

# **Künstliche Intelligenz in der Risikoorgan- und Zielvolumenkonturierung im Rahmen der Bestrahlungsplanung**

Inaugural-Dissertation  
zur Erlangung des Doktorgrades  
der Hohen Medizinischen Fakultät  
der Rheinischen Friedrich-Wilhelms-Universität  
Bonn

**Gustavo Renato Sarria Vargas**

aus Lima, Peru

2024

Angefertigt mit der Genehmigung  
der Medizinischen Fakultät der Universität Bonn

1. Gutachter: PD Dr. med. Christopher Schmeel
2. Gutachter: Prof. Dr. Björn-Peter Krüger

Tag der Mündlichen Prüfung: 23. April 2024

Aus der Klinik und Poliklinik für Strahlentherapie und Radioonkologie  
Direktorin: Univ.-Prof. Dr. med. Eleni Gkika

## **Widmung**

An Gott, der mich auf dieser Reise begleitet. Meinen Eltern und Großeltern für ihre selbstlose Unterstützung während all dieser Jahre. Meiner Frau für ihre bedingungslose Hilfe und die geopfert Zeit, die es mir ermöglicht hat, meine Karriere weiterzuentwickeln. Meinen Kollegen, die mich mit offenen Armen empfangen haben und dafür sorgten, dass ich mich von Anfang an wie zu Hause fühlte.



## Inhaltsverzeichnis

	<b>Abkürzungsverzeichnis</b>	6
<b>1.</b>	<b>Deutsche Zusammenfassung</b>	7
1.1	Einleitung	7
1.2	Material und Methoden	9
1.3	Ergebnisse	15
1.4	Diskussion	18
1.5	Zusammenfassung	23
1.6	Literaturverzeichnis der deutschen Zusammenfassung	25
<b>2.</b>	<b>Veröffentlichung</b>	29
	Abstract	29
	Introduction	29
	Materials and Methods	30
	Results	31
	Discussion	32
	References	35
<b>3.</b>	<b>Danksagung</b>	36
<b>4.</b>	<b>Lebenslauf</b>	37

## Abkürzungsverzeichnis

KI - AI	Kunstliche Intelligenz – Artificial intelligence
ESTRO	European Society for Radiation Oncology
ASTRO	American Society for Radiation Oncology
PK - PC	Prostatakarzinome – Prostate cancer
BK - BC	Brustkrebs – Breast cancer
KHT - HNC	Kopf- und Halstumoren – Head and neck cancer
DICOM	Digital Imaging and Communications in Medicine
MD	Arzt – Medical doctor
OAR	Risikoorgane – Organs at risk
CTV	Klinische Zielvolumina - Clinical target volume
K - C	Korrekturen - Corrections
DSC	Dice-Ähnlichkeitskoeffizient – Dice similarity coefficient
SD	Standardabweichung – Standard deviation
LAD	Left anterior descending coronary artery – ramus interventricularis anterior
$D_{\text{volumen}}$	Dosis (Gray) an einem bestimmten Volumen
$V_{\text{Dosis}}$	Belastung eines Volumens mit einer bestimmten Dosis
DVH	Dosis-Volumen-Histogramm

# 1. Deutsche Zusammenfassung

## 1.1 Einleitung

Künstliche Intelligenz (KI - AI) gewinnt in der medizinischen Versorgung immer mehr an Bedeutung und ist gleichsam auch mit Fortschritten im medizinischen Bereich verbunden. In dieser neuen Ära eröffnen sich zahlreiche Möglichkeiten zur Verbesserung der medizinischen Arbeitsabläufe und gleichsam zur Entlastung des Personals. Die Onkologie ist ein interdisziplinäres Fachgebiet, in dem sich die derzeitige Forschung auch intensiv mit dem Nutzen der KI befasst. Auch und gerade in dem hochtechnisierten Fach der Strahlentherapie sind die Einflüsse der KI, im Gegensatz zu anderen onkologischen Teilbereichen, schon seit Jahren zugegen (Vandewinckele et al. 2020).

So ist die Strahlentherapie in hohem Maße von technologischen Fortschritten abhängig. Ständige Fortschritte, sowohl software- als auch hardwareseitig, sind für Radioonkologen nichts Ungewöhnliches. Klassischerweise war die Sicherstellung präziser und sicherer Behandlungen sowie die Durchführung von Qualitätssicherungsanalysen eine Aufgabe, die nur von menschlichen Arbeitskräften übernommen werden konnte (Munbodh et al. 2022). In diesem Zusammenhang ist die Konturierung von Risikoorganen und Behandlungszielvolumina von großer Bedeutung in der Strahlentherapie, um diese Ziele zu erreichen, wobei verschiedene Gadgets zur Durchführung geeigneter Behandlungen eingesetzt werden können (Hernandez et al. 2020).

Die sogenannte Konturierung von Zielvolumina und Risikoorganen bezeichnet die Interpretation der Bilddaten sowie die folgende Festlegung des zu bestrahlenden Tumorgewebes und der vor Strahlung zu schonenden, gesunden Normalgewebe (Risikoorgane). Diese Aufgaben können durchaus zeitaufwändig und variabel sein, da verschiedene Strahlentherapeuten möglicherweise nicht übereinstimmende Ergebnisse liefern, je nach bestehendem Fachwissen, Erfahrung und Schwierigkeitsgrad des einzelnen Behandlungsfalles (Patrick et al. 2021). Der Prozess beinhaltet die Interpretation komplexer medizinischer Bilddaten und kritische Entscheidungen darüber, wie, wo und mit welcher Intention die Bestrahlung ansetzen soll. Die Zielvolumenkonturierung unterliegt, sofern klassischmanuell durchgeführt, einer substantziellen Variabilität zwischen verschiedenen Beobachtern, d. h. verschiedene

Kliniker können dasselbe Krankheitsbild unterschiedlich konturieren, was gegebenenfalls zu Unstimmigkeiten und potenziellen Fehlern führen kann (van der Veen et al. 2019, Patrick et al. 2021). In vorhergehenden Studien wurde diese Variabilität der Ergebnisse unter Strahlentherapeuten bereits untersucht – so wurden beispielsweise in einer klinischen Studie zu Thoraxmalignomen teils erhebliche Abweichungen bei der Festlegung der Behandlungsvolumina festgestellt, was unter Umständen zu Fehlbehandlungen führen könnte (Tsang et al. 2019). Dennoch hat die KI bislang von Menschen ausgeführte Aufgaben teilweise ersetzen und die Gesamteffizienz auf allen Workflow-Ebenen verbessern können (Vandewinckele et al. 2020).

Zu den modernen KI-Lösungen zählen Deep-Learning-Algorithmen wie z. B. Convolutional Neural Networks, die mit Hilfe großer Datensätze medizinischer Bilddaten trainiert werden. Zu ihren Merkmalen gehören beispielsweise das Erkennen von Mustern in medizinischen Bildern und das Erlernen der Unterscheidung verschiedener Strukturen wie auch die von bösartigem oder gesundem Gewebe (Wong et al. 2020). KI befindet sich immer noch in der fortwährenden Entwicklung und muss mitunter noch zahlreiche Unzulänglichkeiten überwinden, unabhängig von den bisher bereits erzielten Fortschritten. Nach wie vor sind zudem Studien zur Validierung von KI-Algorithmen erforderlich, um deren Zuverlässigkeit und Genauigkeit nachzuweisen.

Kürzlich durchgeführte klinische Untersuchungen haben die Position der KI innerhalb der Strahlentherapie und insbesondere bei der Konturierung von gesunden Normalgeweben gestärkt, indem sie dabei eine Leistungsverbesserung sowohl beim Zeitaufwand als auch bei der Variabilität zwischen den Konturierern zeigen konnten (Hosny et al. 2022). In diesem Sinne ist der Vergleich manueller, von Menschen vorgenommener Konturen, mit denen der KI derzeit die beste Methode zur Bewertung der Ergebnisse im Hinblick auf Reproduzierbarkeit und Genauigkeit (Savjani et al. 2022). Darüber hinaus besteht immer noch eine gewisse Zurückhaltung vor einem möglichen Ersatz menschlicher Leistungen durch KI: denn bis heute hat noch kein KI-System beweisen können, dass es trotz seiner Genauigkeit bei der Konturierung das Urteilsvermögen von Radioonkologen, Physikern oder Dosimetristen vollständig ersetzen kann. Insgesamt muss KI, zumindest gegenwärtig noch, als unterstützendes Werkzeug zur Verbesserung von Arbeitsabläufen und zur Arbeitserleichterung betrachtet werden und nicht als ein äquivalenter Ersatz für menschliches Fachwissen oder Verantwortung.

### a. Fragestellung

In der dieser Dissertation zu Grunde liegenden Publikation wurde auf die oben angeführten Punkte eingegangen und dabei die Hypothese aufgestellt, dass die KI-gestützte Autokonturierung das Potenzial innehaben könnte, Konturen zu erzeugen, die vergleichbar mit denjenigen, die manuell von einem erfahrenen Strahlentherapeuten erzeugt wurden, sind und dabei gleichzeitig signifikante Schwankungen in der Dosisverteilung vermeiden und erheblich Arbeitszeit einsparen kann. (Sarria et al. 2024)

## 1.2 Material und Methoden

### a. Konturierung von Zielvolumina und Risikoorganen

Das eigentliche Tumorumfang (engl. *Gross target volume* - GTV) stellt die grundlegendste Ebene der Zielvolumendefinition in der Strahlentherapie dar. Es handelt sich um eine dreidimensionale Darstellung des sichtbaren und palpablen Tumors, der durch verschiedene diagnostische Bildgebungsverfahren wie CT-, MRT- oder PET-Scans identifiziert wird. Die genaue Bestimmung des GTV ist von größter Bedeutung, da dieses als primäres Ziel für die Strahlentherapie dient. Radiologen, Onkologen und Medizinphysiker arbeiten bei der Definition des GTV eng mit den Radioonkologen zusammen und stellen sicher, dass die genauen Grenzen des Tumors bzw. der Tumorausdehnung gut nachvollziehbar berichtet und ggf. auch visualisiert werden.

Peripher des GTVs befindet sich das sogenannte klinische Zielvolumen (engl. *Clinical target volume* - CTV), welches in der Regel einen Sicherheitssaum um das GTV herum einschließt, um eine mögliche mikroskopische Tumorinvasion bei der Bestrahlungsplanung zu berücksichtigen. Dieser Sicherheitssaum berücksichtigt damit die aktuell nur diffizil mittels bildgebender Verfahren einzugrenzende Tumorausbreitung. Die CTV-Definition basiert auf klinisch-pathologischen Daten zur Tumorinvasion und Ausbreitung sowie klinischer Beurteilung und kann auf der Grundlage patientenspezifischer Faktoren und Behandlungsziele individualisiert an den Einzelfall angepasst werden.

Schließlich fügt das Planungszielvolumen (engl. *Planning target volume* - PTV) einen weiteren, zusätzlichen Saum um das CTV hinzu, um etwaigen Unsicherheiten im Strahlentherapieprozess Rechnung zu tragen. Diese Unsicherheiten können auf Faktoren wie Fehlern bei der Patientenlagerung, Organbewegungen und Schwankungen bei der Strahlenabgabe zurückzuführen sein. Durch die Anwendung eines solchen additiven Sicherheitssaumes um das CTV herum, stellt das PTV sicher, dass die Strahlendosis den vorgesehenen Behandlungsbereich auch bei Vorliegen solcher Unsicherheiten ausreichend abdeckt. Präzise Ausrichtungs- und Immobilisierungstechniken sowie moderne Technologien wie die bildgesteuerte Strahlentherapie spielen eine entscheidende Rolle bei der Minimierung dieser Unsicherheiten und der Optimierung der PTV-Definition.

