff on a patient-by-patient basis. Such processes would have to be incorporated into any workflow.

One of the key questions that would need to be considered if further integration is to take place is whether to promote “off the shelf” AI-contouring use or rather to encourage hospitals to use local training datasets. The previous paragraph noted that there are different protocols between hospitals, at least for some structures, and SCMs have suggested that off the shelf use would encourage harmonisation. It may also save time and resource, as re-training algorithms could be very demanding in terms of number of ‘perfectly’ manually contoured datasets required. Finally, re-training could result in errors or bias, if staff doing this are inexperienced, or could perpetuate unwarranted variation in local institutions.

On the other hand, the SCMs also noted that there is a risk in off the shelf use that if only commercially trained software packages are available, especially for rare tumour sites, this may not be an economically viable model for a commercial manufacturer to create and maintain. SCMs also noted that local training may be useful for site-specific requirements and specific patient cohorts, such as may be found in paediatrics. Alternatively, there may be a middle way between “off the shelf” and local training. One SCM did suggest that if training were to occur, then perhaps this would best be done at a national level, to ensure consistency across centres. Finally, a barrier to off the shelf use, or NHS wide solutions, is that treatment protocols and equipment can differ significantly between hospitals.

13.3. Ongoing studies

Ongoing studies were identified for three of the technologies, either through company submissions or EAG searches, and are listed below in Table 8. It is uncertain whether any of them will address the evidence gaps and integration issues described.

Table 8: Ongoing studies

DLCExpert	Limbus	MVision
[redacted]	CTRI/2019/09/021316[96] Evaluation of performance of artificial intelligence based auto-contouring software in delineation of tumor and organs-at-risk for image-based radiotherapy planning.	[redacted]

In addition to the studies described in Table 8, NICE have informed the EAG that, as part of the NHS England AI awards[*], King's technology evaluation centre (KiTEC) is setting up two studies to gather further evidence on a number of AI-based auto-contouring technologies. One study will focus on the qualitative aspects, including acceptability of the contours and ease of integration into the treatment planning workflow. The other will gather evidence to help inform a future cost-effectiveness analysis. There studies are planned to report within two years, by 2025.

13.4. Key areas for evidence generation

Given the gaps and issues raised in this section, the EAG presents some specific evidence generation recommendations in Table 9. Any proposed trials should follow the best practice recommendations of the Radiotherapy Trials Quality Assurance (RTTQA).[†]

Table 9: Evidence generation recommendations

Research question	Possible study design	Outcomes
1. Which technology or technologies are most suitable for NHS use?	Comparative cohort studies of two or more included technologies in a prospective RWE setting with manual contours as a reference standard and blinded assessment.	Time-saving metrics, dosimetrics, and qualitative assessment (blinded scale scoring).

[*] For more information, see the AI in Health and Care Award website, available at: https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ai-health-and-care-award-winners/

[†] For more information, see https://rttrialsqa.org.uk/

2. Do AI-based auto-contouring technologies improve process efficiency and/or patient outcomes in the NHS?	Cluster randomised controlled trial or, failing that, a before and after cohort study	Throughput, patient outcomes, adverse events.
3. What is the cost effectiveness of AI auto-contouring interventions to manual contouring or atlas-based contouring, within an NHS context?	Cluster randomised controlled trial or, failing that, a before and after cohort study	Time saving metrics, technology costs, resource use estimates (clinician time, staff training costs), patient reported outcomes i.e. QALYs (if appropriate)

Abbreviations: AI artificial intelligence, NHS National Health Service, QALY quality-adjusted life year, RWE real world evidence

14. CONCLUSIONS

Based on clinical opinion to the EAG and evidence from published literature, AI auto-contouring produces contours that are mostly either ready to use or just need minor edits. However, AI software is less able to consistently create such accurate contours for certain organs at risk, typically organs with a small volume or that have an elongated shape, meaning that at the present all AI produced auto-contours should be checked by a clinician before use. I.e., that auto-contouring facilitates, rather than replaces, manual contouring. Where AI-contours were compared with atlas contours, the AI approach consistently outperformed atlas contouring.

AI auto-contouring (including editing and reviewing time) also appears to result in time savings when compared to current contouring approaches used in clinical practice (albeit there is considerable time variation based on tumour site and the structures typically contoured). However, due to the lack of published cost effectiveness evidence and heterogeneity in the pricing and reimbursement strategy for each technology, it was not possible to draw firm conclusions on the cost effectiveness of AI auto-contouring compared to manual or atlas-based segmentation approaches.

There was not sufficient evidence to draw a conclusion on which of the eleven technologies were either most clinical or cost-effective. The three technologies with the largest evidence base were: DLCExpert (Mirada Medical) and Limbus Contour (Limbus AI, AMG Medtech), followed by MIM Contour ProtégéAI (MIM Software).

More robust evidence is required for the following: 1) AI intervention costs in clinical practice and how these technologies impact on healthcare resource use and/or patient outcomes in a UK context. 2) The impact of local NHS training sets rather than an "off the shelf" approach, from both a clinical and harmonisation/cost-effectiveness point of view. 3) AI auto-contouring effectiveness for body structures beyond head and neck and the pelvis, and identification of those organs at risk that are particularly susceptible to being poor contoured. 4) Relative clinical and cost-effectiveness of auto-contouring using MRI vs CT scans. 5) Direct, head-to-head trials between alternative AI auto-contouring technologies.

15. REFERENCES

1. The Royal College of Radiologists. RCR Clinical oncology census report 2021 2021. Available from: https://www.rcr.ac.uk/rcr-clinical-oncology-census-report-2021.

2. National Institute for Health and Care Excellence. Artificial intelligence auto-contouring for radiotherapy treatment planning: early value assessment: final scope 2023. Available from: https://www.nice.org.uk/guidance/indevelopment/gid-hte10015/documents.

3. ESTRO. Guidelines. Available from: https://www.estro.org/Science/Guidelines.

4. The Royal College of Radiologists. Radiotherapy target volume definition and peer review, second edition – RCR guidance 2022. Available from: https://www.rcr.ac.uk/publication/radiotherapy-target-volume-definition-and-peer-review-second-edition-rcr-guidance.

5. DICOM PS3.1 2023b - Introduction and Overview. 1 Scope and Field of Application: DICOM; 2023. Available from: https://dicom.nema.org/medical/dicom/current/output/chtml/part01/chapter_1.html#sect_1.1.

6. Ginn JS, Gay HA, Hilliard J, Shah J, Mistry N, Mohler C, et al. A clinical and time savings evaluation of a deep learning automatic contouring algorithm. Medical dosimetry : official journal of the American Association of Medical Dosimetrists. 2023;48(1):55-60.

7. Hu Y, Nguyen H, Smith C, Chen T, Byrne M, Archibald-Heeren B, et al. Clinical assessment of a novel machine-learning automated contouring tool for radiotherapy planning. J Appl Clin Med Phys. 2023:e13949.

8. Marschner S, Datar M, Gaasch A, Xu Z, Grbic S, Chabin G, et al. A deep image-to-image network organ segmentation algorithm for radiation treatment planning: principles and evaluation. Radiat Oncol. 2022;17(1):129.

9. Peng JL, McDonald DG, Godwin W, Warwick L, Roles SA, Maynard M, et al. Clinical Feasibility of Commercial Artificial Intelligence-Based Auto Contouring of Target Volumes and Organs-at-Risk in Breast Cancer Patients. International Journal of Radiation Oncology Biology Physics. 2022;114(3):e585.

10. Ginn J, Gay H, Hilliard J, Shah J, Mistry N, Mohler C, et al. A Clinical and Time Savings Evaluation of a Commercial Deep Learning Automatic Contouring Algorithm. Med Phys. 2022;49(6):e746.

11. Maduro Bustos L, Doyle L, Nurbag, ova D, Noonan J, Sarkar A, et al. To Evaluate the Clinical Implementation Feasibility of the Siemen's Auto-Contouring

Deep-Learning Solution, AI-Rad Companion Organs RT. Med Phys. 2022;49(6):e461.

12. Blanchard P, Gregoire VG, Petit C, Milhade N, Allajbej A, Nguyen TVF, et al. A Blinded Prospective Evaluation Of Clinical Applicability Of Deep Learning-Based Auto Contouring Of OAR For Head and Neck Radiotherapy. International Journal of Radiation Oncology Biology Physics. 2020;108(3):e780-e1.

13. Nachbar M, Russo ML, Boeke S, Wegener D, Boldt J, Butzer S, et al. Development and quantitative evaluation of AI-based pelvic MRI autocontouring for adaptive MRgRT. Radiother Oncol. 2021;161:S58-S9.

14. Buatti JS, Kirby N, Li R, De Oliveira M, Kabat C, Papanikolaou N, et al. Evaluation of Automated and Manual Contours for Head and Neck Cancers. Med Phys. 2022;49(6):e674-e5.

15. Costea M, Biston M, Gregoire V, Saruut D. Comparison of segmentation algorithms for organs at risk delineation on head-and-neck CT images. Radiother Oncol. 2021;161:S704-S5.

16. Rivera S, Lombard A, Pasquier D, Wong S, Limkin E, Auzac G, et al. AI-driven quality insurance for delineation in radiotherapy breast clinical trials. Radiother Oncol. 2020;152:S953.

17. Gregoire V, Blanchard P, Allajbej A, Petit C, Milhade N, Nguyen F, et al. Deep learning auto contouring of OAR for HN radiotherapy: a blinded evaluation by clinical experts. Radiother Oncol. 2020;152:S379-S80.

18. Leyva M, Wang D, McAllister N, Gutierrez A, Tolakanahalli R. Evaluation of a Commercial Convolution Neural Network Based Auto-Segmentation Software. Med Phys. 2022;49(6):e674.

19. Bice N, Patel B, Milien P, McCarthy A, Cheng P, Rembish J, et al. Statistical Evaluation of a Commercial Deep-Learning-Based Automatic Contouring Software. Med Phys. 2022;49(6):e466.

20. Marasco J, Hendley S, Wong J, Granatowicz A, Besemer A, Zhou S, et al. Evaluation of An Auto-Segmentation Tool On Full Field-Of-View and Limited Field-Of-View Cone Beam Computed Tomography. Med Phys. 2022;49(6):e727-e8.

21. Hague C, Beasley W, McPartlin A, Owens S, Price G, Saud H, et al. Clinical evaluation of deep learning autocontouring in prostate and head and neck cancer. Radiother Oncol. 2020;152:S950.

22. van Dijk LV, Van den Bosch L, Aljabar P, Peressutti D, Both S, J H M Steenbakkers R, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2020;142:115-23.

