MASSACHUSETTS
GENERAL HOSPITAL

RADIATION ONCOLOGY

Evaluation Methods for Auto-Segmentation and Expected Results

Greg Sharp
July 13, 2022

AAPM 2022
JULY 10–14 | WASHINGTON, DC
64TH ANNUAL MEETING & EXHIBITION

CELEBRATING MEDICAL PHYSICS
TRANSFORMING HUMAN HEALTH

Disclosures

- No conflict of interest

Outline

- What do we want to measure?
- What can we measure?
- What should we expect?

What do we want to measure?

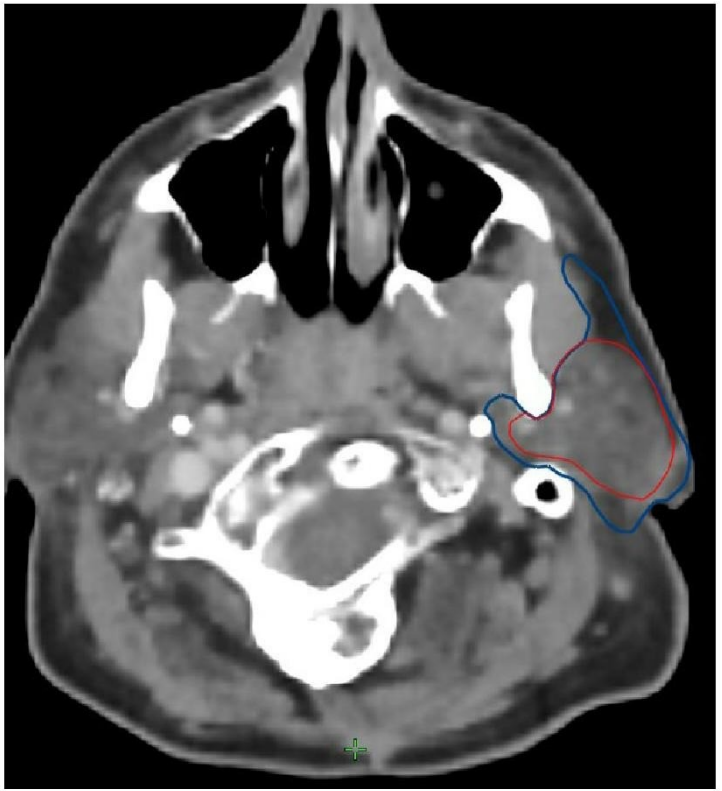
What do we want to measure?

know about our segmentation?

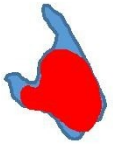
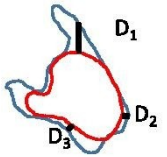
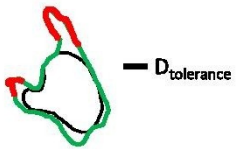
What do we want to measure?

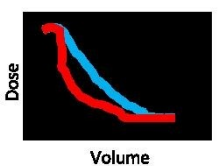
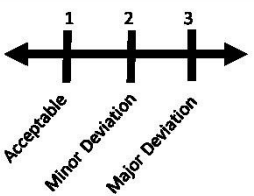
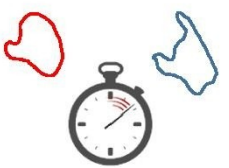
know about our segmentation?

- Is it good enough?
  - Does it need manual correction?
  - Does it affect treatment plan quality?
- Did it save us time?



Sherer 2021

Geometric Methods	Overlap	Boundary Distance	Path Length	Surface Agreement
				
Key examples	Volumetric DSC, JSC	Average and Maximum (Hausdorff) Distances	Added Path Length	Surface DSC
Strengths	Easy to compute	Sensitive to point positions	Better correlation with time spent contouring	Better correlation with time spent contouring
Weaknesses	Low sensitivity for complex boundaries	Does not account for proportion of contour requiring edits	May be less appropriate with "brush" contouring	Requires prespecified tolerance threshold

Other Methods	Dosimetric	Qualitative Scoring	Time
			
Strengths	Allows calculation of relevant validated parameters (e.g. parotid mean, lung V20)	Validated to be predict outcomes (clinical trials)	Reflects impact on clinical workflow
Weaknesses	Requires treatment planning (variable)	Subjective, review can be time consuming	Speed may not reflect quality

Time saving ?