In unserer Klinik wurde ein interner Wettbewerb unter fünf Strahlentherapeuten durchgeführt, um festzustellen, welcher Arzt die präzisesten und schnellsten Organkonturen auf Grundlage von Musterfällen gemäß ESTRO-, ASTRO- oder Global-Harmonization-Group- Leitlinien für die Konturierung vornehmen konnte (Brouwer et al. 2015, Offersen et al. 2016, Mir et al. 2020, Hall et al. 2021). Dieser Experte sollte anschließend alle manuellen Konturen der in dieser Studie eingeschlossenen Patienten übernehmen, welche als Grundlage für den nachfolgenden Vergleich dienen sollten.

Zuvor in unserer Klinik behandelte Fälle von Prostatakarzinomen (PK), Brustkrebs (BK) sowie Kopf- und Halstumoren (KHT) wurden dazu nach DICOM-Datensätzen durchsucht. Diese Fälle umfassten adjuvante Behandlungsfälle für Brustkrebs, primäre Prostatakarzinome und beide Behandlungsmodalitäten für Kopf- und Halstumoren und wurden unabhängig von anatomischen Varianten nach dem Zufallsprinzip ausgewählt, bis die gewünschte Zielanzahl erreicht war. Nur Fälle mit nicht-anatomischen Elementen bzw. Besonderheiten (bspw. Bolusmaterial, andere als orale Prothesen, Herzschrittmacher usw.) wurden ausgeschlossen. Nach entsprechender Selektion wurden diese anonymisiert und die DICOM-Sets für den geplanten Vergleich zwischen den Gruppen verdreifacht. Die Konturen des Arztes wurden auf dem ersten DICOM-Bilddatensatz in einer einheitlichen und kontrollierten Umgebung (ungestörte Umgebung; einheitliche technische Hard- und Softwareausstattung) durchgeführt, um ein ideales Arbeitsumfeld zu simulieren. Zu diesem Zweck wurde die Verwendung der softwareseitig vorhandenen Tools für die automatische Linienführung und Interpolation bei der Konturierung

zugelassen. Zu Vergleichszwecken wurde die Konturierungszeit gemessen und registriert. Zu den Basisstrukturen gehörten in allen Fällen die Risikoorgane und die jeweils für die spezifische Tumorentität passenden und zweckdienlichen CTVs in Prostatakarzinomen- und Brustkrebs-Segmenten, angelehnt an die uniformen Empfehlungen internationaler Leitlinien (Offersen et al. 2016, Mir et al. 2020, Hall et al. 2021). Die entsprechenden Strukturen sind in **Tab. 1** aufgeführt.

**Tab. 1: Struktur- und Dosisrichtwerte pro anatomischem Segment. \*Brustkonturen wurden sowohl als Risikoorgane als auch als CTV verwendet. (c): kontralateral, (b): bilateral.**

Segment	Strukturen	Dosisrichtwerte
Mamma	Lungen	D <sub>50%</sub> (b)
		D <sub>35%</sub> (b)
		D <sub>15%</sub> (b)
	Mammae*	D <sub>5%</sub> (c)
		D <sub>0,035ccm</sub> (c)
	Herz	D <sub>mean</sub>
		D <sub>10ccm</sub>
		D <sub>5ccm</sub>
	LAD	D <sub>mean</sub>
		V <sub>30Gy</sub>
	Rückenmark	D <sub>0.035ccm</sub>
D <sub>2ccm</sub>		
Schilddrüse	D <sub>mean</sub>	
Ösophagus	D <sub>mean</sub>	
Prostata	Dünndarm	V <sub>45Gy</sub>
	Blase	V <sub>70Gy</sub>
		V <sub>45Gy</sub>
	Analkanal	D <sub>mean</sub>
	Femora	V <sub>50%</sub> (b)
		D <sub>0.035ccm</sub> (b)
	Bulbus penis	V <sub>50%</sub>
	Rektum	V <sub>72%</sub>
		V <sub>70%</sub>
		V <sub>45%</sub>
	Cauda equina	D <sub>0.035ccm</sub>
Lymphabflusswege CTV	-	
Samenblasen	-	
Prostata	-	
Hals-Nasen-Ohren	Plexūs	D <sub>3ccm</sub> (b)
		D <sub>0.035ccm</sub> (b)
	Mundhöhle	D <sub>mean</sub>
	Ösophagus	D <sub>mean</sub>
V <sub>35Gy</sub>		

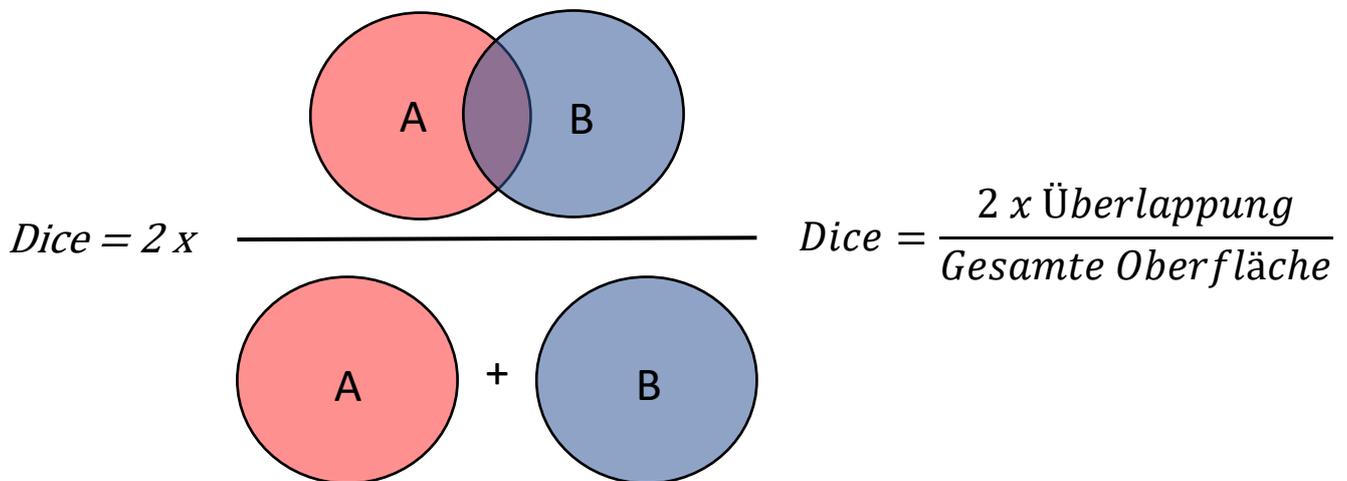
	Glandula submandibularis	$D_{\text{mean}} (b)$
	Schilddrüse	$D_{\text{mean}}$
	Larynx	$D_{\text{mean}}$
		$D_{0.035\text{ccm}}$
	Lippen	$D_{\text{mean}}$
		$D_{0.035\text{ccm}}$
	Constrictor pharyngis	$D_{\text{mean}}$
	Glandulae parotideae	$D_{\text{mean}} (b)$
		$D_{7\text{ccm}} (b)$
	Rückenmark	$D_{0.035\text{ccm}}$
Unterkiefer	-	

Nach der Erstellung der ärztlichen Standardkonturen wurden mit Limbus Contour v.1.6.0 (Limbus AI Inc., Regina, SK, Kanada) Deep-Learning-basierte Autokonturen für die verbleibenden zwei DICOM-Kopien erstellt, die einen reinen KI-Satz und einen KI-Satz mit zusätzlichen ärztlichen Korrekturen (KI+K) ergaben. Erstere blieben unverändert, während letztere vom Arzt zusätzlich auch manuell korrigiert wurden, um Strukturen zu erhalten, die den tatsächlichen klinischen Erfordernissen optimal entsprechen. **Tab. 2** führt einen Überblick über die Vergleichsgruppen und eine kurze Verfahrensbeschreibung für jedes anatomische Segment auf.

**Tab. 2: Vergleichsgruppen**

Vergleichsgruppen	Beschreibung
Arzt (MD)	Manuelle Konturen durch einen Arzt und gleichzeitig Goldstandard zu Vergleichszwecken
Künstliche Intelligenz (KI)	Automatische Konturen alleinig mit der Software generiert
Künstliche Intelligenz + Arzt-Korrekturen (KI+K)	Manuelle Korrekturen der mit der Software generierten Konturen durch einen Arzt

Die Unterschiede zwischen allen Konturierungsgruppen wurden mittels des Dice-Ähnlichkeitskoeffizienten (engl. *Dice* similarity coefficient - DSC) bewertet. **Abb. 1** stellt eine graphische Beschreibung des Dice-Ähnlichkeitskoeffizienten dar.



**Abb 1. Graphische Darstellung des Dice-Ähnlichkeitskoeffizienten**

Ein Wert von 0,7 oder weniger wurde als substanzieller, geometrischer Unterschied gewertet. Korrekturen von AI mit einem Unterschied von  $\geq 0,1$  wurden ebenfalls als substanzieller Unterschied gewertet. Alle Konturen wurden mit Eclipse v. 15.5 (Varian Medical Systems, Palo Alto, CA, USA) erstellt.

b. Vergleich von Planung und Dosisabgabe

In der Strahlentherapie wird gemeinhin die auf ein bestimmtes Ziel applizierte Dosis in Gray (Gy) gemessen und die Dosisverteilung nach ihrer Reichweite bzw. Belastung in einem Risikoorgan oder einem Behandlungsvolumen bewertet. Zu diesem Zweck werden verschiedene Parameter verwendet, die entweder als Dosis (D), die ein bestimmtes Volumen erreicht (z.B. 0,1 ccm, 5%, Durchschnitt, usw.) oder als Volumen (V), das einer bestimmten Dosis ausgesetzt ist (z.B. 30-, 50-, 70-Gy, usw.), beschrieben werden können.

In dieser Studie wurde die Dosisverteilung unverändert auf die ärztlichen Konturen übertragen, um die Abdeckung bzw. Übereinstimmung mit den manuellen Konturen beurteilen zu können. Danach wurden diese auch in unveränderter Form auf die KI- und KI+K-Konturen kopiert und die Differenzen der Risikoorganexposition mit Strahlung wurde mit der Ausgangssituation (dem derzeitigen Goldstandard = vollständig

manuelle ärztliche Konturierung) verglichen, um die Höhe etwaiger Abweichungen in der Organexposition zu ermitteln. Für jedes Segment (z.B. ein Risikoorgan) (**Tab. 1**) wurden mehrere Dosisbelastungsgrenzen (*engl. constraints*) gewählt, die in der Praxis als Dosisgrenzwerte Anwendung finden und sich an rezenten Studien oder Leitlinien orientieren.

c. Bewertung des Zeitaufwands

Die Zeiten für die Verarbeitung von Arzt-, KI- und KI+K-Konturen wurden zu deskriptiven Zwecken erfasst und der Zeitaufwand (Personenarbeitsstunden) wurde für den Vergleich zwischen den Gruppen berechnet.