23. Vaassen F, Hazelaar C, Canters R, Peeters S, Petit S, van Elmpt W. The impact of organ-at-risk contour variations on automatically generated treatment plans for NSCLC. Radiother Oncol. 2021;163:136-42.

24. Walker Z, Bartley G, Hague C, Kelly D, Navarro C, Rogers J, et al. Evaluating the Effectiveness of Deep Learning Contouring across Multiple Radiotherapy Centres. Physics and imaging in radiation oncology. 2022;24:121-8.

25. Vaassen F, Boukerroui D, Looney P, Canters R, Verhoeven K, Peeters S, et al. Real-world analysis of manual editing of deep learning contouring in the thorax region. Physics and imaging in radiation oncology. 2022;22:104-10.

26. Brouwer CL, Boukerroui D, Oliveira J, Looney P, Steenbakkers RJHM, Langendijk JA, et al. 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. 2020;16:54-60.

27. Brunenberg EJL, Steinseifer IK, van den Bosch S, Ka, ers JHAM, Brouwer CL, et al. External validation of deep learning-based contouring of head and neck organs at risk. Physics and imaging in radiation oncology. 2020;15:8-15.

28. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2018;126(2):312-7.

29. van de Glind H, van Bruggen IG, Langendijk JA, Both S, Brouwer CL. No need for manual adjustments of deep learning segmentation in oropharyngeal cancer? Radiother Oncol. 2022;170:S474-S5.

30. Alty K, Marshall D, Bird A, Powis R, Webster G. Evaluation of the dosimetric impact of autodelineation uncertainties in prostate radiotherapy. Radiother Oncol. 2022;170:S1462.

31. Boukerroui D, Baker J, McQuinlan Y, Riegel A, Cao Y, Gooding M, et al. Investigating the Expected Impact of Auto-Contouring in Clinical Practice: A Cohort Analysis. Med Phys. 2022;49(6):e692.

32. Vaassen F, Canters R, Lubken I, Mannens J, Van Elmpt W. Large scale analysis of the clinical implementation of deep learning contouring in the thorax region. Radiother Oncol. 2021;161:S767-S8.

33. Gibbons E, Hoffmann M, Chick B, Hodgson A, Marjoribanks J, Westhuyzen J. Clinical evaluation of deep learning and atlas-based auto-segmentation for organs at risk. J Med Imaging Radiat Oncol. 2021;65:246-7.

34. Geng H, Men K, Yom SS, Xia P, Xiao Y. Deep Learning Auto-Segmentation Model for HN005 Contour Quality Assurance. International Journal of Radiation Oncology Biology Physics. 2021;111(3):e506-e7.

35. Brunenberg E, Lazareva A, Steinseifer I, Smeenk RJ, Aljabar P, Monshouwer R. Global and local evaluation of deep learning contouring of prostate CT. Radiother Oncol. 2020;152:S940.

36. South C, Navarro C, Rickard DJ, Lynch J, Wood K, Nisbet A, et al. A Prospective Clinical Evaluation of Mirada DLCExpert Auto-Contouring for Head and Neck OARs. Radiother Oncol. 2020;152:S238.

37. Liu A, Li R, Han C, Du D, Sampath S, Amini A, et al. Comparative Clinical Evaluation Of Deep-Learning-Based Algorithms In Auto-Segmentation Of Organs-At-Risk For Head And Neck Cancers. International Journal of Radiation Oncology Biology Physics. 2020;108(3):e817.

38. Poortmans P, Henry A, re A, Aljabar P, Baggs R, Gooding M, et al. Deep Learning for automatic contouring of clinical target volumes in breast cancer patients. Radiother Oncol. 2019;133.

39. Aljabar P, Peressutti D, Brunenberg E, Smeenk R, Van Leeuwen R, Gooding M. Comparison of auto-contouring methods for regions of interest in prostate CT. Radiother Oncol. 2018;127:S218-S9.

40. Bakker H, Peressutti D, Aljabar P, Van Dijk LV, Van Den Bosch L, Gooding M, et al. Quantitative evaluation of deep learning contouring of head and neck organs at risk. Radiother Oncol. 2018;127:S217-S8.

41. Gooding M, Smith A, Peressutti D, Aljabar P, Evans E, Gwynne S, et al. PV-0531: Multi-centre evaluation of atlas-based and deep learning contouring using a modified Turing Test. Radiother Oncol. 2018;127:S282-S3.

42. Duan J, Bernard M, Downes L, Willows B, Feng X, Mourad WF, et al. Evaluating the clinical acceptability of deep learning contours of prostate and organs-at-risk in an automated prostate treatment planning process. Med Phys. 2022;49(4):2570-81.

43. Chen W, Li Y, Dyer B, on A, Feng X, Rao S, et al. Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. Radiat Oncol. 2020;15(1):176.

44. Radici L, Ferrario S, Borca VC, Cante D, Paolini M, Piva C, et al. Implementation of a Commercial Deep Learning-Based Auto Segmentation Software in Radiotherapy: Evaluation of Effectiveness and Impact on Workflow. Life (Basel, Switzerland). 2022;12(12).

45. Wong J, Huang V, Wells D, Giambattista J, Giambattista J, Kolbeck C, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. Radiat Oncol. 2021;16(1):101.

46. Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, et al. 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 : journal of the European Society for Therapeutic Radiology and Oncology. 2020;144:152-8.

47. Wong J, Huang V, Giambattista JA, Teke T, Kolbeck C, Giambattista J, et al. Training and Validation of Deep Learning-Based Auto-Segmentation Models for Lung Stereotactic Ablative Radiotherapy Using Retrospective Radiotherapy Planning Contours. Front Oncol. 2021;11:626499.

48. D'Aviero A, Re A, Catucci F, Piccari D, Votta C, Piro D, et al. Clinical Validation of a Deep-Learning Segmentation Software in Head and Neck: An Early Analysis in a Developing Radiation Oncology Center. International journal of environmental research and public health. 2022;19(15).

49. Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. Pract Radiat Oncol. 2021;11(1):e80-e9.

50. Kirkby C, Liu HW, Ghose A, Grendarova P, Seberg S, Li X, et al. Independent Assessment of a Commercial Automated Contouring Software for Lung SBRT. Med Phys. 2022;49(8):5661.

51. Kucharczyk M, Chytyk-Praznik K, Giambattista J, Kolbeck C, Chng N, Bala G, et al. A Randomized Blinded Assessment of a Machine Learning Based Autocontouring Tool for Breast Cancer Radiotherapy Compared to Peer-Reviewed Radiation Oncologist Contours. Radiother Oncol. 2022;174:S11.

52. Coughlan S, Biggar R, Stonton C, Axelsen A, editors. Qualitative evaluation of auto-segmented structures for radiotherapy planning using peer review guidelines. British Institute of Radiology - AI in Practice 2022 Meeting; 2022.

53. Wong J, Huang V, Wells D, Giambattista J, Kolbeck C, Otto K, et al. Implementation of Deep Learning-Based Auto-Segmentation for Radiotherapy Planning Structures: A Multi-Center Workflow Study. Radiother Oncol. 2020;150:S14-S5.

54. Wong J, Kolbeck C, Giambattista J, Giambattista JA, Huang V, Jaswal JK, editors. 2636 Deep Learning-Based Auto-Segmentation for Pelvic Organs at Risk and Clinical Target Volumes in Intracavitary High Dose Rate Brachytherapy. Global Oncology Radiation therapy in a changing world; 2020.

55. Wong J, Huang V, Giambattista JA, Teke T, Atrchian S. Validation of Deep Learning-based Auto-Segmentation for Organs at Risk and Gross Tumor Volumes in Lung Stereotactic Body Radiotherapy. International Journal of Radiation Oncology Biology Physics. 2019;105(1):E140.

56. Fong A, Swift CL, Wong J, McVicar N, Giambattista JA, Kolbeck C, et al. Automatic Deep Learning-based Segmentation of Brain Metastasis on MPRAGE MR

Images for Stereotactic Radiotherapy Planning. International Journal of Radiation Oncology Biology Physics. 2019;105(1):E134.

57. Wong J, Fong A, McVicar N, Smith SL, Giambattista JA, Wells DM, et al., editors. 144 - Comparing Deep Learning-based Auto-segmentation of Organs at Risk and Clinical Target Volumes to Expert Inter-Observer Variability in Radiotherapy Planning. Astro Annual Meeting; 2019; Chicago.

58. Urago Y, Okamoto H, Kaneda T, Murakami N, Kashihara T, Takemori M, et al. Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models. Radiat Oncol. 2021;16(1):175.

59. Lastrucci A, Meucci F, Baldazzi M, Marciello L, Cernusco NLV, Serventi E, et al. Comparative clinical evaluation of auto segmentation methods in contouring of prostate cancer. Onkologia i Radioterapia. 2022;16(3):5-7.

60. Lancellotta V. Evaluation Of Deep-Learning Auto-Segmentation Methods In Cervix Cancer Patients. Radiother Oncol. 2022;170:S265-S6.

61. Tsai P, Huang S, Press R, Shim A, Apinorasethkul C, Chen C, et al. Validation of a Commercial Artificial Intelligence Auto-Segmentation System for Head and Neck Treatment Site. Med Phys. 2022;49(6):e297.

62. Martinez H, Rich B, Young L, Yang F. Head-to-head performance comparison of two deep learning segmentation algorithms for radiotherapy planning: A study in prostate. Med Phys. 2021;48(6):e481.

63. Kruzer A, Wan H, Bending M, Halley C, Darkow D, Pittock D, et al., editors. Comparison of a 3D Convolutional Neural Network Segmentation Method to Traditional Atlas Segmentation for CT Head and Neck Contours. 2020 Joint AAPM | COMP Virtual Meeting Poster Presentation; 2020.

64. Cole NM, Wan H, Niedbala J, Dewaraja YK, Kruzer A, Pittock D, et al., editors. Impact of a 3D Convolution Neural Network Method On Liver Segmentation: An Accuracy and Time-Savings Evaluation. 2020 Joint AAPM | COMP Virtual Meeting Poster Presentation; 2020.

65. Halley C, Wan H, Kruzer A, Pittock D, Darkow D, Butler M, et al., editors. Improved Auto-Segmentation for CT Male Pelvis: Comparison of Deep Learning to Traditional Atlas Segmentation Methods. 2020 Joint AAPM | COMP Virtual Meeting Poster Presentation; 2020.

66. Kuisma A, Ranta I, Keyrilainen J, Suilamo S, Wright P, Pesola M, et al. Validation of automated magnetic resonance image segmentation for radiation therapy planning in prostate cancer. Physics and imaging in radiation oncology. 2020;13:14-20.