YES	NO
La Maccia 2012 Daisne 2013 Langmack 2014 Walker 2014	Simmat 2012 Thomson 2014

Time saving ?

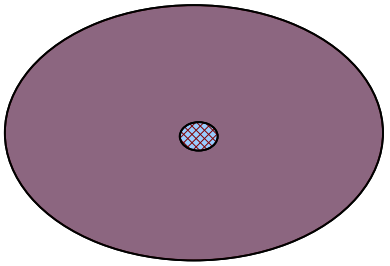
	Manual (min)	Edited (min)
Levendag (2017) Lustberg (2018) Van der Veen (2019) Cha (2021)	180 20 34 40	53 8 23 28

** Time savings decreases with proficiency/experience

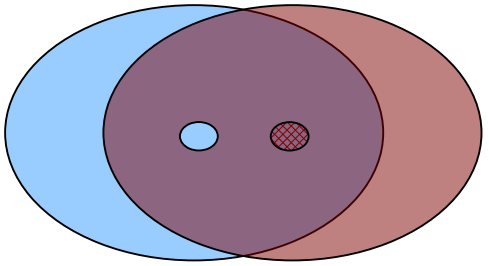
Qualitative scoring

- Typical system
  - 1: Accept
  - 2: Minor corrections needed
  - 3: Unusable
- Turing test
  - Manual segmentation
  - Automatic segmentation