d. Statistische Auswertung

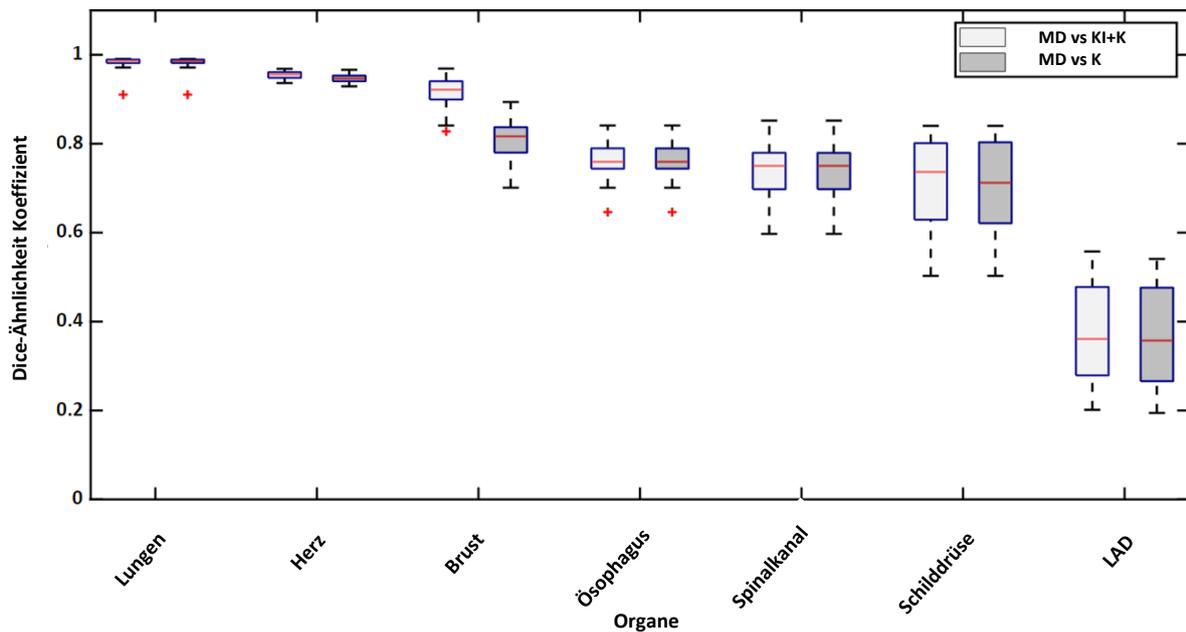
Die mittleren DSC-Werte plus Standardabweichung (*engl. Standard deviation - SD*) wurden für jedes untersuchte anatomische Segment in KI und KI+K erfasst und zu beschreibenden Zwecken mit den Arzt-Werten abgeglichen. Die Unterschiede in den sich ergebenden mittleren OAR-Dosen für jede Struktur wurden mit dem ungepaarten t-Test für unabhängige Variablen verglichen. Ein Wert von  $p < 0,05$  wurde als signifikanter Unterschied gewertet. Die Analyse wurde mit MATLAB R2020b (The MathWorks Inc. Natick, MA, USA) durchgeführt.

e. Ethische Erklärung

Aufgrund des retrospektiven Charakters dieser Studie wurde vor Initiation dieser Studie kein Votum der Ethikkommission eingeholt. Alle Patientendaten wurden anonymisiert. Die Untersuchung wurde nach den Grundsätzen der Deklaration von Helsinki durchgeführt.

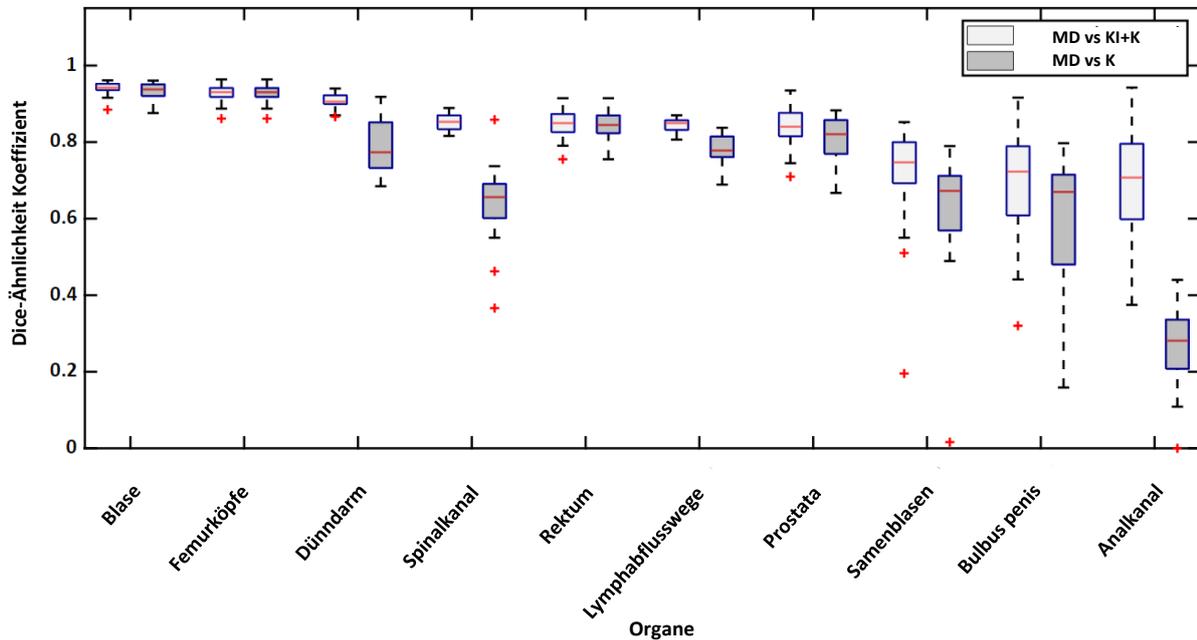
### 1.3 Ergebnisse

Für die Analyse wurden insgesamt 75 Fälle bearbeitet (jeweils 25 für jedes anatomische Segment). Für die Brustkrebs-Strukturen betrug der Gesamtmittelwert der DSC 0,82 (SD  $\pm$  0,19) für KI+K und 0,79 (SD  $\pm$  0,18) für KI. Alle Strukturen wiesen DSC-Werte über 0,7 auf, mit Ausnahme des Ramus interventricularis anterior [engl. *left anterior descending artery (LAD)*] mit 0,37 (SD  $\pm$  0,11) und 0,36 (SD  $\pm$  0,11) für KI+K bzw. KI (**Abb. 2**).



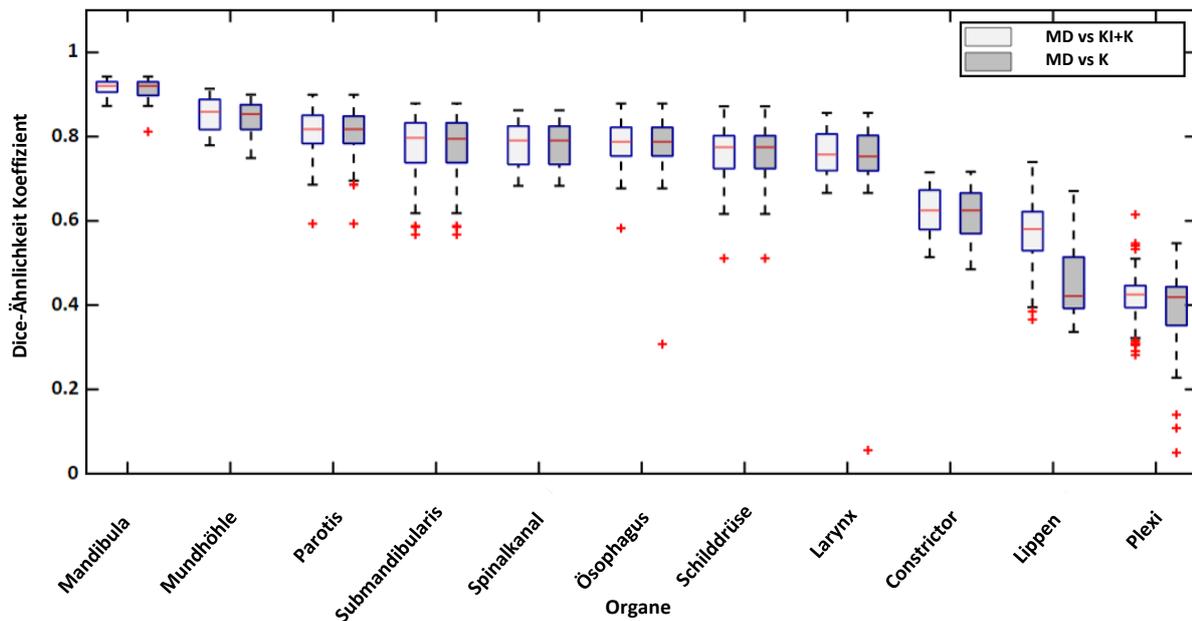
**Abb. 2:** Werte der Dice-Ähnlichkeitskoeffizienten bei Brustkrebs. Arzt-Standardkonturen (MD) wurden mit KI und KI plus Arzt-Korrekturen (KI+K) verglichen. Modifiziert übernommen aus Sarria et al. *Artificial Intelligence-Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics. Adv Radiat Oncol 2024; 9,3: 101394.*

Für Prostatakarzinome lagen die durchschnittlichen DSC bei 0,84 (SD  $\pm$  0,09) und 0,75 (SD  $\pm$  0,18). Größere Unterschiede wurden für den Analkanal (0,68 SD  $\pm$  0,14 und 0,27 SD  $\pm$  0,1) und den Penisbulbus (0,69 SD  $\pm$  0,13 und 0,60 SD  $\pm$  0,17) festgestellt. Die Cauda equina und die Samenblasen zeigten nur in der KI-Gruppe größere Unterschiede (0,64 SD  $\pm$  0,09 und 0,62 SD  $\pm$  0,15) (**Abb. 3**).



**Abb. 3: Dice-Ähnlichkeitskoeffizientenwerte bei Prostatakarzinomen. Arzt-Grundlinienkonturen wurden mit KI und KI plus Arzt-Korrekturen verglichen. Modifiziert übernommen aus Sarria et al. Artificial Intelligence-Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics. Adv Radiat Oncol 2024; 9,3: 101394.**

Bei den Kopf-Halstumoren lagen die durchschnittlichen DSC insgesamt bei 0,72 (SD  $\pm$  0,15) und 0,70 (SD  $\pm$  0,16). Größere Unterschiede wurden bei den Plexus brachialis rechts (0,42 SD  $\pm$  0,07 und 0,37 SD  $\pm$  0,11) und links (0,43 SD  $\pm$  0,06 und 0,41 SD  $\pm$  0,09), den Lippen (0,57 SD  $\pm$  0,09 und 0,45 SD  $\pm$  0,09) und dem Pharynx (0,62 SD  $\pm$  0,06 und 0,61 SD  $\pm$  0,07) festgestellt (**Abb. 4**).



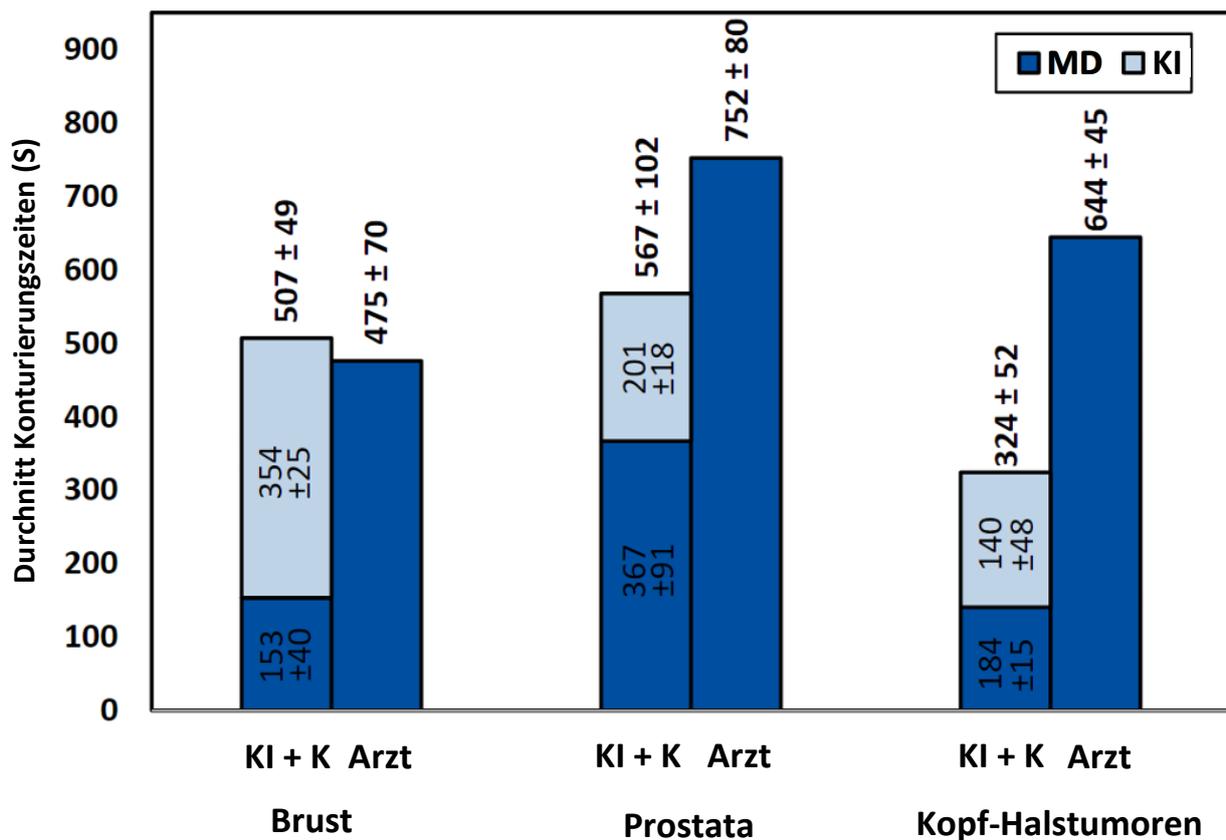
**Abb. 4: Dice-Ähnlichkeitskoeffizientenwerte bei Kopf-Hals-Tumoren. Arzt-Standardkonturen wurden mit KI allein und KI plus Arzt-Korrekturen verglichen. Modifiziert übernommen aus Sarria et al. Artificial Intelligence-Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics. Adv Radiat Oncol 2024; 9,3: 101394.**

Umfangreichere Korrekturen waren für den Analkanal, den Darm, die Cauda equina, die Brüste und die Lippen erforderlich, welche zusammen genommen 17,1 % aller in dieser Studie berücksichtigten Strukturen ausmachten.