67. Maspero M, Tyyger MD, Seevinck PR, Tijssen RHN, Intven MPW, Van Den Berg CAT. Feasibility of MR-only rectum radiotherapy using a commercial prostate sCT generation solution. Radiother Oncol. 2018;127:S1140-S1.

68. Strolin S, Santoro M, Paolani G, Ammendolia I, Arcelli A, ra, et al. How smart is artificial intelligence in organs delineation? Testing a CE and FDA-approved Deep-Learning tool using multiple expert contours delineated on planning CT images. Front Oncol. 2023;13:1089807.

69. Kiljunen T, Akram S, Niemela J, Loyttyniemi E, Seppala J, Heikkila J, et al. A Deep Learning-Based Automated CT Segmentation of Prostate Cancer Anatomy for Radiation Therapy Planning-A Retrospective Multicenter Study. Diagnostics (Basel, Switzerland). 2020;10(11).

70. Suresh R, Niemela J, Akram S, Valdman A, Olsson CE. A Comparative Study Between AI-Generated, Real-Life Clinical as Well as Reference Rectal Volumes Defined in Accordance With the Swedish National STRONG Guidelines in Prostate Cancer Radiotherapy. International Journal of Radiation Oncology Biology Physics. 2021;111(3):e138.

71. Heikkila J, Viren T, Virsunen H, Vuolukka K, Voutilainen L, Sawabi R, et al. Comparison of different autosegmentation software for left-sided breast cancer patients. Radiother Oncol. 2020;152:S149-S50.

72. Oktay O, Nanavati J, Schwaighofer A, Carter D, Bristow M, Tanno R, et al. Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers. JAMA network open. 2020;3(11):e2027426.

73. Almberg SS, Lervag C, Frengen J, Eidem M, Abramova TM, Nordstr, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol. 2022;173:62-8.

74. Rigaud B, Anderson BM, Yu ZH, Gobeli M, Cazoulat G, Soderberg J, et al. Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer. Int J Radiat Oncol Biol Phys. 2021;109(4):1096-110.

75. Liu M, Granville D, Wilson B. Independent evaluation of RayStation's deep learning autosegmentation model for structures in the male pelvis. Med Phys. 2022;49(8):5634.

76. Sidorski G, Mazurier J, Pichon B, Pinel B, Jimenez G, Gallocher O, et al. Artificial intelligence-based contouring and planning algorithms for prostate cancer radiotherapy. Phys Med. 2021;92:S192-S3.

77. Borkvel MA, Gerskevits E, Adamson M, Kiitam I, Kolk MK, Poldveer MM. Evaluating the efficiency gain in using artificial intelligence based automatic segmentation tools in radiotherapy treatment planning. Phys Med. 2022;104:S129.

78. Rong Y, Chen Q, Yuan L, Qi X, Latifi K, Yang X, et al. Cross-Comparison of Multi-Platform AI Auto-Segmentation Tools Using Independent Multi- Institutional Datasets for Head and Neck Cancer. Med Phys. 2022;49(6):e781.

79. Liao Y, Injerd R, Tolekidis G, Joshi N, Turian J. Performance Evaluation of AI-Based Automatic Segmentation Modules for Head and Neck Cancer Patients. Med Phys. 2022;49(6):e671.

80. Yuan L, Chen Q, Rong Y, Al-Hallaq H, Benedict S, Cai B, et al. An Independent Evaluation of Six Commercially Available Deep Learning-Based Auto Segmentation Platforms Using Large Multi- Institutional Datasets. Med Phys. 2022;49(6):e403.

81. Gorgisyan J, Bengtsson I, Lempart M, Lerner M, Wiesl, er E, et al. Evaluation of two commercial deep learning OAR segmentation models for prostate cancer treatment. Radiother Oncol. 2022;170:S1582-S3.

82. Doolan P, Charalambous S, Roussakis Y, Leczynski A, Ferentinos K, Strouthos I, et al. A comparison of three commercial AI contouring solutions. Radiother Oncol. 2021;161:S1221-S2.

83. Hague C, McPartlin A, Lee LW, Hughes C, Mullan D, Beasley W, et al. An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2021;158:112-7.

84. Lin D, Lapen K, Sherer MV, Kantor J, Zhang Z, Boyce LM, et al. A Systematic Review of Contouring Guidelines in Radiation Oncology: Analysis of Frequency, Methodology, and Delivery of Consensus Recommendations. Int J Radiat Oncol Biol Phys. 2020;107(4):827-35.

85. eContour. Published Consensus Contouring Guidelines. Available from: https://www.econtour.org/references.

86. Mackay K, Bernstein D, Glocker B, Kamnitsas K, Taylor A. A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy. Clin Oncol (R Coll Radiol). 2023;35(6):354-69.

87. Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z, et al. The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. Radiation Oncology. 2021;16(1):113.

88. Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol. 2019;135:130-40.

89. Poel R, Rüfenacht E, Hermann E, Scheib S, Manser P, Aebersold DM, et al. The predictive value of segmentation metrics on dosimetry in organs at risk of the brain. Med Image Anal. 2021;73:102161.

90. Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical evaluation of deep learning and atlas based auto-contouring of bladder and rectum for prostate radiotherapy. Pract Radiat Oncol. 2020.

91. Vickers AJ, Thiruthaneeswaran N, Coyle C, Manoharan P, Wylie J, Kershaw L, et al. Does magnetic resonance imaging improve soft tissue sarcoma contouring for radiotherapy? BJR Open. 2019;1(1):20180022.

92. Postma AA, Das M, Stadler AA, Wildberger JE. Dual-Energy CT: What the Neuroradiologist Should Know. Curr Radiol Rep. 2015;3(5):16.

93. MAUDE Adverse Event Report: RAYSEARCH LABORATORIES AB (PUBL)EUGENIAVAEGEN 18C, RAYSTATION RADIATION THERAPY TREATMENT PLANNING SYSTEM. 2023.

94. Field Safety Notice, Medical Device Correction #109886: RayStation, RayPlan 9A, 9B, 10A, 10B, 11A, 11B and 12A including service packs. 2023.

95. Tang J, Cui Y, Li Q, Ren K, Liu J, Buyya R. Ensuring Security and Privacy Preservation for Cloud Data Services. ACM Comput Surv. 2016;49(1):Article 13.

96. Singh A, Dr Anusha R, Dr Priyanka A, Dr Umesh V, Mr Shrinidhi GC, Dr Asha K, et al. Evaluation of Artificial Intelligence based software for radiotherapy planning. 2019.

97. Ahn SH, Yeo AU, Kim KH, Kim C, Goh Y, Cho S, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. Radiat Oncol. 2019;14(1):213.

98. Alves N, Dias JM, Rocha H, Ventura T, Mateus J, Capela M, et al. Assessing the need for adaptive radiotherapy in head and neck cancer patients using an automatic planning tool. Reports of practical oncology and radiotherapy : journal of Greatpoland Cancer Center in Poznan and Polish Society of Radiation Oncology. 2021;26(3):423-32.

99. Aoyama T, Shimizu H, Kitagawa T, Yokoi K, Koide Y, Tachibana H, et al. Comparison of atlas-based auto-segmentation accuracy for radiotherapy in prostate cancer. Physics and Imaging in Radiation Oncology. 2021;19:126-30.

100. Azria D, Boldrini L, De Ridder M, Fenoglietto P, Gambacorta MA, Gevaert T, et al. AI surpassing human expert: a multi-centric evaluation for organ at risk delineation. Radiother Oncol. 2022;170:S408-S10.

101. Bakx N, Bluemink H, Hagelaar E, van der Sangen M, Theuws J, Hurkmans C. Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer. Physics and Imaging in Radiation Oncology. 2021;17:65-70.

102. Byun HK, Chang JS, Choi MS, Chun J, Jung J, Jeong C, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol. 2021;16(1):203.

103. Canters R, Taasti V, Hattu D, van Loon J, De Ruysscher D. Development and validation of an automated robust treatment planning method for lung proton therapy. Radiother Oncol. 2021;161:S356-S8.

104. Chan JW, Kearney V, Haaf S, Wu S, Bogdanov M, Dixit N, et al. Deep learning-based autosegmentation for head and neck radiotherapy. International Journal of Radiation Oncology Biology Physics. 2018;101(2):E17.

105. Chen W, Wang C, Zhan W, Jia Y, Ruan F, Qiu L, et al. A comparative study of auto-contouring softwares in delineation of organs at risk in lung cancer and rectal cancer. Sci Rep. 2021;11(1):23002.

106. Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2020;153:139-45.

107. Chuter R, Whitehurst P, Whitfield G, Lines D, Vasquez Osorio E, Green A, et al. Auto-contouring software comparison for brain SRS patients. Radiother Oncol. 2018;127:S556-S7.

108. Crouzen J, Zindler J, Wiggenraad R, Mast M, Lemmouy S, Hooijdonk CGV, et al. Quality and efficiency of automated organ at risk delineation on MRI in the brain. Radiother Oncol. 2021;161:S887-S8.

109. Ewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol. 2020;153:55-66.

110. Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. Med Phys. 2019;46(5):2169-80.

111. Feng X, Tustison NJ, Patel SH, Meyer CH. Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. Front Comput Neurosci. 2020;14:25-.

112. Feng X, Bernard ME, Hunter T, Chen Q. Improving accuracy and robustness of deep convolutional neural network based thoracic OAR segmentation. Phys Med Biol. 2020;65(7):07NT1-NT1.

113. Feng X, Chen Q. Organ-Specific Segmentation Versus Multi-Class Segmentation Using U-Net. Auto-Segmentation for Radiation Oncology: CRC Press; 2021. p. 125-32.

114. Fi, ra C. Fully-automated genetic treatment planning for volumetric modulated arc therapy: An Italian multi-center validation for prostate cancer. Med Phys. 2019;46(6):e343.

115. Fong E, Stewart D, Wong W, Smee R. Accuracy of automated segmentation of the central nervous system for radiotherapy planning. J Med Imaging Radiat Oncol. 2021;65:253.

116. Gan Y, Langendijk JA, Oldehinkel E, Sc, urra D, Sijtsema N, et al. A novel semi auto-segmentation method for head and neck adaptive radiotherapy. Radiother Oncol. 2021;161:S1412-S4.

117. Garcia Perez A, Reigosa Montes S, Lopez Medina A, Teijeiro Garcia AG, Vazquez Rodriguez J, Salvador Gomez FJ, et al. Feasibility of CBCT positioning in an MRI only Prostate Treatment Planning. Radiother Oncol. 2018;127:S1143-S4.