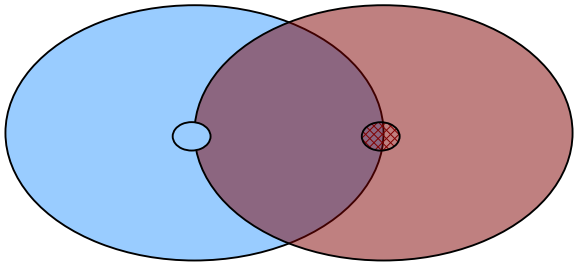
Dice coefficient



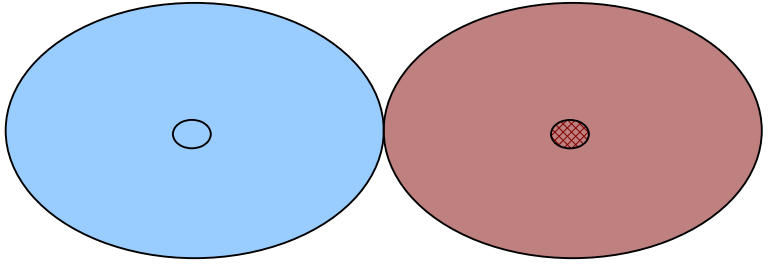
1



0.63



0.31



ZERO

Jaccard index vs Dice coefficient

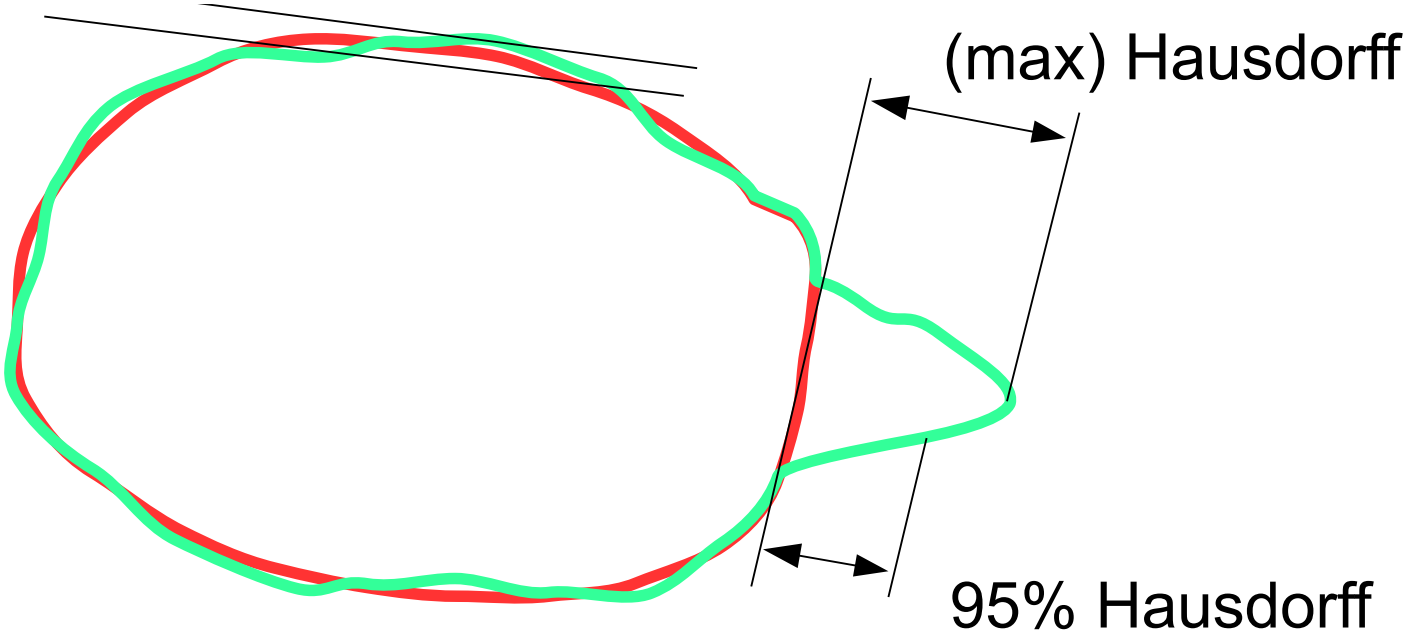
$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