In Bezug auf die Risikoorganexposition wurden signifikante Unterschiede in der Dosisverteilung für Brustkrebs-KI in der kontralateralen Brust  $D_{0,1\text{ccm}}$  festgestellt ( $p=0,041$ ). Für Prostatakarzinome- KI+K und KI im mittleren Darm  $V_{45\text{Gy}}$  ( $p=0,0007$  und  $p<0,0001$ ), Analkanal  $D_{\text{mean}}$  ( $p=0,0051$  und  $p=0,0013$ ) und Cauda equina  $D_{0,035\text{ccm}}$  (nur KI,  $p<0,0001$ ). Bei Kopf- und Halstumoren waren nur die Lippen  $D_{0,035\text{ccm}}$  ( $p=0,02$  und  $p=0,014$ ) signifikant unterschiedlich. Das bedeutet, dass bei 10,4 % aller Strukturen Abweichungen in der Dosisabsorption vorlagen.

Die Zeitaufwandsanalyse ergab durchschnittliche ärztliche Konturierungszeiten (Sekunden) von 475 (SD  $\pm$  70), 752 (SD  $\pm$  80) und 644 (SD  $\pm$  45) für Brustkrebs,

Prostatakarzinome und Kopf- und Halstumoren. Ebenso betragen die mittleren Korrekturzeiten für KI-generierte Konturen 153 (SD  $\pm$  40), 367 (SD  $\pm$  91) und 187 (SD  $\pm$  15), während die KI-Verarbeitungszeiten 354 (SD  $\pm$  25), 201 (SD  $\pm$  18) und 140 (SD  $\pm$  48) betragen. Dies entspricht einer durchschnittlichen relativen Reduktion des zur Konturierung erforderlichen Zeitaufwands von 68 %, 51 % bzw. 71 % für jedes der untersuchten anatomischen Segmente (**Abb. 5**).



**Abb. 5: Durchschnitt ( $\pm$ SD) Konturierungszeiten in Sekunden für Arzt-, KI- und KI+K-Konturen in allen drei untersuchten anatomischen Segmenten. Modifiziert übernommen aus Sarria et al. Artificial Intelligence-Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics. Adv Radiat Oncol 2024; 9,3: 101394.**

#### 1.4 Diskussion

Convolutional Neural Networks haben sich im Bereich der Automatisierung in der Medizin als sehr vielversprechend erwiesen. Gerade in der Analyse von Bildern und der

Erkennung von Mustern sind Sie bereits robust, was sie für die Aufgabe der automatischen Zielvolumenkonturierung optimal erscheinen lässt (Poortmans et al. 2020). Die Fortschritte der letzten Jahre haben zu einer Umstellung von der atlasbasierten Autokonturierung hin zu Deep-Learning-basierten Autokonturen geführt, die sich insgesamt besser den Anforderungen der täglichen Praxis anpassen und die zuvor genannten sogar übertreffen könnten, indem sie bisweilen auch Korrekturen an klinisch validierten Modellen vorschlagen (Rhee et al. 2019).

In der vorliegenden Studie wurde die Leistungsfähigkeit der KI-basierten Limbus-Konturierungssoftware mit der eines sehr erfahrenen Arztes verglichen und die Auswirkungen auf die klinische Routine bewertet. Die drei untersuchten anatomischen Regionen (Segmente) wurden als repräsentative Tumorentitäten ausgewählt, die global am häufigsten in der strahlentherapeutischen Routine anzutreffen sind.

Die DSC-Analyse zeigte dabei insgesamt akzeptable Korrelationen zwischen den Arzt-, KI- und KI+K-Konturen. Die meisten signifikanten Unterschiede wurden bei kleinen oder langen Organen festgestellt, wenn auch ohne größere klinische Bedeutung. Darüber hinaus ist zu berücksichtigen, dass alle Korrekturen (KI+K), bis zum Erreichen klinisch akzeptabler Standards, nach dem Urteil des Arztes vorgenommen wurden. Die niedrigen Korrekturraten der KI-generierten Konturen (17,1 % der Organe wiesen eine DSC-Änderung  $> 0,1$  auf) zeugen bereits von einer ausgesprochen hohen Genauigkeit, die überwiegend nur geringe oder gar keine Korrekturen erfordert. Darüber hinaus und unerwarteter Weise zeigte die dosimetrische Analyse zudem auf, dass sich im Rahmen der medizinphysikalischen Bestrahlungsplanung zwischen den rein KI-generierten Konturen und den zusätzlich ärztlich korrigierten Konturen kaum signifikante Änderungen der klinisch relevanten DVH-Parameter ergaben, diese also de facto keine Auswirkungen auf den endgültigen Bestrahlungsplan hatten. In Brustkrebs-Fällen wurden wesentliche Unterschiede bei den LAD-Konturen festgestellt. Dies ist relativ häufig der Fall, wie bereits in verschiedenen Studien zur Bewertung dieser Struktur festgestellt wurde und hängt in hohem Maße von der Qualität der Bildgebung und der individuellen anatomischen Gegebenheiten ab (Harms et al. 2021, Spoor et al. 2021, Chin et al. 2023). Dennoch stufen wir diese Unterschiede als klinisch untergeordnet und damit wenig relevant ein, da die meisten von ihnen auf Diskrepanzen bei der Breite, dem oberen oder unteren Ende

zurückzuführen waren. Weitere geringfügige Unterschiede wurden bei den Risikoorganen Speiseröhre, der Schilddrüse und dem Rückenmark festgestellt, die ebenfalls mit den zuvor genannten Faktoren zusammenhängen. Bemerkenswert ist, dass die KI+K-Brustkonturen mit großer Ähnlichkeit zur ärztlichen, manuellen Konturierung in CTVs umgewandelt wurden (DSC 0,92). Diese Ergebnisse stimmen mit anderen früheren Studien zu diesem Thema überein (Choi et al. 2020, Wong et al. 2020, Buelens et al. 2022). Es soll an dieser Stelle nicht unerwähnt bleiben, dass in diesem Segment aus technischen Gründen keine Konturen der Lymphbahnen berücksichtigt wurden. Die zum Zeitpunkt der Studie verfügbare Softwareversion rechnete mit den ASTRO -Konturen, während in der Studie die Konturen routinemäßig gemäß der hiesig bevorzugt verwendeten europäischen ESTRO-Empfehlungen durchgeführt werden. In Bezug auf die Prostatakarzinome wurden die deutlichsten Unterschiede im Bereich des Analkanals, den Lymphbahnen, dem Bulbus penis und der Samenblasen festgestellt. Im Falle des Analkanals zeigten sich erhebliche Abweichungen zwischen den Konturen von Arzt und KI. Ein Grund dafür könnte die von der Software verwendete „Schablone“ sein, die offensichtlich von den Konsensrichtlinien der Global Harmonization Group (Mir et al. 2020) abweicht und kleinere Strukturen erzeugt, die den äußeren Schließmuskel nicht erfassen. Ähnliche Differenzen zum Ideal wurden auch bei den Samenblasen beobachtet, was am Ehesten auf deren variable Anatomie zurückzuführen ist. Wahrscheinlich führten Unterschiede im Gesamtvolumen und in der kranialen oder kaudalen Länge zu niedrigeren DSC-Werten als bei den übrigen Risikoorganen, die meist jedoch nicht klinisch relevant sind (z. B. Cauda equina). Hervorzuheben ist, dass die Lymphbahnen mit einer geringfügigen Bearbeitung eine zufriedenstellende Konfiguration erreichten (DSC 0,85). Diese Ergebnisse ähneln denen von Cha et al., obwohl in unsere Analyse deutlich mehr Risikoorgane einbezogen wurden, was für zusätzliche Gewissheit und Aussagekraft sorgt (Cha et al. 2021). Für das Kopf- und Halstumoren-Segment wurde aufgrund der größeren Inhomogenität bei der Volumenvorgabe keine CTV-Analyse berücksichtigt. Hier wurden zudem die meisten Unterschiede in den DSC-Werte festgestellt, was möglicherweise auf die vergleichsweise nur geringen Organvolumina in der Kopf-Hals-Region zurückzuführen ist. Dennoch lagen die meisten von ihnen nur leicht unter einem DSC-Wert von 0,8 und über dem in dieser Studie definierten unteren Grenzwert von 0,7. In bestimmten Fällen, wie z. B. beim Plexus brachialis, waren die Abweichungen höher. Aus praktischen Gründen schließt die Plexuskontur die knöchernen

Strukturen zwischen den Foramina ein. Damit soll eine homogene Dosisverteilung entlang des Plexus gewährleistet und mögliche Hotspots in der Nähe minimiert werden. In diesem Sinne bietet der KI-Algorithmus eine anatomischere Abgrenzung, die die Wirbelkörper respektiert und auch den internationalen Guidelines entspricht (Mir et al. 2020). Darüber hinaus waren diejenigen Plexusanteile zwischen den beiden Skalenusmuskeln, die potenziell näher an den behandelten Volumina liegen, in allen drei Gruppen recht ähnlich. Trotz der DSC-Ergebnisse und aus klinischer Sicht sind diese Unterschiede jedoch nur als wenig relevant zu erachten. In Bezug auf die in dieser Studie angewendeten Dosisgrenzwerte muss hierzu auch festgehalten werden, dass der willkürlich gewählte Cut-off-Point des Dice-Koeffizienten in Anlehnung an frühere Veröffentlichungen (Dolz et al. 2016, Aoyama et al. 2021, Pera et al. 2023) festgelegt wurde, obwohl bis heute kein eindeutiger Konsens erzielt wurde. Ein Grund dafür dürfte sicherlich der Unterschied zwischen den bloßen Messwerten einerseits und der klinischen Anwendbarkeit andererseits sein, da erstere die letzteren nicht zwangsläufig garantieren (Sherer et al. 2021, Rhee et al. 2022). Folglich sind diese Werte primär als Referenz zu betrachten, sollten aber die klinische Entscheidungsfindung nicht unmittelbar beeinflussen. Weitere Metriken wurden für die Analyse nicht berücksichtigt, z. B. die sogenannte Hausdorff-Distanz, da lineare Metriken keinen angemessenen Überblick über die Volumenüberlappung bieten.