118. Ghimire K, Chen Q, Feng X. Patch-Based 3D UNet for Head and Neck Tumor Segmentation with an Ensemble of Conventional and Dilated Convolutions: Springer International Publishing; 2021 2021. 78-84 p.

119. Ghimire K, Chen Q, Feng X. Head and Neck Tumor Segmentation with Deeply-Supervised 3D UNet and Progression-Free Survival Prediction with Linear Model: Springer International Publishing; 2022 2022. 141-9 p.

120. Giaddui T, Glick A, Bollinger D, Zhong H, Phillips H, Nunez F, et al. Improving treatment planning quality, consistency, and efficiency using rapid and autoplanning: A feasibility study based on the NRG-HN002 clinical trial. International Journal of Radiation Oncology. 2016;96(2):E653.

121. Giaddui T, Bollinger D, Glick A, Zhong H, Phillips H, Nunez F, et al. Toward improving treatment planning quality and efficiency using knowledge engineering and autoplanning: A study based on NRGHN001 clinical trial. International Journal of Radiation Oncology. 2016;96(2):E656-E7.

122. Gleeson I, Bolger N, Chun H, Hutchinson K, Klodowska M, Mehrer J, et al. Implementation of automated personalised breast radiotherapy planning techniques with scripting in Raystation. The British journal of radiology. 2023;96(1144):20220707.

123. Gooding MJ, Boukerroui D, Vasquez Osorio E, Monshouwer R, Brunenberg E. Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy. Physics and imaging in radiation oncology. 2022;24:152-8.

124. Gungor G, Klausner G, Gur G, Serbez I, Temur B, Caffaro A, et al. AI-based OAR delineation in brain T1w-MRI: Overcoming Inter- and Intra-observer variability. Radiother Oncol. 2022;170:S1674-S5.

125. Hammers JE, Pirozzi S, Lindsay D, Kaidar-Person O, Tan X, Chen RC, et al. Evaluation of a commercial DIR platform for contour propagation in prostate cancer patients treated with IMRT/VMAT. J Appl Clin Med Phys. 2020;21(2):14-25.

126. He Y, Zhang S, Luo Y, Yu H, Fu Y, Wu Z, et al. Quantitative Comparisons of Deep-learning-based and Atlas-based Auto- segmentation of the Intermediate Risk Clinical Target Volume for Nasopharyngeal Carcinoma. Current medical imaging. 2022;18(3):335-45.

127. Hedrick SG, Petro S, Ward A, Morris B. Validation of automated complex head and neck treatment planning with pencil beam scanning proton therapy. J Appl Clin Med Phys. 2022;23(2):e13510.

128. Hern, ez S, Parkes J, Burger H, Nguyen C, Rhee D, et al. Automating Contouring and Treatment Planning for Pediatric 3D-Craniospinal Irradiation Therapy. Med Phys. 2022;49(6):e336-e7.

129. Johansson W, Opp D, Tichacek C, Libby B, Zhang G, Redler G, et al. Evaluation of Two Treatment Planning Systems for Single Isocenter Multiple Metastases Stereotactic Radiosurgery Treatment Planning and Delivery. Med Phys. 2022;49(6):e968.

130. Khalifa A, Winter J, Navarro I, McIntosh C, Purdie TG. Domain adaptation of automated treatment planning from computed tomography to magnetic resonance. Phys Med Biol. 2022;67(12).

131. Loap P, Botticella A, De Marzi L, Levy A, Bolle S, Colame S, et al. AI-based cardiac sub-structures segmentation for safer radiotherapy planning. Cancer Res. 2023;83(5).

132. Lu SL, Xiao FR, Cheng JC, Yang WC, Cheng YH, Chang YC, et al. Randomized Multi-Reader Evaluation of Automated Detection and Segmentation of Brain Tumors in Stereotactic Radiosurgery with Deep Neural Networks. Neuro Oncol. 2021;23(9):1560-8.

133. Magallon-Baro A, Milder MTW, Granton PV, den Toom W, Nuyttens JJ, Hoogeman MS. Impact of Using Unedited CT-Based DIR-Propagated Autocontours on Online ART for Pancreatic SBRT. Front Oncol. 2022;12:910792.

134. Nash D, McWilliam A, Palmer AL, Vasquez Osorio E. The impact of using propagated contours for automatic replanning on CBCTs for H&N radiotherapy. Radiother Oncol. 2022;170:S349-S51.

135. Newman N, Stathakis S, Thorwarth D, Zips D, Nachbar M, iban S, et al. Clinical evaluation of organs at risk automatic-segmentation for T2-weigthed MRI. Radiother Oncol. 2022;170:S296-S7.

136. Nicolae A, ru, Semple M, Lu L, Smith M, Chung H, et al. Conventional vs machine learning-based treatment planning in prostate brachytherapy: Results of a Phase I randomized controlled trial. Brachytherapy. 2020;19(4):470-6.

137. O'Hara C, Bird D, Speight R, Andersson S, Nilsson R, Al-Qaisieh B. Assessment of CBCT based synthetic CT generation accuracy for adaptive radiotherapy planning. Radiother Oncol. 2022;170:S342-S3.

138. Okada H, Ito M, Minami Y, Nakamura K, Asai A, Adachi S, et al. Automatic one-click planning for hippocampal-avoidance whole-brain irradiation in RayStation. Medical dosimetry : official journal of the American Association of Medical Dosimetrists. 2022;47(1):98-102.

139. Shelley CE, Bolt MA, Hollingdale R, Chadwick SJ, Barnard AP, Rashid M, et al. Implementing cone-beam computed tomography-guided online adaptive radiotherapy in cervical cancer. Clinical and translational radiation oncology. 2023;40:100596.

140. Sibolt P, Andersson LM, Calmels L, Sjostrom D, Bjelkengren U, Geertsen P, et al. Clinical implementation of artificial intelligence-driven cone-beam computed tomography-guided online adaptive radiotherapy in the pelvic region. Physics and imaging in radiation oncology. 2021;17:1-7.

141. Tsui G, Tsang DS, McIntosh C, Purdie TG, Bauman G, Laperriere N, et al. Automated Machine-Learning Radiotherapy Planning for Pediatric and Adult Brain Tumours. Journal of Medical Imaging and Radiation Sciences. 2021;52(2):S3.

142. van De Sande, D., Bluemink H, Kneepkens E, Bakx N, Hagelaar E, Sharabiani M, et al. Development of artificial intelligence based treatment planning for locally advanced breast cancer. Radiother Oncol. 2021;161:S656-S8.

143. Visak J, Inam E, Meng B, Wang S, Parsons D, Nyugen D, et al. Evaluating machine learning enhanced intelligent-optimization-engine (IOE) performance for ethos head-and-neck (HN) plan generation. J Appl Clin Med Phys. 2023:e13950.

144. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Med Phys. 2020;47(9):e929-e50.

145. Wang J, Yang C, Qu B, Ma L, Fan W, Liu B, et al. Clinical evaluation of atlas and deep learning based automatic contouring for nasopharyngeal carcinoma. Med Phys. 2019;46(6):e448.

146. Wang JY, Zheng QZ, Zhang HJ, Yang W, Qu BL, Xu SP. Comparative study of two software tools on the atlas-based auto-segmentation of organs-at-risk in cervical cancer. Chin J Cancer Prev Treat. 2019;26(24):1889-94.

147. Yang Y, Shao K, Zhang J, Chen M, Chen Y, Shan G. Automatic Planning for Nasopharyngeal Carcinoma Based on Progressive Optimization in RayStation Treatment Planning System. Technol Cancer Res Treat. 2020;19:1533033820915710.

148. Yedekci Y, Gultekin M, Sari SY, Yildiz F. Automatic contouring using deformable image registration for tandem-ring or tandem-ovoid brachytherapy. Journal of Contemporary Brachytherapy. 2022;14(1):72-9.

149. Yedekci Y, Gultekin M, Sari SY, Yildiz F. Improving normal tissue sparing using scripting in endometrial cancer radiation therapy planning. Radiation and environmental biophysics. 2023.

150. Zhong Y, Yang Y, Fang Y, Wang J, Hu W. A Preliminary Experience of Implementing Deep-Learning Based Auto-Segmentation in Head and Neck Cancer: A Study on Real-World Clinical Cases. Front Oncol. 2021;11:638197.

16. APPENDICES

Appendix A: Searches for clinical and cost effectiveness evidence

Table 10: Resources searched for clinical and cost effectiveness studies

Database/Resource	Host	Date Searched	Results
MEDLINE and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily	Ovid	03.05.23	241
Embase	Ovid	03.05.23	837
Cochrane Database of Systematic Reviews	Cochrane Library: Wiley	04.05.23	5
Cochrane CENTRAL		04.05.23	62
INAHTA HTA database	https://database.inahta.org/	04.05.23	6
Company websites		04.05.23	32
ClinicalTrials.gov	http://www.clinicaltrials.gov/	03.05.23	56
WHO ICTRP	https://trialsearch.who.int/	03.05.23	12
NICE Guidelines		04.05.23	6
SIGN Guidelines		04.05.23	0
MHRA	https://www.gov.uk/drug-device-alerts		1
MAUDE	https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm		1
ScharrHUD	https://www.scharrhud.org/	03.05.23	0
CEA Registry	https://cear.tuftsmedicalcenter.org/	03.05.23	0
Total records retrieved			1259
Total records after deduplication			933

Ovid MEDLINE(R) ALL

#	Searches	Results
1	(AI-rad companion* or "Art-plan" or Autocontour or DLCexpert* or DLC-expert* or "DLC expert*" or INTContour or "INT Contour" or limbusAI or "limbus AI" or "limbus-AI" or "Limbus Contour" or "MIM contour" or ProtegeAI or MRCAT or	1305

	MVision* or Osairis or Osiris or "Ray station*" or Raystation* or “workflow box”).af.	
2	("AMG medtech" or mirada or philips or raysearch or "ray search" or "Oncology systems" or therapanacea or "MIM software" or radformation or carina or "Siemens Healthineers" or MVision or "Cambridge University Hospitals NHS Foundation Trust").in.	21511
3	1 or 2	22752
4	Organs at Risk/	4663
5	("organ*?at?risk*" or "organs-at-risk").ti,ab,kw,kf.	5931
6	exp *Radiotherapy/	113501
7	("clinical target volume" or CTV or "target volume" or "planning target volume" or PTV or "gross tumour volume" or "gross tumor volume" or GTV).ti,ab,kw,kf.	19191
8	(radiotherap* or irradiation* or "gamma knife" or "cyberknife" or "linear accelerator" or linac or wbrt or (radiation adj2 (therap* or dose*))).ti,ab.	496046
9	or/4-8	539523
10	((AI or intelligen* or auto* or radiomic*) and (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*)).ti,ab.	141438
11	(((deep* or machine*) adj2 learn*) and (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*)).ti,ab.	25391
12	or/10-11	155988
13	3 and 9 and 12	157
14	((AI or intelligen* or (deep adj2 learn*) or (machine adj2 learn*)) adj3 (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*)).ti,ab. /freq=2	749
15	14 and 9	102
16	13 or 15	241

