Bridging Geometric Accuracy and Clinical Utility:

A Framework for Assessing Deep Learning Auto-Segmentation Tools

Zach Gude, PhD[1]; Jun Lian, PhD[1]; Shiva Das, PhD[1];
[1] University of North Carolina Medical Center, Department of Radiation Oncology

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Purpose

To integrate (1) **qualitative treatment-planner assessments** with (2) **quantitative spatial metrics** to establish a framework for evaluating the **clinical utility** of deep learning (DL)-based auto-segmentation tools.

Introduction

A.I. tools are increasingly powerful and prevalent in many clinical processes. But for highly impactful decisions, the non-intuitive and unpredictable failure modes of A.I. requires expert oversight / intervention to ensure quality care. Therefore, the quality of a tool should not only measure its independent performance, **but its impact when combined with user expertise.**

Organ at risk (OAR) contouring can be a laborious bottle-neck in RT workflows. **RayStation 2024A SP3**[1] includes a built-in deep learning-based (DL) auto-segmentation tool, capable of quickly contouring a large library of OARs. But individual OARS present unique contouring challenges *(volume, contrast, curvature complexity, inter-slice transitions, inter-patient variability, etc.)*. These challenges may impact human planners and DL tools differently.

Which OARs would this DL auto-segmentation tool be useful?
Importantly, how do you define "useful" in this clinical context?

Materials & Methods

Dataset Overview

50 patient CT-sims were retrospectively contoured with RayStation's DL auto-segmentation tool, creating **349 clinically approved OAR structures with a DL counterpart,** spanning **35 unique organs** across **5 broad treatment-sites**

"Clinical" Structures: Manually contoured, MD approved, & used in Tx	**"DL" Structures:** OARs generated by *RayStation's DL Auto-Segmentation tool*

Qualitative Analysis: 5 teams of *1 dosimetrist and 1 DABR physicist* each score the utility of 1 body-sites **DL Structures,** based on edits required

- ***0** – Major changes, **1** – Moderate changes, **2** – Minor-to-no changes necessary*
- *Qualitative score + correction time & other observations to determine **"Clinical Utility"***

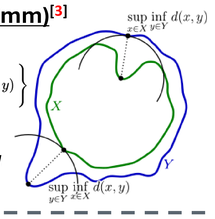
Quantitative Analysis: MIM Version 7.3.7[2] calculated **spatial agreement** *between "**Clinical**" and "**DL**" structures*

Dice Score (3D)

$$\frac{2 * |X \cap Y|}{|X| + |Y|} = \frac{2 \times}{X + Y}$$

Hausdorff Dist. (mm)[3]

$$d_H(X,Y) := \max\left\{ \sup_{x \in X} d(x,y),\ \sup_{y \in Y} d(X,y) \right\}$$



Maximum minimum dist between surface 1 & surface 2

Linear Regression – *Matlab's fitlm*[4]

Model Formula → Relationship between **predictors** and **outcomes**
*($y = \beta_0 + \beta_1 X_1$, where β_0 & β_1 are **coefficients**, & X_1 is the **predictor**)*

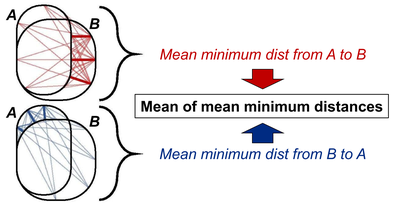
***Coeff p-Value* →** T-test for likelihood that a *specific* predictor can explain outcomes when controlling for all predictors *(< 0.05 statistically significant & smaller the better!)*

F-Statistic* →** *Quantify how much the output variance can be explained by any predictor variable, compared to null-hypothesis **(larger the better!)

R^2* →** Coefficient of determination ***(0-1, larger the better)

RMSE* →** Root Mean Squared Error ***(smaller the better!)

Mean Surface Dist. (mm) [5]



Qualitative Analysis: Clinical Utility (CU) Score

High Clinical Utility *(should be used)*

#	OAR Name	Score	STD
1	Bladder	1.55	0.69
2	Bone_Mandible	2.00	0.00
3	Brain	2.00	0.00
4	Cavity_Oral	1.78	0.44
5	Cochlea_L	2.00	0.00
6	Cochlea_R	2.00	0.00
7	Eye_L	2.00	0.00
8	Eye_R	2.00	0.00
9	Femur_Head_L	1.82	0.40
10	Femur_Head_R	1.60	0.70
11	Glnd_Lacrimal_L	2.00	0.00
12	Glnd_Lacrimal_R	2.00	0.00
13	Kidney_L	1.65	0.58
14	Kidney_R	1.83	0.50
15	Lens_L	2.00	0.00
16	Lens_R	2.00	0.00
17	Liver	1.66	0.49
18	Lung_L	1.92	0.19
19	Lung_R	1.87	0.23
20	OpticNrv_L	2.00	0.00
21	OpticNrv_R	2.00	0.00
22	Prostate	1.57	0.53
23	Rectum	1.73	0.65
24	SeminalVes	1.57	0.53