Stattdessen wurde versucht, die Unterschiede in der Dosisverteilung zwischen den drei Untersuchungsgruppen zu messen, um einen besseren Eindruck der potenziell klinischen Auswirkungen zu erhalten. Indem die tatsächlich *in praxi* bestrahlten Behandlungspläne mit den Arzt-Konturen fusioniert wurden und diese sodann als Standardkonturen angenommen wurden, kann abschließend festgestellt werden, dass nur fünf von insgesamt 48 Dosis-Volumen-Parametern, je nach Segmentierungsmethode, variierten. Diese Änderungen traten vor allem bei Organen mit großem Volumen oder irregulärer Anatomie auf. Da hauptsächlich von statistischen Unterschieden berichtet wird, können diese entweder niedrigere oder höhere Expositionsraten bedeuten. Unabhängig davon war die überwiegende Mehrheit der Unterschiede minimal und diese können auch als klinisch nicht relevant eingestuft werden, was die auf dem DSC-Koeffizienten basierenden Ergebnisse unterstützt.

Die Zeitaufwandsanalyse ergab interessante Unterschiede in allen drei Gruppen. So war die KI-Ausgabe bei den Prostatakarzinomen- und Kopf- und Halstumoren-Konturen offensichtlich schneller. Zwar war sie auch in den Brustkrebs-Sets schneller, doch war die Zeitersparnis insgesamt geringer. Dies dürfte mit der Anzahl der Strukturen pro Segment und deren Komplexität zusammenhängen. Zudem ist die Interpolation (automatisches Füllen von Lücken, die durch die Konturierung nur bestimmter Abschnitte und nicht des gesamten Organs oder der gesamten Struktur entstehen) und die integrierten Auto-Konturierungstools bei größeren Organen, wie z. B. der Lunge oder der Brust, in der Regel nur wenig fehleranfällig.

Wie erwartet waren die KI+K-Zeiten erheblich kürzer als die des Arztes, was zu einer wesentlichen Reduzierung der Arbeitsstunden führte. Es muss in diesem Kontext jedoch darauf hingewiesen werden, dass die Arzt-Konturen in einer kontrollierten und damit optimalen Umgebung für die konturierenden Ärzte durchgeführt wurden, welche real in der täglichen Praxis durch etwaige Störeinflüsse wie zwischenzeitige Unterbrechungen praktisch unmöglich zu reproduzieren ist. Überdies verfügt der in dieser Studie eingebundene Facharzt über große Erfahrungen auf dem Gebiet der Konturierung und ist seit vielen Jahren als Trainer für Konturierungen tätig. Selbstredend wurde dieses ideale Szenario gewählt, um die KI so stark wie möglich zu herauszufordern. In einer realistischeren Arbeitsumgebung könnte der Nutzen daher noch deutlicher ausfallen, wenn zusätzlich Störfaktoren berücksichtigt werden.

KI-Algorithmen weisen nach wie vor Limitationen und damit verbundene Unzulänglichkeiten auf. So können sie beispielsweise mit Bildmaterial von geringer Qualität, Artefakten, anatomischen Varianten oder unüblichen Positionierungseinstellungen Schwierigkeiten haben. Es besteht ferner ein gewisses Risiko, sich in der Routine zu sehr auf Autokonturierungsalgorithmen zu verlassen, was zu mangelnder Sorgfalt bei der gründlichen Bild- bzw. Konturenbeurteilung führen oder gar in Behandlungsfehlern resultieren könnte. Um diese Risiken zu mindern, ist es stets erforderlich, eine ärztliche, manuelle Überprüfung von Autokonturen in den Behandlungsplanungsprozess einzubeziehen. Daher wurde in unserer Klinik die Strategie KI+K als Standard-Workflow eingeführt. Eine weitere wichtige Überlegung ist das Risiko einer Dequalifizierung von Radioonkologen und Physikern, da eine manuelle

Konturierung von Risikoorganen und Zielvolumina mit dem zunehmenden Einsatz von KI-Lösungen weniger häufig notwendig wird. Diesem Risiko kann nur durch regelmäßige Schulungen und entsprechende Praxiseinsätze begegnet werden, um Autokonturen bei Bedarf überprüfen und anzupassen zu können. Nur so kann sichergestellt werden, dass die KI als unterstützendes, zeitsparendes Werkzeug zur Verbesserung und Erleichterung der klinischen Arbeitsabläufe eingesetzt wird und eben nicht als Ersatz für menschliches Fachwissen.

Die dieser Dissertation zu Grunde liegende Studie birgt einige Schwächen. Da das Studiendesign nur ein einziges Zentrum und einen einzigen Arzt vorsah, könnten die Ergebnisse beim Versuch einer Reproduktion der Ergebnisse gegebenenfalls variieren. Dazu dürften unter Umständen auch unterschiedliche Software- und Hardwareversionen beitragen. Außerdem sind die Ergebnisse möglicherweise nicht mit denen der MRT-gestützten Strahlentherapie vergleichbar, da diese zu noch genaueren Segmentierungen führen könnte (Sritharan et al. 2022). Ungeachtet dessen hat die KI-gestützte Konturierung das Potenzial, im Vergleich zur rein manuellen Konturierung, ärztliche Arbeitsstunden einzusparen und die Konsistenz der Konturierung sowie die Arbeitsabläufe zu verbessern. Wenn sich diese Technologie weiterentwickelt, hat sie das Potenzial, den Bereich der Strahlentherapie grundlegend zu verändern.

## 1.5 Zusammenfassung

Die Autokonturierungstechnologie hat die Strahlentherapie bereits verändert. Ihre Fähigkeit Konsistenz, Genauigkeit und Effizienz im Konturierungsprozess zu gewährleisten, hat zu besseren Behandlungsergebnissen und möglicherweise auch zu geringeren Nebenwirkungen für die Patienten geführt. Mit stetiger Weiterentwicklung der Technologie wird das Autokonturieren eine noch bedeutendere Rolle bei der Gestaltung der Zukunft der Präzisionsstrahlentherapie spielen. In dieser Dissertation wurden potenzielle Vorteile für bestimmte anatomische Regionen ermittelt. Unter anderem Konsistenz und Genauigkeit, da die automatische KI-basierte Konturierung eine reproduzierbare und präzise Abgrenzung der klinischen Zielvolumina und der Risikoorgane gewährleistet. Diese Konsistenz verringert die Variabilität zwischen den

Strahlenonkologen und minimiert das Fehlerpotenzial bei der Behandlungsplanung. Darüber hinaus ist die Effizienz durch die automatische Konturierung erhöht und die für die Konturierung benötigte Zeit wird erheblich reduziert, so dass sich die Ärzte auf die Optimierung des Behandlungsplans konzentrieren können. Diese Effizienz ist besonders in Fällen mit komplexer Anatomie wertvoll. Außerdem könnte durch die Verringerung der Variabilität zwischen den Beobachtern auch eine potenzielle Verringerung der unerwünschten Ereignisse erreicht werden, da die Konturen zwischen verschiedenen Patienten konsistenter werden. Eine genaue Definition der Risikoorgane in der Umgebung eines Zielvolumens könnte dazu beitragen, dieses Ziel zu erreichen. Beim Vergleich der Dosisverteilung von Arzt- und KI-generierten Zielvolumina wurden keine signifikanten Unterschiede in der Exposition der Risikoorgane beobachtet. Dies ergänzt die Evidenz, die durch den Vergleich der Reproduzierbarkeit der Risikoorgane zwischen den drei untersuchten Gruppen gewonnen wurde und gibt zusätzlich Aufschluss über den tatsächlichen klinischen Einfluss und die Auswirkungen, die KI auf die Strahlentherapiepraxis haben könnte.

Mit Unterstützung durch KI ist es bereits möglich, Arbeitszeit und Arbeitskräfte für andere Aufgaben als die Konturierung einzusetzen, ohne dabei die Qualität zu beeinträchtigen oder das Wohlergehen der Patienten zu gefährden. Nach Ansicht des eingebundenen Konturierungsexperten wird dieses Konzept durch einen nur minimalen Korrekturbedarf noch verstärkt. Eine auf dieser Arbeit aufbauende Kosten-Nutzen-Analyse könnte diese Untersuchung sinnvoll ergänzen und die Rolle der KI in der Strahlentherapie weiter definieren.

## 1.6 Literaturverzeichnis

- Aoyama T, Shimizu H, Kitagawa T, Yokoi K, Koide Y, Tachibana H, Suzuki K and Kodaira T. Comparison of atlas-based auto-segmentation accuracy for radiotherapy in prostate cancer. *Phys Imaging Radiat Oncol* 2021; 19: 126-130
- Brouwer CL, Steenbakkers RJ, Bourhis J, Budach W, Grau C, Gregoire V, van Herk M, Lee A, Maingon P, Nutting C, O'Sullivan B, Porceddu SV, Rosenthal DI, Sijtsema NM and Langendijk JA. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol* 2015; 117,1: 83-90
- Buelens P, Willems S, Vandewinckele L, Crijns W, Maes F and Weltens CG. Clinical evaluation of a deep learning model for segmentation of target volumes in breast cancer radiotherapy. *Radiother Oncol* 2022; 171: 84-90
- Cha E, Elguindi S, Onochie I, Gorovets D, Deasy JO, Zelefsky M and Gillespie EF. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiother Oncol* 2021; 159: 1-7
- Chin V, Finnegan RN, Chlap P, Otton J, Haidar A, Holloway L, Thwaites DI, Dowling J, Delaney GP and Vinod SK. Validation of a Fully Automated Hybrid Deep Learning Cardiac Substructure Segmentation Tool for Contouring and Dose Evaluation in Lung Cancer Radiotherapy. *Clin Oncol (R Coll Radiol)* 2023; 35,6: 370-381
- Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, Chang JS and Kim JS. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiother Oncol* 2020; 153: 139-145
- Dolz J, Kirisli HA, Fechter T, Karnitzki S, Oehlke O, Nestle U, Vermandel M and Massoptier L. Interactive contour delineation of organs at risk in radiotherapy: Clinical evaluation on NSCLC patients. *Med Phys* 2016; 43,5: 2569
- Hall WA, Paulson E, Davis BJ, Spratt DE, Morgan TM, Dearnaley D, Tree AC, Efstathiou JA, Harisinghani M, Jani AB, Buyyounouski MK, Pisansky TM, Tran PT, Karnes RJ, Chen RC, Cury FL, Michalski JM, Rosenthal SA, Koontz BF, Wong AC, Nguyen PL, Hope TA, Feng F, Sandler HM and Lawton CAF. NRG Oncology Updated International Consensus Atlas on Pelvic Lymph Node Volumes for Intact and Postoperative Prostate Cancer. *Int J Radiat Oncol Biol Phys* 2021; 109,1: 174-185

Harms J, Lei Y, Tian S, McCall NS, Higgins KA, Bradley JD, Curran WJ, Liu T and Yang X. Automatic delineation of cardiac substructures using a region-based fully convolutional network. *Med Phys* 2021; 48,6: 2867-2876

Hernandez V, Hansen CR, Widesott L, Back A, Canters R, Fusella M, Gotstedt J, Jurado-Bruggeman D, Mukumoto N, Kaplan LP, Koniarova I, Piotrowski T, Placidi L, Vaniqui A and Jornet N. What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. *Radiother Oncol* 2020; 153: 26-33

Hosny A, Bitterman DS, Guthier CV, Qian JM, Roberts H, Perni S, Saraf A, Peng LC, Pashtan I, Ye Z, Kann BH, Kozono DE, Christiani D, Catalano PJ, Aerts H and Mak RH. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *Lancet Digit Health* 2022; 4,9: e657-e666

Mir R, Kelly SM, Xiao Y, Moore A, Clark CH, Clementel E, Corning C, Ebert M, Hoskin P, Hurkmans CW, Ishikura S, Kristensen I, Kry SF, Lehmann J, Michalski JM, Monti AF, Nakamura M, Thompson K, Yang H, Zubizarreta E, Andratschke N and Miles E. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. *Radiother Oncol* 2020; 150: 30-39

Munbodh R, Roth TM, Leonard KL, Court RC, Shukla U, Andrea S, Gray M, Leichtman G and Klein EE. Real-time analysis and display of quantitative measures to track and improve clinical workflow. *J Appl Clin Med Phys* 2022; 23,9: e13610

Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Sola AB, Kirova YM, Pignol JP, Remouchamps V, Verhoeven K, Weltens C, Arenas M, Gabrys D, Kopek N, Krause M, Lundstedt D, Marinko T, Montero A, Yarnold J and Poortmans P. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer, version 1.1. *Radiother Oncol* 2016; 118,1: 205-208

Patrick HM, Souhami L and Kildea J. Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention. *Acta Oncol* 2021; 60,2: 229-236

Pera O, Martinez A, Mohler C, Hamans B, Vega F, Barral F, Becerra N, Jimenez R, Fernandez-Velilla E, Quera J and Algara M. Clinical Validation of Siemens' Syngo.via Automatic Contouring System. *Adv Radiat Oncol* 2023; 8,3: 101177

Poortmans PMP, Takanen S, Marta GN, Meattini I and Kaidar-Person O. Winter is over: The use of Artificial Intelligence to individualise radiation therapy for breast cancer. *Breast* 2020; 49: 194-200

Rhee DJ, Akinfenwa CPA, Rigaud B, Jhingran A, Cardenas CE, Zhang L, Prajapati S, Kry SF, Brock KK, Beadle BM, Shaw W, O'Reilly F, Parkes J, Burger H, Fakie N, Trauernicht C, Simonds H and Court LE. Automatic contouring QA method using a deep learning-based autocontouring system. *J Appl Clin Med Phys* 2022; 23,8: e13647

Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, Garden AS, Peterson CB, Beadle BM and Court LE. Automatic detection of contouring errors using convolutional neural networks. *Med Phys* 2019; 46,11: 5086-5097

Sarria GR, Kugel F, Roehner F, Layer J, Dejonckheere C, Scafa D, Koeksal M, Leitzen C and Schmeel LC. Artificial Intelligence-Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics. *Adv Radiat Oncol* 2024; 9,3: 101394

Savjani RR, Lauria M, Bose S, Deng J, Yuan Y and Andrearczyk V. Automated Tumor Segmentation in Radiotherapy. *Semin Radiat Oncol* 2022; 32,4: 319-329

Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, Dahele M and Gillespie EF. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol* 2021; 160: 185-191

Spoor DS, Sijtsema NM, van den Bogaard VAB, van der Schaaf A, Brouwer CL, Ta BDP, Vliegenthart R, Kierkels RGJ, Langendijk JA, Maduro JH, Peters FBJ and Crijs APG. Validation of separate multi-atlases for auto segmentation of cardiac substructures in CT-scans acquired in deep inspiration breath hold and free breathing. *Radiother Oncol* 2021; 163: 46-54

Sritharan K, Dunlop A, Mohajer J, Adair-Smith G, Barnes H, Brand D, Greenlay E, Hijab A, Oelfke U, Pathmanathan A, Mitchell A, Murray J, Nill S, Parker C, Sundahl N and Tree AC. Dosimetric comparison of automatically propagated prostate contours with manually drawn contours in MRI-guided radiotherapy: A step towards a contouring free workflow? *Clin Transl Radiat Oncol* 2022; 37: 25-32

Tsang Y, Hoskin P, Spezi E, Landau D, Lester J, Miles E and Conibear J. Assessment of contour variability in target volumes and organs at risk in lung cancer radiotherapy. *Tech Innov Patient Support Radiat Oncol* 2019; 10: 8-12

van der Veen J, Gulyban A and Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiother Oncol* 2019; 137: 9-15

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

Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, Kolbeck C, Giambattista J, Gondara L and Alexander A. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol* 2020; 144: 152-158

## Scientific Article

# Artificial Intelligence–Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics



Gustavo R. Sarria, MD,<sup>a,\*</sup> Fabian Kugel, PhD,<sup>a</sup> Fred Roehner, PhD, MD,<sup>a</sup> Julian Layer, MD,<sup>a,b</sup> Cas Dejonckheere, MD,<sup>a</sup> Davide Scafa, MD,<sup>a</sup> Muentaz Koeksal, MD,<sup>a</sup> Christina Leitzen, MD,<sup>a</sup> and Leonard Christopher Schmeel, MD<sup>a</sup>

<sup>a</sup>Department of Radiation Oncology; and <sup>b</sup>Institute of Experimental Oncology, University Hospital Bonn, University of Bonn, Bonn, Germany

Received 20 June 2023; accepted 9 October 2023

**Purpose:** The study's purpose was to compare the performance of artificial intelligence (AI) in auto-contouring compared with a human practitioner in terms of precision, differences in dose distribution, and time consumption.

**Methods and Materials:** Datasets of previously irradiated patients in 3 different segments (head and neck, breast, and prostate cancer) were retrospectively collected. An experienced radiation oncologist (MD) performed organs-at-risk (OARs) and standard clinical target volume delineations as baseline structures for comparison. AI-based autocontours were generated in 2 additional CT copies; therefore, 3 groups were assessed: MD alone, AI alone, and AI plus MD corrections (AI+C). Differences in Dice similarity coefficient (DSC) and person-hour burden were assessed. Furthermore, changes in clinically relevant dose-volume parameters were evaluated and compared.

**Results:** Seventy-five previously treated cases were collected (25 per segment) for the analysis. Compared with MD contours, the mean DSC scores were higher than 0.7 for 74% and 80% of AI and AI+C, respectively. After corrections, 17.1% structures presented DSC score deviations higher than 0.1 and 10.4% dose-volume parameters significantly changed in AI-contoured structures. The time consumption assessment yielded mean person-hour reductions of 68%, 51%, and 71% for breast, prostate, and head and neck cancer, respectively.

**Conclusions:** In great extent, AI yielded clinically acceptable OARs and certain clinical target volumes in the explored anatomic segments. Sparse correction and assessment requirements place AI+C as a standard workflow. Minimal clinically relevant differences in OAR exposure were identified. A substantial amount of person-hours could be repurposed with this technology.

© 2023 The Author(s). Published by Elsevier Inc. on behalf of American Society for Radiation Oncology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## Introduction

Artificial intelligence (AI) is being increasingly related to current advances in medical care worldwide. Automation

of processes and usage optimization of both material and human resources are 2 aspects to highlight among its advantages. The extent of its capabilities is a subject of research in different medical areas, such as oncology. In this regard, many opportunities arise in the field of radiation therapy (RT).<sup>1</sup>

Closely linked to technological development, workflows in RT depend in great manner on soft- and hardware availability. Quality assurance, treatment precision, and patient

Sources of support: This work had no specific funding.

Data will be shared upon reasonable request.

\*Corresponding author: Gustavo R. Sarria, MD; E-mail: [gustavo.sarria@ukbonn.de](mailto:gustavo.sarria@ukbonn.de)

<https://doi.org/10.1016/j.adro.2023.101394>

2452-1094/© 2023 The Author(s). Published by Elsevier Inc. on behalf of American Society for Radiation Oncology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

safety are important items related to the latter, which have been historically performed by human practitioners.<sup>2</sup> As a part of these, a correct segmentation of organs at risk (OARs) and treatment target volumes defines a starting point for the subsequent involvement of other technological components.<sup>3</sup> This duty commonly lies on the professional team of an RT department and is particularly time-consuming, which could widely vary according to expertise and case difficulty. The process involves interpreting complex medical images and making critical decisions about where to target the radiation delivery. Additionally, manual contouring is subject to interobserver variability (ie, different clinicians may contour the same image differently), leading to inconsistencies and potential errors.<sup>4,5</sup> Nonetheless, AI has partly replaced human-performed tasks, enhancing the overall efficacy at all workflow levels.<sup>1</sup>

Modern AI solutions include deep learning algorithms, such as convolutional neural networks (CNNs), which are trained on large datasets of medical images. CNNs can identify patterns in medical images and learn to distinguish between different anatomic structures, such as tumors and healthy tissue.<sup>6</sup> Despite these recent advances, AI still faces certain challenges. There is a need for validation studies to ensure that the AI algorithms are accurate and reliable. These should focus on comparing AI-powered autocontouring with manual contouring by experts to ensure that the AI algorithm produces comparable or better results.<sup>7</sup> On the other hand, there are also concerns about a potential replacement of human expertise with AI-powered autocontouring. Although AI can enhance the accuracy and consistency of contouring, it cannot replace the knowledge and clinical judgment of radiation oncologists, physicists, or radiation therapists. Therefore, it is essential to ensure that AI is used as a tool to support and optimize clinical decision-making rather than as a replacement for human expertise.

Given the abovementioned facts, we deemed relevant to address these concerns and hypothesized that AI-powered autocontouring has the potential to produce human-like contours, while avoiding major deviations in dose distribution and saving significant person-hours.

## Methods and Materials

### Target and OAR delineation

An internal assessment at our department among 5 radiation oncologists was performed to determine which practitioner could achieve the most precise and swift delineations based on template cases defined according to the European Society for Radiotherapy and Oncology, the American Society for Radiation Oncology, or the Global Harmonization Group contouring guidelines.<sup>8-11</sup> This expert would assume thereafter all manual contours, which were to be considered the baseline standard for comparison.

Cases previously treated at our department encompassing prostate (PC), breast (BC), and head and neck cancer (HNC) were screened for DICOM datasets. These cases encompassed adjuvant treatment cases for BC, primary for PC, and both treatment modalities for HNC, and were selected randomly until reaching the target number, regardless of anatomic variants. Only cases with nonanatomic elements (bolus, prostheses other than oral, pacemakers, etc) were excluded. After selection, they were anonymized and DICOM sets were tripled for intergroup comparison. The medical doctor contours (MD) were defined on the first DICOM set, in a controlled environment assuring no distractions and dedicated time, to emulate an ideal scenario. To this end, using built-in autodelineation and interpolation tools was allowed. For comparison purposes, the total contouring time was measured and registered. Baseline structures included OARs in all cases and certain clinical target volumes (CTVs) in PC and BC segments, according to recommendations from international guidelines.<sup>8-10</sup> These structures are listed in [Table 1](#). Once the MD-baseline contours were created, deep learning-based autocontours were generated on the remaining 2 DICOM set copies with Limbus Contour v.1.6.0 (Limbus AI Inc), yielding an AI-only and AI plus MD corrections (AI+C) sets. The former remained unmodified, while the same MD manually corrected the latter, in order to achieve structures better fitting the actual clinical practice, if necessary. Differences between all contouring groups were assessed by means of the Dice similarity coefficient (DSC). A 0.7 or lower value was deemed as a geometric major difference. Corrections from AI with a  $\geq 0.1$  difference were considered major as well. All delineations were created with Eclipse version 15.5 (Varian Medical Systems).

### Planning and dose delivery comparison

The original dose distribution was retrieved and placed unmodified on the MD sets for baseline assessment. Furthermore, these were also rigidly copied on the AI and AI +C contours and raw differences in OAR exposure were compared with the baseline, even if the constraints were not met, as the purpose was to identify the magnitude of deviations in organ exposure. Several constraints corresponding to common practice were adopted for each segment ([Table 1](#)), according to different published trials or guidelines ([Appendix E1](#)).

### Time-consumption assessment

Times from AI processing, MD, and AI+C were collected for descriptive purposes and person-hours were calculated for comparison between groups.