Embase

#	Searches	Results
1	(AI-rad companion* or "Art-plan" or Autocontour or DLCexpert* or DLC-expert* or "DLC expert*" or INTContour or "INT Contour" or limbusAI or "limbus AI" or "limbus-AI" or "Limbus Contour" or "MIM contour" or ProtegeAI or MRCAT or MVision* or Osairis or Osiris or "Ray station*" or Raystation* or "workflow box").af.	3029
2	("AMG medtech" or mirada or philips or raysearch or "ray search" or "Oncology systems" or therapanacea or "MIM software" or radformation or carina or "Siemens Healthineers" or MVision or "Cambridge University Hospitals NHS Foundation Trust").in.	24497

3	1 or 2	27362
4	Organs at Risk/ or exp clinical target volume/	15992
5	("organ*?at?risk*" or "organs-at-risk").ti,ab,kw,kf.	13249
6	exp *Radiotherapy/	256603
7	("clinical target volume" or CTV or "target volume" or "planning target volume" or PTV or "gross tumour volume" or "gross tumor volume" or GTV).ti,ab,kw,kf.	43208
8	(radiotherap* or irradiation* or "gamma knife" or "cyberknife" or "linear accelerator" or linac or wbrt or (radiation adj2 (therap* or dose*))).ti,ab.	692274
9	or/4-8	769069
10	((AI or intelligen* or auto* or radiomic*) and (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*)).ti,ab.	204823
11	(((deep* or machine*) adj2 learn*) and (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*)).ti,ab.	33577
12	or/10-11	222878
13	3 and 9 and 12	596
14	((AI or intelligen* or (deep adj2 learn*) or (machine adj2 learn*)) adj3 (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*)).ti,ab. /freq=2	1197
15	14 and 9	325
16	13 or 15	842

Cochrane Library

#1 (("AI-rad" NEXT companion*) OR Art-plan OR Autocontour OR DLCexpert* OR DLC-expert* OR ("DLC" NEXT expert*) OR INTContour OR "INT Contour" OR limbusAI OR "limbus AI" OR limbus-AI OR "Limbus Contour" OR "MIM contour" OR ProtegeAI OR MRCAT OR MVision* OR Osairis OR Osiris OR ("Ray" NEXT station*) OR Raystation* OR “workflow box”) 89

#2 ("AMG medtech" OR mirada OR philips OR raysearch OR "ray search" OR "Oncology systems" OR therapanacea OR "MIM software" OR radformation OR carina OR "Siemens Healthineers" OR MVision OR "Cambridge University Hospitals NHS Foundation Trust") 1620

#3 #1 OR #2 1693

#4 ("organ* at risk*":ti,ab OR organs-at-risk:ti,ab) 124

#5 MeSH descriptor: [Radiotherapy] explode all trees 10296

#6 ("clinical target volume":ti,ab OR CTV:ti,ab OR "target volume":ti,ab OR "planning target volume":ti,ab OR PTV:ti,ab OR "gross tumour volume":ti,ab OR "gross tumor volume":ti,ab OR GTV:ti,ab) 1734

#7 (radiotherap*:ti,ab OR irradiation*:ti,ab OR "gamma knife":ti,ab OR cyberknife:ti,ab OR "linear accelerator":ti,ab OR linac:ti,ab OR wbrt:ti,ab OR (radiation:ti,ab NEAR/2 (therap*:ti,ab OR dose*:ti,ab))) 41486

#8 #4 OR #5 OR #6 OR #7 44223

#9 ((AI:ti,ab OR intelligen*:ti,ab OR auto*:ti,ab OR radiomic*:ti,ab) AND (contour*:ti,ab OR autocontour*:ti,ab OR ("auto" NEXT contour*):ti,ab OR segment*:ti,ab OR plan*:ti,ab OR optimi*:ti,ab)) 10968

#10 (((deep*:ti,ab OR machine*:ti,ab) NEAR/2 learn*:ti,ab) AND (contour*:ti,ab OR autocontour*:ti,ab OR ("auto" NEXT contour*):ti,ab OR segment*:ti,ab OR plan*:ti,ab OR optimi*:ti,ab)) 784

#11 #9 OR #10 11342

#12 #3 AND #8 AND #11 25

#13 ((AI:ti,ab OR intelligen*:ti,ab OR (deep:ti,ab NEAR/2 learn*:ti,ab) OR (machine:ti,ab NEAR/2 learn*:ti,ab)) NEAR/3 (contour*:ti,ab OR autocontour*:ti,ab OR ("auto" NEXT contour*):ti,ab OR segment*:ti,ab OR plan*:ti,ab OR optimi*:ti,ab)) 196

#14 #13 and #8 46

#15 #12 or #14 67

= 5 reviews and 62 trials

INAHTA

((radiotherap* or irradiation* or "gamma knife" or "cyberknife" or "linear accelerator" or linac or wbrt*) AND (contour* or autocontour* or "auto contour*" or segment* or plan* or optimi*) AND (AI or intelligen* or auto* or radiomic* or machine*)) OR ((AI-rad companion* or "Art-plan" or Autocontour or DLCexpert* or DLC-expert* or "DLC expert*" or INTContour or "INT Contour" or limbusAI or "limbus AI" or "limbus-AI" or "Limbus Contour" or "MIM contour" or ProtegeAI or MRCAT or MVision* or Osairis or Osiris or "Ray station*" or Raystation* or “workflow box”))

= 6 hits

ClinicalTrials.gov

Search string	Results
autocontour/all studies	0
autocontouring/all studies	0

artificial intelligence AND radiotherapy AND contour/all studies	6
machine learning AND radiotherapy AND contour/all studies	1
artificial intelligence AND radiotherapy AND plan/all studies	18
machine learning AND radiotherapy AND plan/all studies	16
artificial intelligence AND radiotherapy AND optimization/all studies	4
machine learning AND radiotherapy AND optimization/all studies	4
AI-rad companion/all studies	0
Art-plan/all studies	1
Autocontour/all studies	0
DLCexpert/all studies	0
DLC-expert/all studies	0
DLC expert/all studies	0
INTContour/all studies	0
INT Contour/all studies	1
limbusAI/all studies	0
limbus AI/all studies	0
limbus-AI/all studies	0
Limbus Contour/all studies	1
MIM contour/all studies	3
ProtegeAI/all studies	0
MRCAT/all studies	0
MVision/all studies	0
Osairis/all studies	0
Raystation/all studies	1
workflow box/all studies	0

ICTRP (basic search)

Search string	Results
Autocontour	0
Autocontouring	0
artificial intelligence AND radio* AND contour*	2
machine learning AND radiotherapy AND contour	0
artificial intelligence AND radio* AND plan*	6
machine learning AND radio* AND plan*	2

artificial intelligence AND radio* AND optimi*	0
machine learning AND radio* AND optimi*	1
AI-rad companion	0
Art-plan	0
Autocontour	0
DLCexpert	0
DLC-expert	0
DLC expert	0
INTContour	0
INT Contour	0
limbusAI	0
limbus AI	0
limbus-AI	0
Limbus Contour	1
MIM contour	0
ProtegeAI	0
MRCAT	0
MVision	0
Osairis	0
Raystation	0
Workflow box	0

CEA Registry

Search string	Results
contour* OR auto-contour* or autocontour*	0
Artificial intelligence	5 (0)
Machine learning	5 (0)

ScharrHUD

Search string	Results
contour* OR auto-contour* or autocontour*	0
Artificial intelligence	0
Machine learning	0

(radio* or radiation) AND (map* or plan* or segment* or contour*)	6 (0)

NICE

contour or contouring or autocontour or autocontouring or artificial intelligence or machine learning (as separate searches)

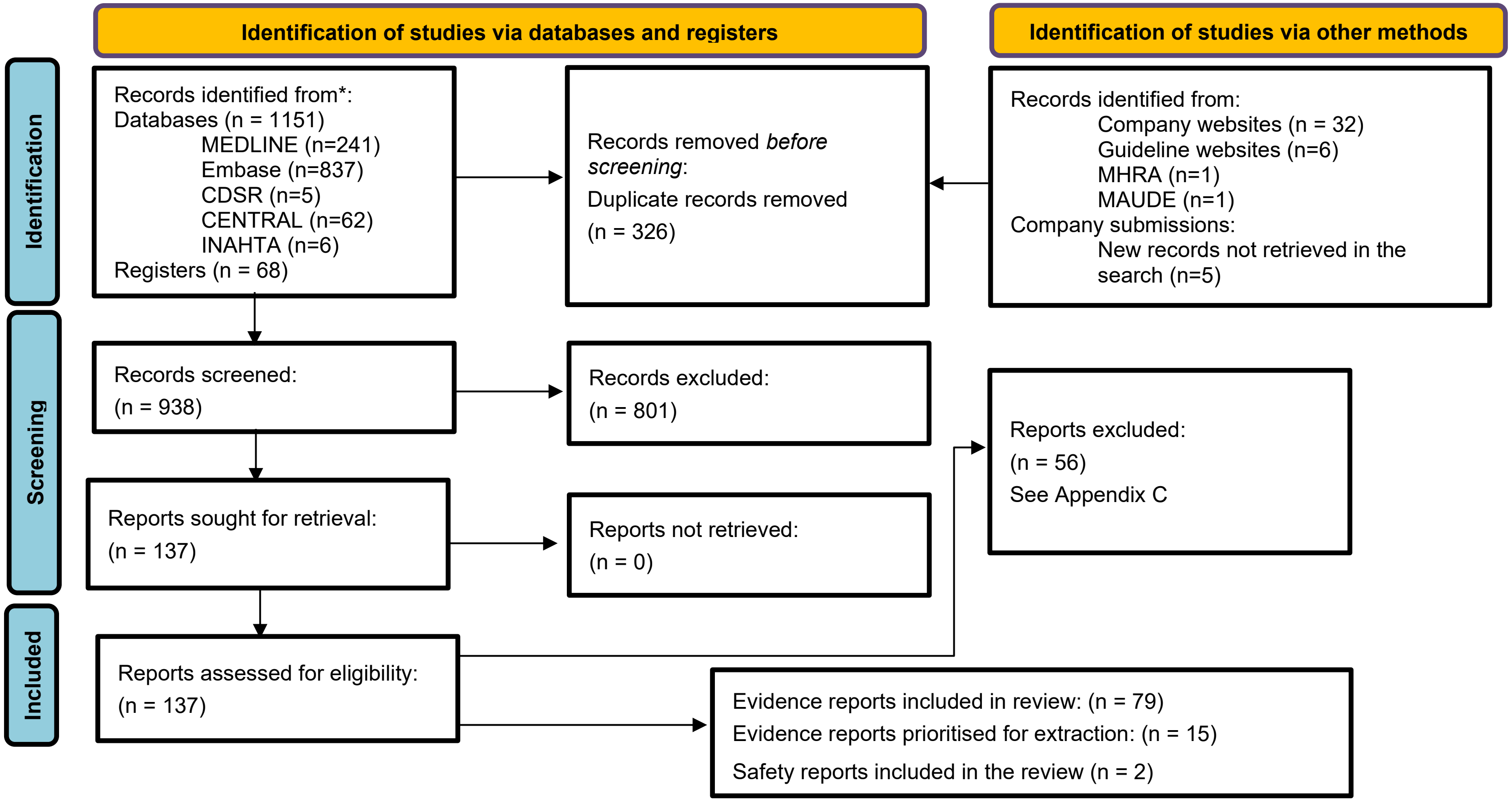
= **6 guidelines**

SIGN

contour or contouring or autocontour or autocontouring or artificial intelligence or machine learning (as separate searches)