$$J = \frac{|X \cap Y|}{|X \cup Y|}$$

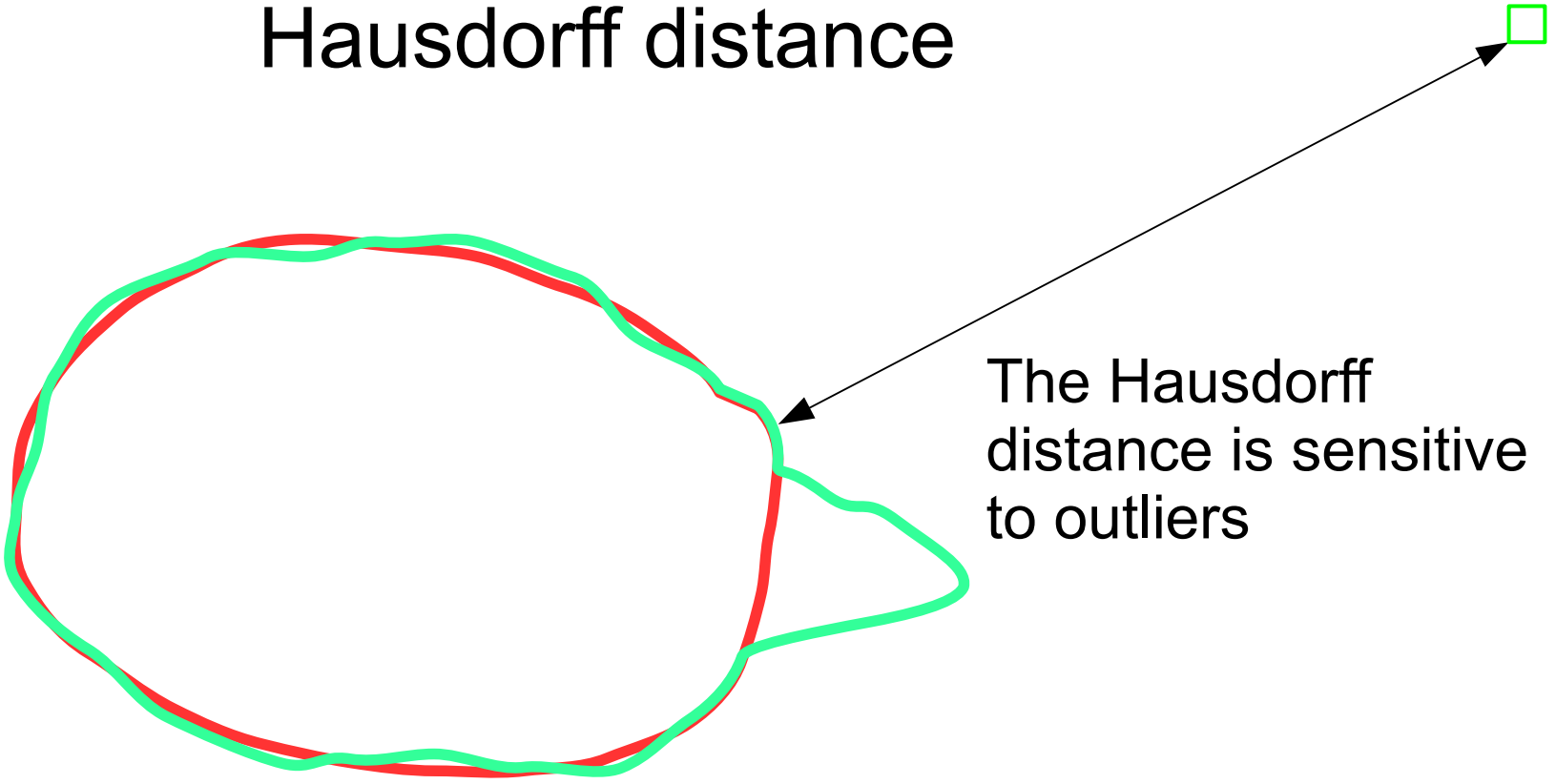
$$= \frac{DSC}{(2 - DSC)}$$

Hausdorff distance

Avg Surface distance

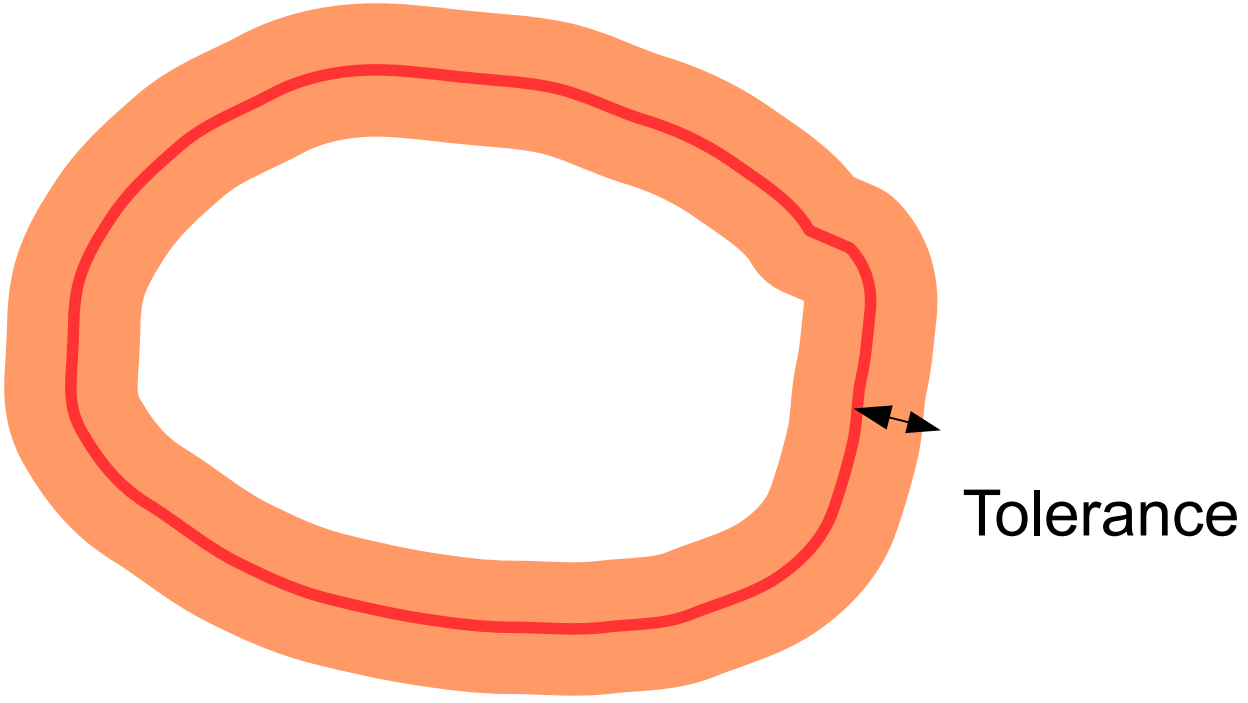