**Table 1** Anatomic segments, structures, and dose constraints

Segment	Structures	Constraint
Breast	Lungs	D <sub>50</sub> (b)
		D <sub>35</sub> (b)
		D <sub>15</sub> (b)
	Breasts*	D <sub>5</sub> (c)
		D <sub>0.1 cc</sub> (c)
	Heart	D <sub>mean</sub>
		D <sub>10</sub>
		D <sub>5</sub>
	LAD	D <sub>mean</sub>
		V <sub>30</sub>
	Spinal cord	D <sub>0.035 cc</sub>
		D <sub>2 cc</sub>
	Thyroid	D <sub>mean</sub>
		Esophagus
Prostate	Bowel	V <sub>45</sub>
		V <sub>45</sub>
	Bladder	V <sub>70</sub>
		V <sub>45</sub>
	Anal canal	D <sub>mean</sub>
	Femurs	V <sub>50</sub> (b)
		D <sub>0.035 cc</sub> (b)
	Penile bulb	V <sub>50</sub>
	Rectum	V <sub>72</sub>
		V <sub>70</sub>
		V <sub>45</sub>
	Cauda equina	D <sub>0.035 cc</sub>
	Lymphatic pathways CTV	-
	Seminal vesicles	-
Prostate	-	
Head and neck	Plexi	D <sub>3 cc</sub> (b)
		D <sub>0.035 cc</sub> (b)
	Oral cavity	D <sub>mean</sub>
	Esophagus	D <sub>mean</sub>
		V <sub>35</sub>
	Submandibular	D <sub>mean</sub> (b)
	Thyroid	D <sub>mean</sub>
	Larynx	D <sub>mean</sub>
		D <sub>0.035 cc</sub>
	Lips	D <sub>mean</sub>
		D <sub>0.035 cc</sub>
Pharyngeal constrictor	D <sub>mean</sub>	

(continued on next page)

**Table 1** (Continued)

Segment	Structures	Constraint
	Parotids	D <sub>mean</sub> (b)
		D <sub>7 cc</sub> (b)
	Spinal cord	D <sub>0.035 cc</sub>
	Mandible	-

*Abbreviations:* b = bilateral; c = contralateral; CTV = clinical target volume; D = dose delivered at relative (%) volume (or absolute if indicated); LAD = left anterior descending artery; V = volume receiving certain dose (Gy).

\* Breast contours were used as both organs at risk and CTV. Structure and constraint list per anatomic segment. The constraint references are listed in the [Appendix E1](#).

## Statistical analysis

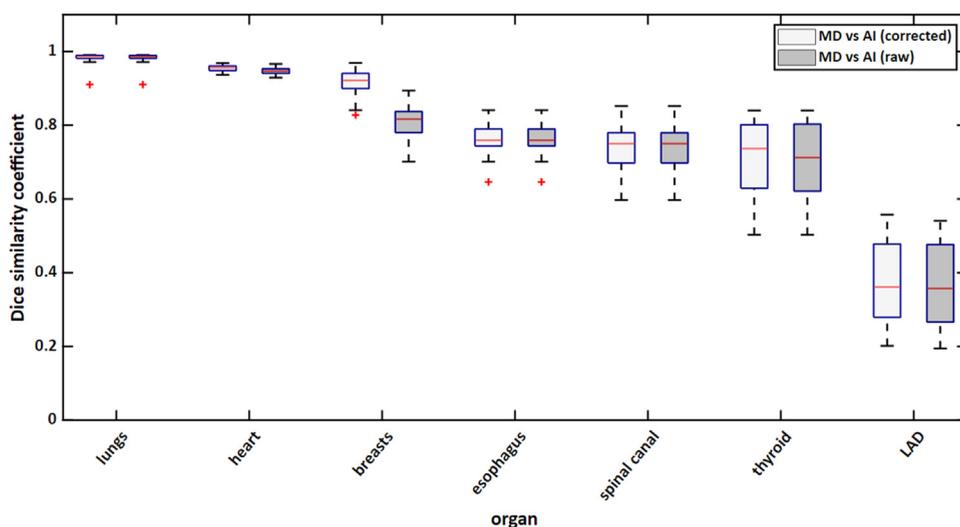
Mean DSC values plus standard deviation (SD) were registered for each studied anatomic segment in AI and AI+C and matched to those MD-generated for descriptive purposes. Differences in the resulting mean OAR-doses for each constraint were compared with the unpaired *t* test for independent variables. A *P* < .05 value determined a significant difference. The analysis was performed with MATLAB R2020b (The MathWorks Inc).

## Ethics statement

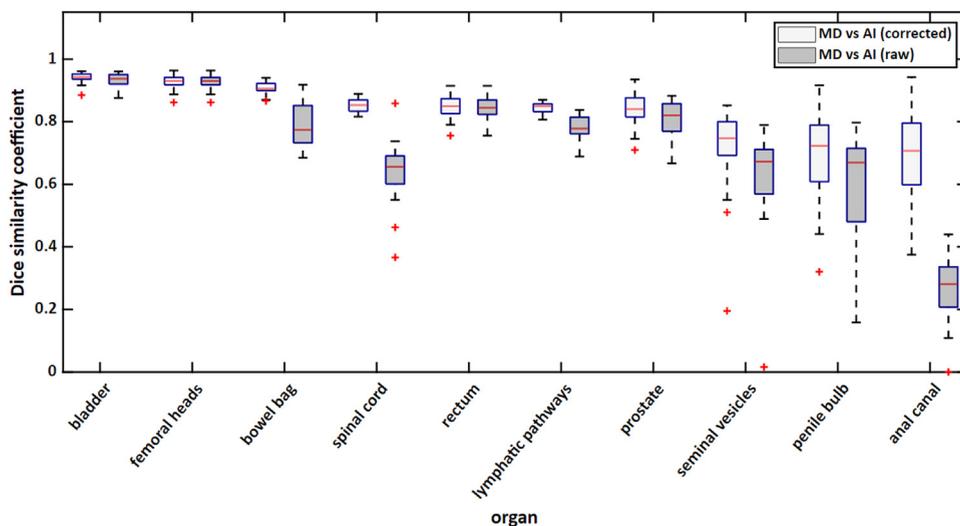
Institutional review board release was obtained before initiation, due to the retrospective nature of this study. All datasets were anonymized before inclusion. This investigation was conducted according to the principles of the Declaration of Helsinki.

## Results

Seventy-five cases were collected (25 for each anatomic segment) for the analysis. For BC structures, the overall mean DSCs were 0.82 (SD ± 0.19) for AI+C and 0.79 (SD ± 0.18) for AI. All structures showed DSC scores higher than 0.7, excepting for the left anterior descending artery (LAD) with 0.37 (SD ± 0.11) and 0.36 (SD ± 0.11) for AI +C and AI, respectively (Fig. 1). For PC, the overall mean DSC were 0.84 (SD ± 0.09) and 0.75 (SD ± 0.18). Major differences were detected for anal canal (0.68, SD ± 0.14, and 0.27, SD ± 0.1) and penile bulb (0.69, SD ± 0.13, and 0.60, SD ± 0.17). The cauda equina and seminal vesicles showed major differences only in the AI group (0.64, SD ± 0.09, and 0.62, SD ± 0.15; Fig. 2). Regarding HNC, the overall mean DSC were 0.72 (SD ± 0.15) and 0.70 (SD ± 0.16). Major differences were found in brachial plexi right (0.42, SD ± 0.07, and 0.37, SD ± 0.11) and left (0.43, SD



**Figure 1** Dice similarity coefficient scores in breast cancer. MD baseline contours were compared with AI alone and AI plus MD corrections. *Abbreviations:* AI = artificial intelligence; MD = experienced radiation oncologist.



**Figure 2** Dice similarity coefficient scores in prostate cancer. MD baseline contours were compared with AI alone and AI plus MD corrections. *Abbreviations:* AI = artificial intelligence; MD = experienced radiation oncologist.

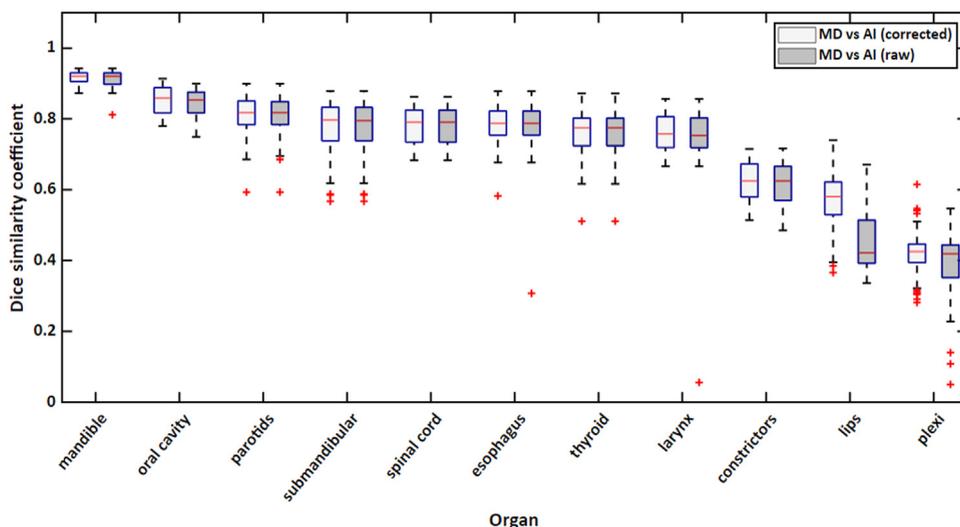
± 0.06, and 0.41, SD ± 0.09), lips (0.57, SD ± 0.09, and 0.45, SD ± 0.09), and pharyngeal constrictor (0.62, SD ± 0.06, and 0.61, SD ± 0.07; Fig. 3). Major corrections were required for anal canal, bowel bag, cauda equina, breasts, and lips, which represent 17.1% of all structures.

In terms of OAR exposure, significant dose distribution differences were identified for BC AI in contralateral breast D<sub>0.1 cc</sub> ( $P = .041$ ). For PC AI+C and AI in mean bowel V<sub>45</sub> ( $P = .0007$  and  $P < .0001$ ), anal canal D<sub>mean</sub> ( $P = .0051$  and  $P = .0013$ ) and cauda equina D<sub>0.035 cc</sub> (AI only,  $P < .0001$ ). In HNC, only mean lips D<sub>0.035 cc</sub> ( $P = .02$  and  $P = .014$ ) were significantly different. These translate into 10.4% of all structures presenting with dose-absorption deviations. The absolute dose differences per each constraint and box plots can be observed in Appendix E1 and E2, respectively.

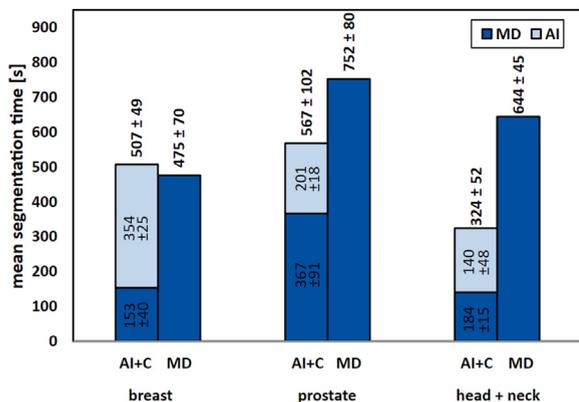
The time consumption analysis yielded mean MD segmentation times (seconds) of 475 (SD ± 70), 752 (SD ± 80), and 644 (SD ± 45), for BC, PC, and HNC, respectively. Similarly, the mean correction times of AI-generated contours were 153 (SD ± 40), 367 (SD ± 91), and 187 (SD ± 15), and the AI processing times were 354 (SD ± 25), 201 (SD ± 18), and 140 (SD ± 48). Proportionally, these represent mean person-hour reductions of 68%, 51% and 71% for each of the explored anatomic segments, respectively (Fig. 4).

### Discussion

CNNs have shown great promise in the field of automation in medicine. They have proven quite robust at analyzing



**Figure 3** Dice similarity coefficient scores in head and neck cancer. MD baseline contours were compared with AI alone and AI plus MD corrections. *Abbreviations:* AI = artificial intelligence; MD = experienced radiation oncologist.



**Figure 4** Mean ( $\pm$  standard deviation) segmentation times in seconds for MD, AI, and AI+C contours in all 3 investigated anatomic segments. *Abbreviations:* AI = artificial intelligence; C = corrections; MD = experienced radiation oncologist.

images and identifying patterns, making them optimal for the task of automated target volume delineation.<sup>12</sup> Advances in the past few years have led to a migration from atlas-based autocontouring to deep learning-based autocontours, which might adjust better to variabilities in daily practice and outperform the former, even suggesting corrections to clinically validated models.<sup>13</sup> In our study, we challenged the capacities of the Limbus Contour software against a highly experienced practitioner and evaluated the reach this could have on clinical routine. The 3 evaluated segments were selected as representative indications for patients treated at any RT department, according to their frequency.