= **0 guidelines**

Appendix B: PRISMA flow diagram


Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from*:
Databases (n = 1151)
MEDLINE (n=241)
Embase (n=837)
CDSR (n=5)
CENTRAL (n=62)
INAHTA (n=6)
Registers (n = 68)
Records removed *before screening*:
Duplicate records removed
(n = 326)
Records identified from:
Company websites (n = 32)
Guideline websites (n=6)
MHRA (n=1)
MAUDE (n=1)
Company submissions:
New records not retrieved in the search (n=5)
Records screened:
(n = 938)
Records excluded:
(n = 801)
Reports excluded:
(n = 56)
See Appendix C
Reports sought for retrieval:
(n = 137)
Reports not retrieved:
(n = 0)
Reports assessed for eligibility:
(n = 137)
Evidence reports included in review: (n = 79)
Evidence reports prioritised for extraction: (n = 15)
Safety reports included in the review (n = 2)

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

Appendix C: List of excluded studies

Author	Reason for exclusion
Ahn, 2019[97]	Wrong intervention
Alves, 2021[98]	Wrong study design
Aoyama, 2021[99]	Wrong intervention
Azria, 2022[100]	Wrong intervention
Bakx, 2021[101]	Wrong intervention
Byun, 2021[102]	Wrong intervention
Canters, 2021[103]	Wrong intervention
Chan, 2018[104]	Wrong intervention
Chen, 2021[105]	Wrong intervention
Choi, 2020[106]	Wrong intervention
Chuter, 2018[107]	Wrong intervention
Crouzen, 2021[108]	Wrong intervention
Ewinckele, 2020[109]	Background article
Feng, 2019[110]	Wrong intervention
Feng, 2020[111]	Wrong intervention
Feng, 2020[112]	Wrong intervention
Feng, 2021[113]	Wrong publication type
Fi, 2019[114]	Wrong intervention
Fong, 2021[115]	Wrong intervention
Gan, 2021[116]	Wrong intervention
Garcia-Perez, 2018[117]	Wrong intervention
Ghimire, 2021[118]	Wrong intervention
Ghimire, 2022[119]	Wrong intervention
Giaddui, 2016[120]	Wrong intervention
Giaddui, 2016[121]	Wrong intervention
Gleeson (2023)[122]	Wrong intervention
Gooding, 2022[123]	Background article
Gungor, 2022[124]	Wrong intervention
Hammers 2020[125]	Wrong intervention
He, 2022[126]	Wrong intervention
Hedrick, 2022[127]	Wrong study design
Hern, 2022[128]	Wrong intervention

Johansson, 2022[129]	Wrong intervention
Khalifa, 2022[130]	Wrong study design
Loap, 2023[131]	Wrong intervention
Lu, 2021[132]	Wrong intervention
Magallon-Baro, 2022[133]	Wrong intervention
Nash, 2022[134]	Wrong intervention
Newman, 2022[135]	Wrong intervention
Nicolae, 2020[136]	Wrong intervention
O'Hara, 2022[137]	Wrong intervention
Okada, 2022[138]	Wrong intervention
Shelley, 2023[139]	Wrong intervention
Sibolt, 2021[140]	Wrong study design
Singh, 2019[96]	Wrong study design
Tsui 2021[141]	Wrong intervention
van de Sande, 2021[142]	Wrong intervention
Visak, 2023[143]	Wrong intervention
Vrtovec, 2020[144]	Background article
Wang, 2019[145]	Wrong intervention
Wang, 2019[146]	Wrong intervention
Yang, 2020[147]	Wrong intervention
Yedekci, 2022[148]	Wrong intervention
Zabel, 2020[90]	Wrong study design
Yedekci 2023[149]	Wrong intervention
Zhong, 2021[150]	Wrong intervention

Appendix D. Additional study results

This table presents results for clinical effectiveness outcomes. Further details compared to the results presented in the main clinical section are provided where relevant. However, there has been a focus on making the results understandable rather than presenting all minutiae.

Table 11: Study results for clinical effectiveness

References	Structures	Results	Authors conclusions
AI-Rad Companion Organs RT			
Ginn 2023[6]	Head & neck Pelvis	**Time-saving metrics**: Editing contours was faster than manual contouring with an average time saving of 43.4% or 11.8 minutes per patient. **Qualitative assessment**: 240 structures were scored, with > 95% of structures receiving a score of 3 (only minor edits needed) or 4 (clinically usable). Of the structures reviewed, only 11 structures needed major revision or to be redone entirely. Structures that needed more revision included the prostate, the oesophagus and the optic nerves. **Geometric analysis**: Dice and Jaccard scores showed high alignment (> 0.8) for relatively large organs such as the lungs, brain, liver, femurs, and kidneys. Smaller elongated structures had lower scores. Poor performing outlier cases included overestimation of the bladder and incorrect truncation of the spinal cord and femur contours.	Our results indicate the evaluated auto-contouring solution has the potential to reduce clinical contouring time. The algorithm's performance is promising, but human review and some editing is required prior to clinical use.
ART-Plan			
Blanchard 2020[12] (abstract)	Head & neck	**Qualitative assessment**: v1.0 (trained on 6,000 cases): 96% of all manual contours were classified as clinically useable (75% as A [acceptable] and 21% as B [acceptable after minor corrections] categories), compared to 95% for auto-contours (56 % and 39 % in A and B, respectively).	This study illustrates the potential of AI for automatic contouring of organs at risk in radiotherapy planning. Automatic contouring with this CE-marked software was very

References	Structures	Results	Authors conclusions
		v2.0 (trained on 21,000 cases): contours classified as clinically useable (A + B) increased significantly, reaching 100% for mandibles, 98% for brain stem, 98% for optic nerve and 92% for submandibular gland, versus 100%, 97%, 63% and 50% for v1.0, respectively. **Geometric analysis**: For optic nerve and submandibular gland, mean DICE score improved from 0.53 to 0.70 and 0.70 to 0.78 between v1.0 and v2.0 of the software, whereas mean HD score decreased by 30% and 17%, respectively.	close to expert contouring and clinically usable in the vast majority of cases.
AutoContour			
Leyva 2022[18] (abstract)	Head & neck	**Geometric analysis**: Good agreement was found between the AI generated contours and manually drawn clinical contours. Nine out of ten had mean surface distance less than 5mm, while DICE scores of greater than 0.7 were found for 60% of the sample included. However, a larger variance in DICE scores was seen for structures with small volume (< 5cc) such as pituitary, chiasm and cochlea, as well as for structures that were manually drawn solely in the area of interest, such as the spinal cord and oesophagus.	AutoContour tool generates clinically acceptable normal structure contours and is efficient in removing inter-user segmentation variability that occurs with manual segmentation but more qualitative research is needed.
DLCExpert			
Hague 2020[21]	Head & neck	**Qualitative assessment**: Scores were from 1 (good agreement') to 7 (gross error); a score of 5 or less was defined as clinically acceptable. Scores were assigned based on scans from three different MRI machines (diagnostic, planning and MR-linac). The mean score using diagnostic scans was 1.4, for planning scans it was 1.9, while for MR-linac it was 5.4. **Geometric analysis**: Automated contours showed good agreement with manual contours on the diagnostic and planning scans for bilateral parotid glands and submandibular glands (with a mean DICE score of 0.8 or above). The agreement was lower for the MR-linac scan for the bilateral parotid glands (mean DICE of 0.70). There	MR -based deep learning auto-contouring models show promise as an aid to clinician OAR contouring. A model trained on diagnostic MR images has been shown to work well on planning images as well as diagnostic images. However, extending this MR-linac images shows that these models remain sensitive to the MR sequence used.

References	Structures	Results	Authors conclusions
		was a lack of overlap between automated and manual contours on the MR-linac images for the left and right submandibular glands (mean DICE of 0.10 and 0)	
Van Dijk 2020[22]	Head & neck	**Time-saving metrics**: Deep learning contouring reduced the delineation time compared to atlas contours. The average delineation time for the expert and beginner observer for the atlas contours were 36 ± 7 and 59 ± 14 minutes, respectively, reducing slightly to 34 ± 6 and 54 ± 8 minutes, respectively, for the deep-learning contours. **Qualitative assessment**: Deep learning contours were more often preferable to the atlas contours overall, were considered to be more precise, and were more often confused with manual contours (in the Turing test). The overall misclassification of manual contours was 41%, with 32% of them being marked as requiring correction. However, manual contours still outperformed both deep learning and atlas contours. Nevertheless, deep learning contour results were within or bordering the inter-observer variability for the manual edited contours. **Dosimetric analysis**: The mean dose differences (|∆mean-dose|) between the glandular manual and auto-contours were lower for deep learning contours (0.9 ± 1.3 Gy) than for atlas contours (1.9 ± 2.7 Gy). Likewise, the mean dose differences decreased significantly for all upper digestive tract and airway organs, except for the right buccal mucosa. For the CNS, mean dose differences were comparable between atlas and deep learning contours, while deep learning contouring reduced the mean dose distance in the carotid arteries. **Geometric analysis**: For glandular organs at risk, DICE values for deep learning contours significantly improved over atlas, with the largest difference for the thyroid gland, where DICE increased from 0.60 ± 0.15 (atlas) to 0.83 ± 0.08 (deep learning). Similarly, for the upper digestive tract and airway organs, DICE values were significantly higher for deep learning compared to atlas for all except for the oral cavity. For the CNS, DICE values were slightly higher for atlas than for deep learning contouring (DICE >0.86 and >0.84, respectively) for the brainstem, cerebrum, and	The deep learning contouring, trained on a large head and neck cancer patient cohort, outperformed atlas contouring for the majority of head and neck outcomes of interest.