Hausdorff distance

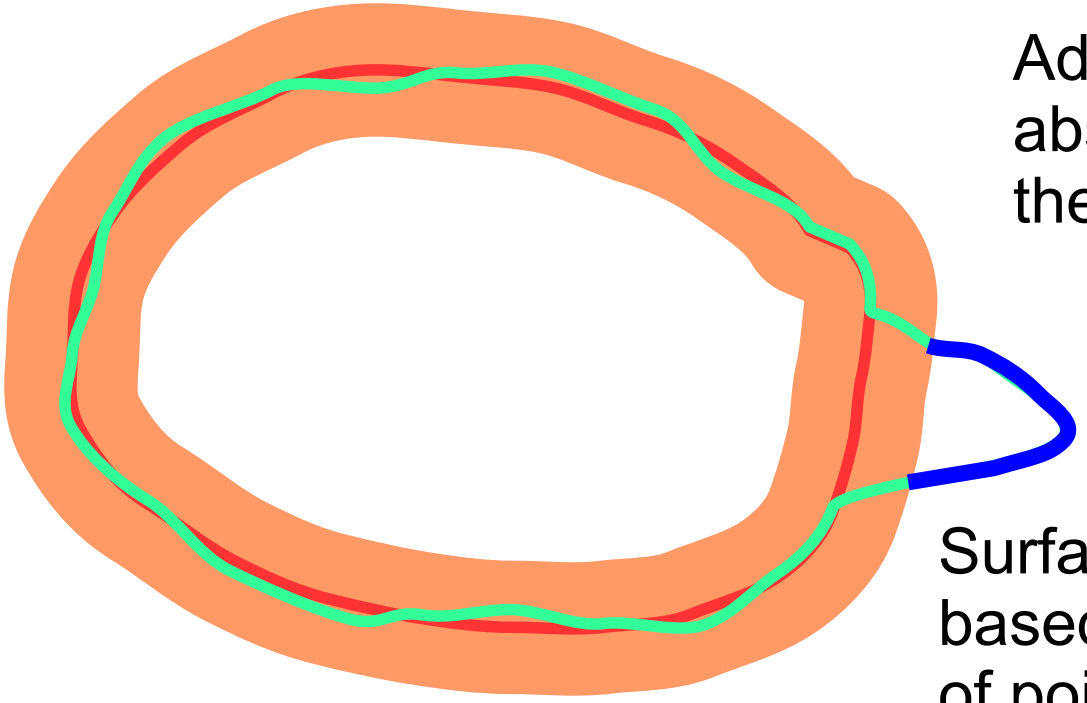


The Hausdorff distance is sensitive to outliers

Surface Dice / Added path length



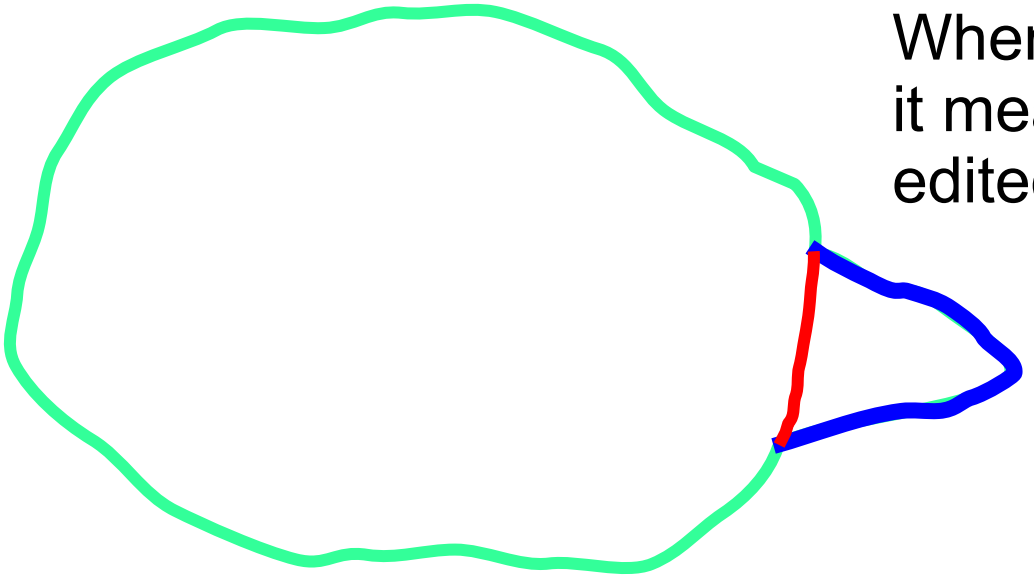
Surface Dice / Added path length