The DSC is a useful statistical tool for contour comparison, by taking 2 different segmentations and measuring their overlap ratio. These values range from 0 to 1, where 1 represents full accordance between both structures.

Although occasionally this might be subject to bias, depending on factors such as structure volume or number of voxels, it still represents one of the strongest available methods for performing these measurements.<sup>14</sup> In our analysis, the DSC showed overall acceptable correlations between the MD, AI, and AI+C delineations. Most significant differences were identified in small or long organs, albeit without major clinical meaning. In addition, it should be considered that all corrections (AI+C) were done until achieving clinically acceptable standards, as per the MD’s judgment. The low correction rates of the AI-generated contours (17.1% of organs had DSC change >0.1) demonstrates a quite impressive accuracy, requiring minor to no correction at all. Moreover, the dosimetric analysis showed that, even after modifications, these almost never led to significant changes in the clinically relevant DVH parameters. In BC cases, the most remarkable differences were identified in LAD contours. This is relatively common, as pointed out in different studies assessing this structure, and highly dependent on imaging quality and anatomic configuration.<sup>15-17</sup> Nevertheless, the MD deemed these differences of little clinical relevance, as most of them were related to discrepancies at width, upper end, or lower end. Further minor differences were identified in esophagus, thyroid, and spinal cord, also linked to the previously mentioned. Noteworthy, the AI+C breast contours were converted into CTVs with great similarity to the standard MD (DSC 0.92). Our results are consistent with other previous studies addressing this same topic.<sup>6,18,19</sup> We must mention that no lymphatic pathway contours were considered in this segment, due to technical reasons. The software version available at the moment of the study counted with the American Society for Radiation Oncology templates, while our contours are routinely performed according to the European Society for Radiotherapy and Oncology recommendations. In

regard to PC, the most notable differences were observed in the anal canal, lymphatic pathways, penile bulb, and seminal vesicles. The first showed substantial deviations between MD and AI contours. A reason for this might be related to the template used by the software, which evidently differs from the Global Harmonization Group consensus guidelines,<sup>8</sup> creating smaller structures not reaching the outer sphincter. A similar situation was observed for the seminal vesicles, perhaps due to their variable anatomy. Likely, differences in total volume and cranial or caudal length led to lower DSC scores in the remaining OARs, mostly not clinically relevant (eg, cauda equina). Worth highlighting, the lymphatic pathways reached a satisfactory configuration with minor editing (DSC 0.85). These outcomes resemble those of Cha et al, although we included more OARs in our analysis, providing additional robustness.<sup>20</sup> For the HNC segment, no CTV analysis was considered, due to a higher inhomogeneity in volume prescription. Several differences in DSC scores were found in more organs than in the other explored segments, possibly due to the smaller nature of these structures. Yet, most of them were slightly below a 0.8 DSC score and above our 0.7 lower-limit score. Certain specific cases, such as the brachial plexus, presented with higher deviations. For practical reasons, our plexus contour includes the bony structures between foramina. This is to ensure a homogeneous dose distribution along the plexus area and minimize possible hotspots nearby. In this sense, the AI algorithm provides a more anatomic delineation, respecting the vertebral bodies, and in accordance with international guidelines.<sup>8</sup> Moreover, the plexus portions between both scalene muscles, which are the ones potentially closer to the treated volumes, were rather similar in all 3 groups. Clinically judged and despite the DSC outcomes, these differences were considered of minor relevance. Regarding metrics, we deem relevant to stress some points. The arbitrary cut-off point was adopted according to previous publications,<sup>21-23</sup> although no strong consensus has been reached to date. A reason for this might lie on a difference between metrics and clinical applicability, as the former does not necessarily guarantee the latter.<sup>24,25</sup> Consequently, these values must be taken as reference but should not directly affect clinical decision-making. Further metrics were not considered for the analysis (eg, Hausdorff distance) because single linear metrics do not provide a proper oversight of volume overlap. Instead, we deemed relevant to measure the differences in dose distribution among the 3 investigational groups, to achieve a better appreciation of a potential clinical effect. By placing the actual irradiated plan on the MD contours and assuming this as the baseline standard, we found that only 5 of 48 dose-volume parameters varied according to the segmentation method. These changes were located mostly in organs with large size or challenging anatomy. Furthermore, as we mainly report statistical differences, these could mean either lower or higher

exposure rates (details in [Appendix E1](#)). Regardless of this, most differences were minimal and could be deemed clinically not relevant, supporting our findings based on the DSC coefficient.

The time-sparing analysis yielded interesting differences along all 3 groups. For instance, the AI output was evidently faster in the PC and HNC contours. Although it was also faster in the BC sets, this advantage appeared to be minor. This could be related to the number of structures per segment and their complexity. In addition, interpolation and integrated autodelineation tools perform better usually in larger organs, such as lungs or breasts. As expected, the AI+C times were considerably shorter than the MD's, resulting in a significant reduction of person-hours. It must be remarked that the MD contours were defined in a controlled, favorable environment, which is practically impossible to reproduce in daily practice. Moreover, our expert MD has vast experience in the field and has practiced for several years as contouring trainer. We established this ideal scenario in order to challenge the AI as much as possible. In a real-life setting, the benefit could be even more noticeable if other distracting factors are to be accounted for.

However, it is important to note that AI algorithms are not perfect, and there are some limitations to their accuracy. For example, they may struggle with low-quality images, artifacts, anatomic variants, or unusual positioning settings. Additionally, there is a considerable risk of overreliance on autocontouring algorithms, which can potentially lead to errors or skipping thorough assessments. To mitigate these risks, it is mandatory to incorporate manual review of autocontours into the treatment planning process. Therefore, we established AI+C as our standard workflow. Another important consideration is the risk of de-skilling among radiation oncologists and physicists, as manual contouring may be less frequently used with the increased use of AI solutions. To address these issues, it is compulsory to ensure training in the use of AI and acquire the skills necessary to review and adjust autocontours as needed. This can help ensure that the AI is used as a tool to improve clinical workflow, rather than a replacement for human expertise.

Our study carries certain shortcomings. Its single-center and single-practitioner character could yield variable outcomes, when trying to reproduce this methodology. The software and hardware version available at different centers could also contribute to the latter. Besides the abovementioned, our outcomes might not be comparable to those of MRI-guided RT, as these could result in even more accurate segmentations.<sup>26</sup> Notwithstanding, AI-based contouring has the potential to save person-hours, and to improve contouring consistency and workflows in comparison to human-performed manual contouring. As this technology continues to evolve, it has the potential to revolutionize the field of RT.

## Conclusion

In great extent, AI yielded acceptable OARs and certain CTVs in the explored anatomic segments. Sparse correction and assessment requirements place AI+C as a standard workflow. Minimal clinically relevant differences in OAR-exposure were identified. A substantial number of person-hours could be repurposed with this technology.

## Disclosures

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Supplementary materials

Supplementary material associated with this article can be found in the online version at [doi:10.1016/j.adro.2023.101394](https://doi.org/10.1016/j.adro.2023.101394).

## References

1. Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiother Oncol.* 2020;153:55-66.
2. Munbodh R, Roth TM, Leonard KL, et al. Real-time analysis and display of quantitative measures to track and improve clinical workflow. *J Appl Clin Med Phys.* 2022;23:e13610.
3. Hernandez V, Hansen CR, Widesott L, et al. What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. *Radiother Oncol.* 2020;153:26-33.
4. Patrick H M, Souhami L, Kildea J. Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention. *Acta Oncol.* 2021;60:229-236.
5. van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiother Oncol.* 2019;137:9-15.
6. Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol.* 2020;144:152-158.
7. Savjani RR, Lauria M, Bose S, Deng J, Yuan Y, Andrearczyk V. Automated tumor segmentation in radiotherapy. *Semin Radiat Oncol.* 2022;32:319-329.
8. Mir R, Kelly SM, Xiao Y, et al. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. *Radiother Oncol.* 2020;150:30-39.
9. Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer, version 1.1. *Radiother Oncol.* 2016;118:205-208.
10. Hall W A, Paulson E, Davis BJ, et al. NRG Oncology updated International Consensus Atlas on pelvic lymph node volumes for intact and postoperative prostate cancer. *Int J Radiat Oncol Biol Phys.* 2021;109:174-185.
11. Brouwer C L, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol.* 2015;117:83-90.
12. Poortmans PMP, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: The use of artificial intelligence to individualise radiation therapy for breast cancer. *Breast.* 2020;49:194-200.
13. Rhee DJ, Cardenas CE, Elhalawani H, et al. Automatic detection of contouring errors using convolutional neural networks. *Med Phys.* 2019;46:5086-5097.
14. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol.* 2004;11:178-189.
15. Chin V, Finnegan RN, Chlap P, et al. Validation of a fully automated hybrid deep learning cardiac substructure segmentation tool for contouring and dose evaluation in lung cancer radiotherapy. *Clin Oncol (R Coll Radiol).* 2023;35:370-381.
16. Harms J, Lei Y, Tian S, et al. Automatic delineation of cardiac substructures using a region-based fully convolutional network. *Med Phys.* 2021;48:2867-2876.
17. Spoor DS, Sijtsma NM, van den Bogaard VAB, et al. Validation of separate multi-atlases for auto segmentation of cardiac substructures in CT-scans acquired in deep inspiration breath hold and free breathing. *Radiother Oncol.* 2021;163:46-54.
18. Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiother Oncol.* 2020;153:139-145.
19. Buelens P, Willems S, Vandewinckele L, et al. Clinical evaluation of a deep learning model for segmentation of target volumes in breast cancer radiotherapy. *Radiother Oncol.* 2022;171:84-90.
20. Cha E, Elguindi S, Onochie I, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiother Oncol.* 2021;159:1-7.
21. Aoyama T, Shimizu H, Kitagawa T, et al. Comparison of atlas-based auto-segmentation accuracy for radiotherapy in prostate cancer. *Phys Imaging Radiat Oncol.* 2021;19:126-130.
22. Pera O, Martinez A, Mohler C, et al. Clinical validation of Siemens' syngo.via automatic contouring system. *Adv Radiat Oncol.* 2023;8:101177.
23. Dolz J, Kirisli H A, Fechter T, et al. Interactive contour delineation of organs at risk in radiotherapy: Clinical evaluation on NSCLC patients. *Med Phys.* 2016;43:2569.
24. Sherer MV, Lin D, Elguindi S, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol.* 2021;160:185-191.
25. Rhee DJ, Akinfenwa CPA, Rigaud B, et al. Automatic contouring QA method using a deep learning-based autocontouring system. *J Appl Clin Med Phys.* 2022;23:e13647.
26. Sritharan K, Dunlop A, Mohajer J, et al. Dosimetric comparison of automatically propagated prostate contours with manually drawn contours in MRI-guided radiotherapy: A step towards a contouring free workflow? *Clin Transl Radiat Oncol.* 2022;37:25-32.

### **3. Danksagung**

Mein besonderer Dank im Rahmen dieser Arbeit gilt Herrn PD Dr. med. Christopher Schmeel, meinem Doktorvater für die Überlassung des Themas. Als mein Mentor setzte er ein wunderbares Beispiel von unermüdlicher Hilfsbereitschaft, Kameradschaft und Freundlichkeit. Ich bin ihm zu großem Dank verpflichtet.

Mein Dank gilt gleichermaßen auch meinen Eltern und Großeltern, die meine Neugier und meine Interessen stets mit ihrem vollen Einsatz und Liebe unterstützt haben. Ihr Beispiel war mir immer Ansporn und Inspiration. Darüber hinaus danke ich meiner geliebten Frau Adriana für Ihre ständige Geduld, Unterstützung und die entgangene gemeinsame Zeit.

Außerdem danke ich all denen, die mich bei der Erstellung meiner Dissertation begleitet und unterstützt haben, besonders Fabian Kugel, die mir die Fertigstellung dieser Arbeit erst ermöglichten.