References	Structures	Results	Authors conclusions
		spinal cord. Finally, DICE values were substantially higher for deep learning (0.68 ± 0.11) than for atlas (0.29 ± 0.12) for the carotid arteries.	
Vaassen 2021[32]	Thorax (lung cancer)	**Dosimetric analysis**: Dosimetric effect of intra-observer contour variability was highest for Heart D_{max}(3.4 ± 6.8 Gy) and lowest for Lungs D_{mean}(0.3 ± 0.4 Gy). The effect of contour variation on treatment plan evaluation was highest for Heart D_{max}(6.0 ± 13.4 Gy) and oesophagus D_{max}(8.7 ± 17.2 Gy). However, dose differences for the various treatment plans, evaluated against the manual contour, were on average below 1 Gy/1%, and the majority of treatment plans fulfilled the planning objectives. Some patients were assigned doses by auto-contouring above the clinical constraint: this happened for the heart (x3), the lungs (x6) and the spinal cord (x1). For the heart, the clinical constraint was exceeded for patients where the target was located close to, or overlapped with, the heart. For Lungs, this was due to large tumour size. For the spinal cord, the tumour was located next to the spinal cord. **Geometric analysis**: The highest DICE scores were for lungs (1.0), while the lowest DICE score was for the oesophagus (0.46) (although the highest HD score—and therefore poorest alignment according to that metric—was for the heart).	This study shows the potential for procedures to use automatic delineation and planning in the thorax in clinical practice. Dose differences arising from automatic contour variations were of the same magnitude or lower than intra-observer contour variability.
INTContour			
Duan 2022[42]	Pelvis (prostate cancer)	**Qualitative assessment**: In the double-blinded evaluation, 95.7% of the AI contours were scored as either "perfect" (34.8%) or "acceptable" (60.9%), while only one (prostate) case (4.3%) was scored as "unacceptable with minor changes required." The reference contour was picked as the better contour in seven of 23 cases, and the AI contour was picked as the better contour in three cases. The remaining 13 cases were "too close to call". **Dosimetric analysis**: AI contours produced a statistically significant difference in bladder dose (a lower dose). No statistically significant differences were found in other organs at risk. All organs at risk satisfied the dose constraints of RTOG-0815 (contouring guidelines)	Automated treatment plans created from the AI contours produced similar and clinically acceptable dosimetric distributions as those from plans created from reference contours.

References	Structures	Results	Authors conclusions
		Geometric analysis: The AI contours demonstrated good accuracy on organs at risk and prostate contours, with average DICE scores for bladder, rectum, femoral heads, seminal vesicles, and penile bulb of 0.93, 0.85, 0.96, 0.72, and 0.53, respectively. The DICE, HD, and mean surface distance for the prostate were 0.83 ± 0.05, 6.07 ± 1.87 mm, and 2.07 ± 0.73 mm, respectively.	
Limbus Contour			
Radici 2022[44]	Head & neck Pelvis (prostate cancer) Bowel (rectal cancer) Thorax (breast cancer)	**Time-saving metrics**: The maximum time saving, both absolute and relative, was in head and neck contours (80 min, -65%). Time savings were also seen for breast (7 min, -46%); prostate (4 min, -18%), and rectum contours (3 min, -17%). **Dosimetric analysis**: The most relevant difference between auto and manual contours was found in the bowel for rectal cancer treatments: the mean volume covered by the 45 Gy isodose was 10.4 cm^3 for the manually contoured structures versus 289.4 cm^3 for the auto-contoured ones. The reason for this large difference is discussed later in the paper, where the authors noted that differences in the definition of the bowel between automatic and manual contours justified the dosimetric variation observed. Otherwise, dose distributions were similar between auto and manual contours. **Geometric analysis**: The lowest DICE score was for the penile bulb (0.39), while the best results were found for lungs (0.99). Good agreements were found for bladder, heart, and femoral heads, all with values greater than or close to 0.9. Considering all structures, the average DISC score was 0.72	Automatic contouring was able to significantly reduce the time required in the procedure, simplifying the workflow, and reducing interobserver variability. Its implementation was able to improve the radiation therapy workflow in our department.
Wong 2021[45]	Central nervous system Head & neck	**Qualitative assessment**: “Satisfaction scores” could be assigned from 1 (poor) to 5 (high). The mean score for CNS was 4.1, for head and neck it was 4.4, and for prostate it was 4.6. “Editing scores”, on the other hand, ranged from 1 (minimal editing required) to 5 (significant editing required). Most deep learning contours required minimal edits (mean editing score ≤2). The highest editing scores were for optic chiasm (3.4), prostate (2.8) and mandible (2.3).	Previously validated deep learning contouring models for CNS, head and neck, and prostate radiotherapy planning required minimal subjective and objective edits. High user satisfaction suggests that the auto

References	Structures	Results	Authors conclusions
	Pelvis (prostate cancer)	**Geometric analysis**: Unedited DCs were compared to the edited treatment approved contours. Mean DICE HD scores were ≥ 0.90 and ≤ 2.0mm, respectively. The poorest scores were for the optic chiasm.	contours may have served as appropriate starting points.
Wong 2020[46]	Central nervous system Head & neck Pelvis (prostate cancer)	**Time-saving metrics**: The mean auto and manual contouring times were, respectively: 0.4 vs 7.7 min for CNS; 0.6 vs 26.6 min for head and neck; 0.4 vs 21.3 min for prostate. **Geometric analysis**: Geometric analyses focused on inter-observer variation, to determine if the variation between deep learning and manual contours is comparable to variation among manual contours by ROs. For CNS structures, geometric metrics were not significantly different for the optic chiasm. However, variation was greater between deep learning and manual contours vs among manual contours alone for the optic globe (DICE score 0.85 vs 0.87, respectively). For head and neck, geometric metrics were not significantly different for spinal cord, parotid gland, submandibular gland. Variation was greater, however, between deep learning and manual contours vs among manual contours alone for the neck clinical target volume (DICE score 0.72 vs 0.79). For prostate, geometric metrics were not significantly different for seminal vesicles and rectum. However, variation was greater between deep learning and manual contours vs among manual contours alone for bladder (DICE score 0.97 vs 0.96), femoral head (0.91 vs 0.91) and prostate (0.79 vs 0.83).	We found that deep learning contours take significantly less time to generate than manual contours and approximate the expert Inter-observer-variability seen for organs at risk. Deep learning contours for clinical target volumes were less accurate, but given that these volumes highly depend on the clinical scenario and clinical judgement of the treating physician, they would likely serve as a usable starting template for patient specific adjustments.
MIM Contour ProtégéAI			
Urago 2021[58]	Head & neck Pelvis (prostate cancer)	**Time-saving metrics**: The processing time to create delineations was approximately 3 min per case on the atlas-based model and approximately 5 min (range, 3–10 min) per case on the AI-based model for the patients with prostate cancer. For patients with head and neck cancer, the processing times were both approximately 6 min (range, 3–8 min).	The effectiveness of the commercial AI-based model can be expected to improve the segmentation efficiency and to significantly shorten the delineation time.

References	Structures	Results	Authors conclusions
		Qualitative assessment: For prostate cancer patients, some errors were observed in the atlas-based delineations when the boundary between the small bowel or the seminal vesicle and the bladder was unclear. The AI-based delineations, on the other hand, were more consistent with the manual ones. For patients with head and neck cancer, no significant differences were observed between the two models for almost all organs at risk, except small delineations such as the optic chiasm and optic nerve (for both atlas and AI-based delineations). **Geometric analysis**: In both the atlas- and AI-based models, the median DICE score exceeded 0.8, indicating good agreement with the manual delineations. In the AI contours, the median value was closer to 1, and the interquartile range was smaller than that of the atlas contours. Similarly, for HD, the median and interquartile range of AI-based assessment was smaller than that of the atlas-based assessment in both the bladder and the rectum. Mean distance to agreement results were similar to those of HD.	
MRCAT Prostate plus Auto-contouring			
Kuisma 2020[66]	Pelvis (prostate cancer)	**Geometric analysis**: DICE scores (mean, SD) showed high alignment for delineating prostate were 0.84, for bladder they were 0.92, and for rectum 0.86. DICE scores were lower (showing moderate alignment) for seminal vesicles (0.56) and penile bulb (0.69). In repeat assessment, using a second scan taken a median of 8 days after the first scan, consistency of prostate delineation resulted in a mean DICE score of 0.89 for auto-contours, while mean DICE scores for manual delineation was 0.82.	Fully automated MRI segmentation tool showed good agreement and repeatability compared with manual segmentation and was found clinically robust in patients with PC. However, manual review and adjustment of some structures in individual cases remain important in clinical use
MVision Segmentation Service			
Strolin 2023[68]	Head & neck	**Time-saving metrics**: The median (range) time for manual delineation, deep learning-based segmentation, and subsequent manual corrections were 25.0 (8.0-115.0), 2.3 (1.2-8) and 10.0 minutes (0.3-46.3), respectively. The overall time for volume of	Our analysis revealed the positive impact of introducing and validating a novel CE-