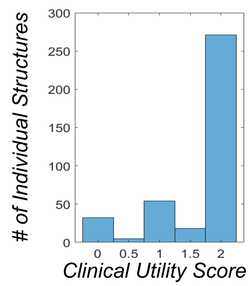
Moderate Clinical Utility *(use with caution)*

#	OAR Name	Score	STD
1	Esophagus	1.68	0.53
2	Glnd_Submand_L	1.78	0.44
3	Glnd_Submand_R	1.56	0.53
4	Heart	1.69	0.42
5	Parotid_L	1.40	0.70
6	Parotid_R	1.50	0.71
7	Stomach	1.54	0.61

Not Clinically Useful *(do not use!)*

#	OAR Name	Score	STD
1	A_LAD	0.54	0.48
2	Brainstem	0.10	0.32
3	Breast_L	0.00	0.00
4	Breast_R	0.00	0.00

Histogram: *Score Distribution of DL-Structures*



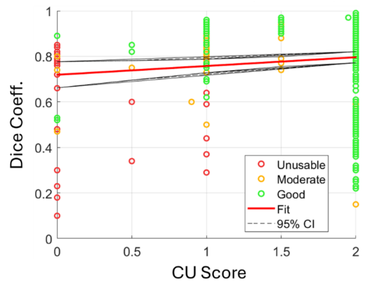
OAR's Grouped by Overall Clinical Utility:

- **24 OARs** have high clinical utility *(Mean Score = 1.85 ± 0.18)*
- **7 OARs** have moderate clinical utility *(Mean Score = 1.59 ± 0.13)*
- **4 OARs** are clinically unusable *(Mean Score = 0.16 ± 0.26)*

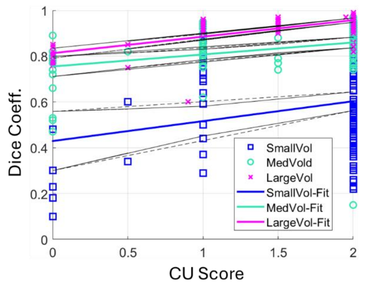
Quantitative Analysis *(Dice / Volume)*

What's the relationship between **Dice Coeff.** & **Clinical Utility** (CU)?

A **Linear Fit:** Dice Vs CU Score (All Data)



B **Linear Fit:** Dice Vs CU Score *Small/Medium/Large Volume Ranges*



(A) All 349 OARS plotted. Color-coded to indicate overall CU for an organ type

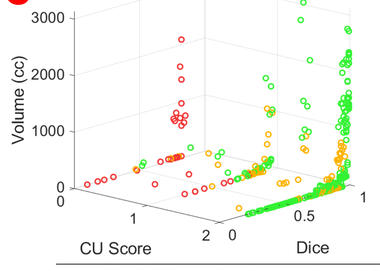
- **Dice** and **CU** aren't highly correlated
- Evenly splitting OARs into 3 volume-based groups improves linear fit
- OAR volume impacts the relationship

Model Formula: $Score = \beta_0 + \beta_1 Dice$

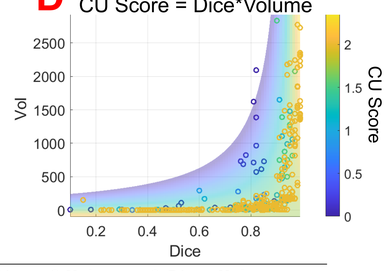
Predictor Variable	F-Stat	β_1 p-Val	R^2	RMSE
(A) Dice *(All Data)*	5.69	0.0176	0.0162	0.638
(B) Dice *(Small Vol)*	6.01	0.0158	0.0501	0.522
(B) Dice *(Medium Vol)*	15.4	1.47e-4	0.12	0.675
(B) Dice *(Large Vol)*	125	5.311e-20	0.522	0.446

How does **Volume** interplay with **Dice Coeff.** and **CU**?