Added path length is absolute length outside the chosen tolerance

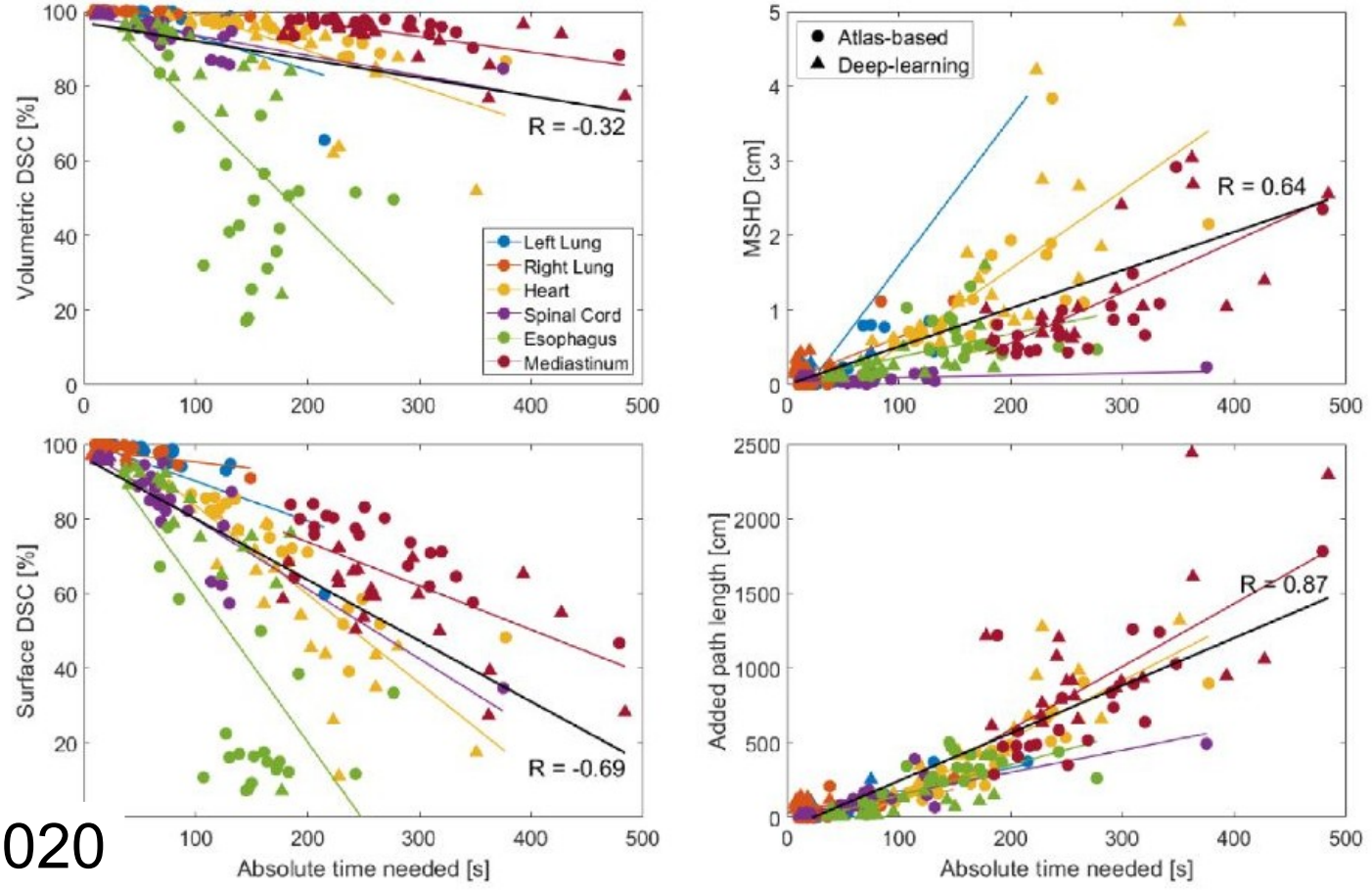
Surface Dice is DSC based on percent of points outside the chosen tolerance