References	Structures	Results	Authors conclusions
	Thorax Abdomen Male and female pelvis	interest retrieving and modification was statistically significantly lower than for manual contouring (p<0.001) **Qualitative assessment**: The deep learning tool was generally appreciated by ROs, with 44% of vote 4 (well done) and 43% of vote 5 (very well done), correlated with the saved time (p<0.001). The average satisfactory grade per district was higher than 4, except for the female pelvis. **Geometric analysis**: Overall, DICE scores increased with the volume of the investigated volume of interest. Median DICE scores, when comparing manually adjusted auto-contours vs unedited auto-contours were higher than 0.8 for all the organs except for the oesophagus and glottis in the head and neck district. The relative volume differences and similarity indexes suggested a better inter-agreement of manually adjusted auto-contours than manually segmented ones.	and FDA- approved commercial deep learning tool for automatic segmentation in terms of; i) a high level of clinicians' satisfaction, particularly for complex cases including large and numerous organs, ii) saving time, and iii) improving the consistencies of volumes of interest amongst different ROs.
OSAIRIS			
Oktay 2020[72]	Pelvis (male) Head & neck	**Time-saving metrics**: Manual segmentation of nine organs at risk took 86.75 min/scan for expert reader and 73.25 min/scan for radiation oncologist. With AI to assist them in reviewing and editing it took 4.98 (95% CI, 4.44-5.52) min/scan for head and neck scans and 3.40 (95% CI, 1.60-5.20) min/scan for prostate scans. The autogenerated contours represented a 93% reduction in time. **Geometric analysis**: The auto-contouring models achieved levels of clinical accuracy within the bounds of expert interobserver variability for 13 of 15 structures (the left and right submandibular glands were the only two structures outside the bounds). The models also performed consistently well according to DICE scores and similar metrics. The lowest DICE score for the head and neck structures was 0.79, for the right submandibular gland, while the highest was 0.96, for the mandible. For pelvic structures the lowest DICE score was 0.73, for the seminal vesicles, and the highest was 0.982, for the left femur.	The models achieved levels of clinical accuracy within expert inter-observer variability while reducing manual contouring time and performing consistently well across previously unseen heterogeneous data sets. With the availability of open-source libraries and reliable performance, this creates significant opportunities for the transformation of radiation treatment planning.
RayStation			

References	Structures	Results	Authors conclusions
Almberg 2022[73]	Thorax (left-sided breast cancer)	**Qualitative assessment**: No or only minor corrections were required for 14% and 71% of the CTVs and 72% and 26% of the OARs, respectively. Major corrections were required for 15% of the CTVs and 2% of the OARs. None of the structures, neither target volumes nor OARs, were scored as ‘‘not usable”. The lungs and sternum did not need any corrections, while the most frequent corrections occurred in the cranial and caudal parts of the structures. **Dosimetric analysis**: VMAT-plans were automatically optimised based on the auto-contours using an in-house developed script in RayStation. Dose coverage (D98; lowest dose to 98% of the volume) for auto-contours was compared to those from the manual reference contours. While some statistically significant differences were found, these differences were considered clinically irrelevant. **Geometric analysis**: The metrics (DICE and HD) for auto-contours were better than manual inter-observer variation for all target volumes, reaching statistical significance for all except the breast. The trend was the same for the OARs: one or both metrics were significantly better for auto-contours than for manual inter-observer variation; the only exception was the thyroid gland.	The model is now clinically implemented at both hospitals and will be combined with other existing models and soon be available world-wide. Alongside implementation, quality assurance and monitoring should be performed, to gain further understanding of both the capabilities and limitations of the model.

Appendix E. SCM responses to cost-effectiveness question

Table 12 SCM responses to the EAGs cost effectiveness questions

Question posed to SCM	SCM 1 response	SCM 2 response	SCM 3 response	SCM 4 response	SCM 5 response	SCM 6 response	SCM 7 response	SCM 8 response
In your experience, does the use of AI auto contouring result in a reduction in clinician time compared to the standard approach used?	Yes (mostly). Apart from structures where a different definition/guidelines may be used Or for (small number) of patients who are known to have unusual anatomy / post-surgical changes in anatomy, which current AI segmentation cannot manage well.	Yes	Yes	If used correctly with the right software, then yes	At the moment its hard to tell, although I expect the degree of time saving will improve. Our experience was that currently there was no benefit.	In general no. OARs are often not contoured by clinicians, but rather by technical staff (radiographers, physicists, dosimetrists) and quickly checked for gross error by clinicians. Introducing AI contouring for OARs is therefore not reducing clinician time in our experience	Yes, very clearly. Both expected time savings for OARs that were previously outlined manually. Also, some unexpected benefits e.g. lung tumour CTV definition, still done manually but time is reduced by being able to exclude local normal structures (outlined with AI) from the CTV using TPS Boolean tools.	Yes; there will be some reduction in clinician time. It could potentially speed up RT pathways and hence enable quicker access to treatment as well as clinician's ability to do more cases per unit of time.
From your perspective, what is the primary	Efficiency and harmonisation gains.	Reduction in delineation time	It saves time in contouring routine organs at risk , good	Frees staff for more complex cases where AI not appropriate	the primary benefit is time saving.	Primary benefit is greater standardisation of contours	Time saving. Hopefully improved consistency too	Reduction in the time required for contouring OARs – may

benefit of using AI- auto contouring, compared to current standard practice	Allows all staff to have a harmonised vocabulary and knowledge to discuss contours. Allows us to realistically design audit or research questions using large numbers of patients, with full contouring (which would be un-feasible if this was all manually contoured).		standardised approach in contouring and labelling as compared to manual process.	and increasing clinical load Standardises contours among observers				facilitate shortening of RT pathways and better access for patients. There may be some harmonisation of practices and reductions in inter-observer variability
Do you foresee any improvement in patient disease progression or survival outcomes from these technologies?	This would be very difficult to demonstrate in a scientifically and statistically rigorous way, and potentially not very ethical to test this prospectively in a full randomised controlled trial. It is **possible** that efficiency gains could	Yes. Accurate Organ at Risk delineation will lead to a reduction in treatment toxicity. AI delineation of target structures may lead to improved clinical outcomes.	Yes, if there is a complete solution from contouring to planning technique.	Could be if shortening treatment pathways, but this is less relevant for prostate cancer. In theory in centres where AI contours are more clinically appropriate than the local clinician contours. Might also help if more accurate	If centres follow current guidelines and peer review all contours (whether human or machine generated), there should be no difference in contours and plan dosimetry. If AI contours are not properly reviewed, it could lead to	Not currently. More likely to reduce toxicity	Not where it is being used to simply replicate manual contouring. Perhaps where it allows structures to be outlined where they weren't before or for techniques to be changed e.g., breast Radiotherapy changing from a VSIM approach	No. OAR delineation would not be anticipated to have an impact on disease progression. Some tools offer prophylactic nodes delineation but I would not anticipate this to have an impact

	shorten patient pathways, and that could influence outcomes in a positive way. Also, any reduction in unwarranted variations (e.g. on contour definition or accuracy) would also have a positive impact on patient outcomes. It is possible that use of AI segmentation on retrospective patient cohorts (where follow up data on disease progression or survival was available) could facilitate audit and research. This sort of retrospective audit/registry trial could be a good way of			bowel etc contouring for toxicity rates.	worse outcomes as it could result in suboptimal treatments. Conversely, if we ever reach a point where AI contours result in significant time saving, this could		to target volume based	

	demonstrating benefits/risks of high quality contouring on patient outcomes.							
From your perspective what would you say were the major impacts on resources (either savings or increases) from using AI auto-contouring? (Please consider both in terms of staff time and any equipment required, or other impacts on resource use)	Has allowed more skill mix in tasks from v experienced staff to other staff groups / less experienced staff members whilst maintaining consistent quality. Has allowed de-bundling of tasks and steps in patient pathway. This allows a much more consistent time for the contouring step in the patient pathway with AI, rather than with manual human expert delineation, reduces delays or waiting time in	Costly to purchase, commission time, lack of formal QA standards. Saving of delineator time (radiographer, dosimetrist or clinician)	Training, Trouble shooting, financial cost, Data protection	Time a massive impact with increasing patient numbers and more time pressure on staff. Especially important given clinical oncology recruitment issues and unfilled training/ consultant posts. Equipment less of an issue at our centre as contouring done remotely now, rather than in a 'planning room.	There will be a significant cost in commissioning and implementing these systems if models need training on local data, or retraining at later dates due to changes eg in protocol, imaging modalities/parameters etc. hardware or cloud computing physics/computing staff to maintain the systems, perform updates etc There could be significant savings. in staff time spent contouring	Some time savings from AI auto contours, but offset by time for manual inspection and minor modifications. As commercial AI auto-contouring models are currently classed as decision aids the responsibility for accuracy lies with the end-user and this limits the time savings available to clinical staff. Additional resources are required to implement AI models initially and for any updates (that are often quite	Reduction in staff time for contouring is the obvious main benefit. The financial cost of the software needs to be considered. Staff time will be required to develop processes for the safe implementation of AI auto-contouring and manage software upgrades. Reduction in staff time spent contouring will greatly outweigh this though	Time saving for clinicians and streamlining of pathways

	the many consecutive steps of patient treatment pathway.				if the systems are good enough	regular as the models are not yet fully mature in most cases).		
On average, how much time do you spend editing AI auto contours?	Varies hugely depending on tumour site and individual structure. Also can be very patient dependant. HN lymph nodes – estimated around 10 mins average editing time	30mins	less than 15 minutes if it is only organ at risk	It would depend on how accurate the contour is, and the structure. IF accurate then a minute or so for prostate. Can be more for SV. Many contours are appropriate for clinical use but we are just so used to ‘tweaking’ them	We have decided not to use these systems currently as we did not feel they are currently worth the expense.	Difficult to say as once AI models are in routinely clinical use, we do not monitor time spent changing AI auto contours. We know from local data that <3% of AI auto contours are grossly incorrect, usually due to unusual patient set-up or anatomy.	It varies greatly for different treatment sites and different structures, and depends on how the AI is implemented, what imaging protocols are used	N/A

Appendix F. Description of cost offset calculator

The purpose of this appendix is to provide a user guide and summary implications of the Excel cost offset calculator for AI auto-contouring.

To use:

- Enter the estimated cost of the AI intervention per radiotherapy planning session. This should include any licence fee and any specific equipment required over and above that required for manual contouring. If the licence fee is on a per centre or per period of time basis rather than per plan then an estimate of the cost per plan must be made based on expected number of plans per period of time.
- Enter the time saved from auto-contouring vs manual. This should be the estimated time to prepare a manual plan less the estimated time to prepare an auto-contouring plan.
- Select the grade of staff who performs the contouring from the dropdown list of options.

Based on this information, an hourly cost for the staff grade is estimated based on 2021 unit costs (the latest available at the time of writing). Multiplying this by the time saved gives an estimate of the value of the time. The net cost is simply the cost of the AI system less the cost of time saved.

For example, suppose the AI system cost £8 per plan and reduced the time taken to prepare a plan by 30 minutes and a registrar usually performed the contouring. At the time of writing, the hourly cost of a registrar is £52, so the value of time saved is £26 (£8 - £26 = -£18). Therefore, the cost saving per plan is £18.

Under this scenario, the AI system can cost up to the £26 value of time saved for it to be cost-neutral to the NHS.

Likewise, If a band 7 radiographer were to perform the contouring (£65ph) and the per-plan cost of AI was only £4, it would have to save only 4/65 = 0.062 of an hour = ~4 minutes for it to be cost neutral. If the per plan cost was £50, then it has to save at least 50/65 = 0.769 of an hour = ~47 minutes for it to be cost-neutral.