C **3D Plot:** *Dice Vs CU Vs Volume*



D **Lin-Reg Model:** CU Score = Dice*Volume



Model Formula: $Score = \beta_0 + \beta_1 Dice + \beta_2 Volume + \beta_3 Dice * Volume$

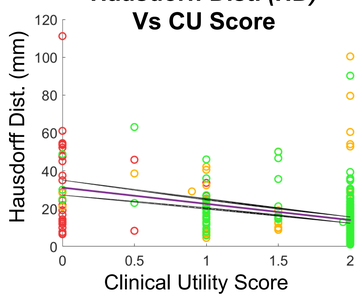
Predictor Variable	F-Stat	p-Value	R^2	RMSE	β_1 p-Val	β_2 p-Val	β_3 p-Val
(D) Dice * Volume	29.1	9.06e-17	0.203	0.576	4.539e-16	1.048e-17	2.357e-17

Volume is known to impact **Dice** scores[6] when evaluating contour accuracy
Importantly, it also impacts the way Clinical Utility is evaluated!

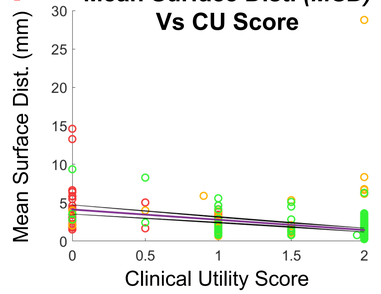
Quantitative Analysis *(HD / MSD)*

How do contour **surface errors** correlate with **CU**?

E **Hausdorff Dist.** ***(HD)*** **Vs CU Score**



F **Mean Surface Dist.** ***(MSD)*** **Vs CU Score**



- **Surface-based metrics** correlate with **CU** scores better than **Dice**
- They provide distinct, yet related geometric agreement insight

Model Formula: $Score = \beta_0 + \beta_1 Predictor$

Predictor Variable	F-Stat	β_1 p-Val	R^2	RMSE
(E) Hausdorff Dist.	56.3	5.41e -13	0.140	0.597
(F) Mean Surface Dist.	59.5	1.35e-13	0.147	0.594

Model Formula: $Score = \beta_0 + \beta_1 HD + \beta_2 MSD + \beta_3 HD * MSD$

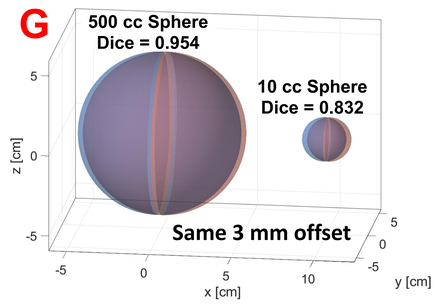
Predictor Variable	F-Stat	p-Val	R^2	RMSE	β_1 p-Val	β_2 p-Val	β_3 p-Val
HD * MSD	49.4	1.49e-26	0.302	0.539	0.0716	2.129e-16	1.161e-14

Combined, they improve correlation to CU Score!

Discussion

What is the **rationale** for these results? Do they provide **insight**?

G



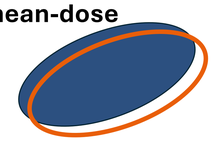
Example (G): *Two sphere-pairs, each with a 3 mm offset, but at different volumes.* **Large OAR with small offset might take a long time to edit.**

- This could **reduce the CU Score**
- But have a **high Dice Coeff.**

H

Impacts max-dose
Large **HD**
Small **MSD**

Impacts mean-dose
Small **HD**
Large **MSD**



Example (H): HD and MSD quantify different dosimetric impacts. **The HD and MSD correlation to clinical utility may change for serial vs parallel.**

Final Takeaway

Quantifying the clinical utility of an auto-segmentation tool requires nuance beyond basic geometric agreement. Accounting for the time of planner's edits, and organ-specific dose constraints might better indicate impact on workflow efficiency.

References

1. *RayStation (Version 2024A SP3) Software. Stockholm, Sweden: RaySearch Laboratories AB.*
2. *MIM Maestro. Version 7.3.7. MIM Software Inc; 2026. https://www.mimsoftware.com*
3. *By Rocchini - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=2918812*
4. *MathWorks. (2026). Statistics and Machine Learning Toolbox (Version R2026a) Software. Natick, MA: The MathWorks, Inc.*
5. *Kiser, K.J., Barman, A., Stieb, S. et al. Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow. J Digit Imaging 34, 541–553 (2021). https://doi.org/10.1007/s10278-021-00460-3*
6. *Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015 Aug 12;15:29. doi: 10.1186/s12880-015-0068-x. PMID: 26263899; PMCID: PMC4533825.*

Thank You!

Qualitative Assessment Teams

Physicists (DABR)
- Jun Lian, Ph.D.
- Shiva Das, Ph.D.
- Cielle Collins, M.S.
- David Fried, Ph.D.
- Leith Rankine, Ph.D.
- Michael Dance, M.S.

Dosimetrists (CMD)
- David Chang
- Heather Baliker
- Jackie Williamson
- Matt Hawkins
- Raina Smith