Surface Dice / Added path length



When tolerance is zero,
it measures the distance
edited (e.g) by a pen tool

Correlation: Time savings

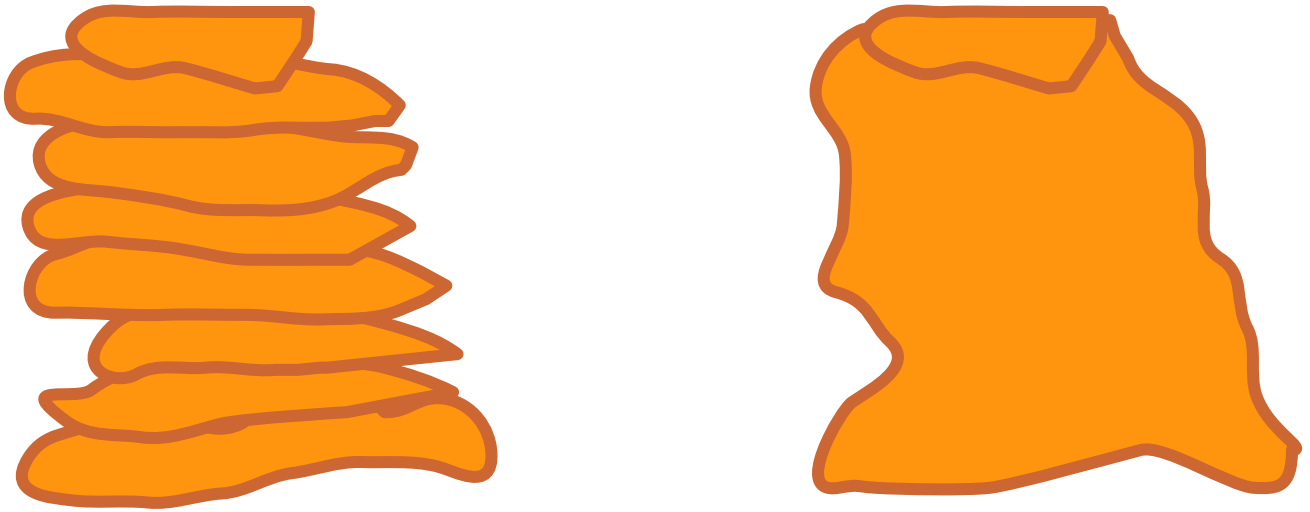


Vaassen 2020

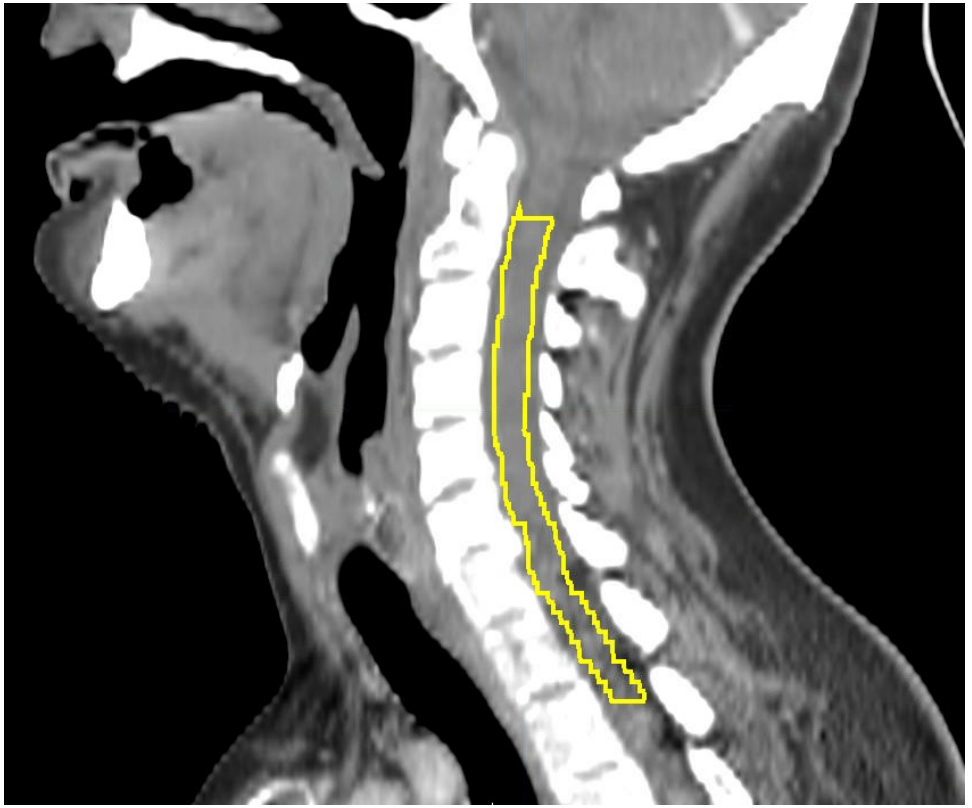
Metrics: In practice

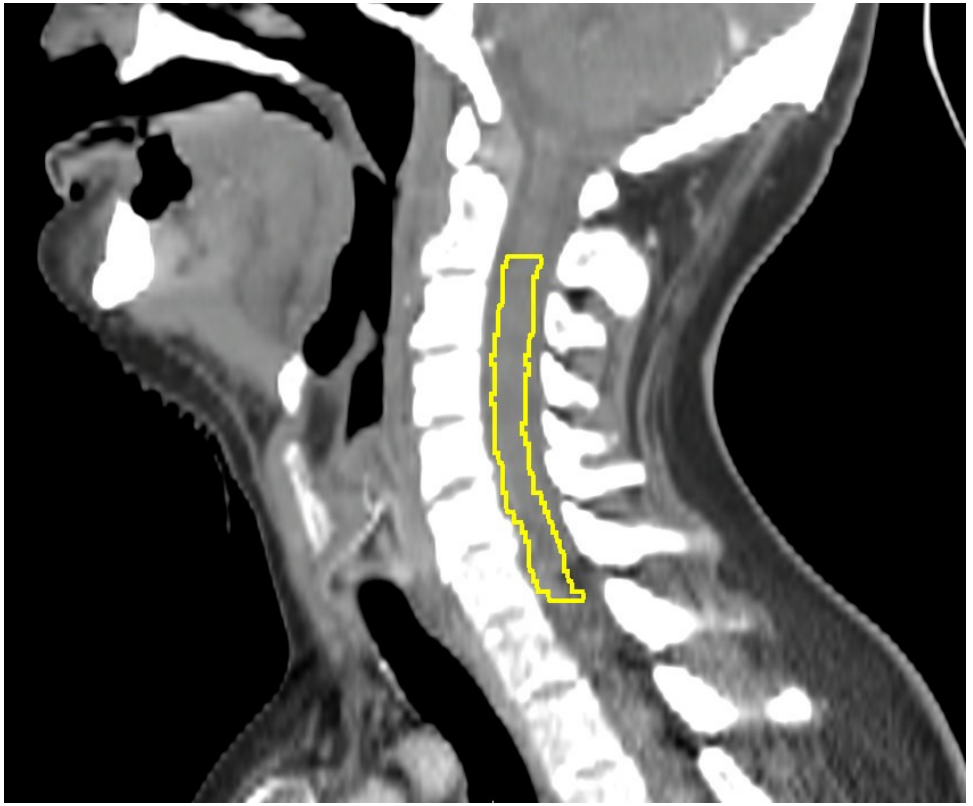
- No consensus on best metric
- DSC
  - Good for historical comparisons
  - Not well correlated to quality
- Hausdorff
  - Sensitive to outliers **(*SAM*)**
  - Consider 95HD instead
- Average surface distance
  - Generally good
- Surface Dice / APL
  - Correlated with editing effort **(*SAM*)**
- Consider multiple metrics

Volumetric vs Slicewise



Tubular structures





Caution on interpreting literature

Variations of the average surface distance (ASD)

ASSD	Average symmetric surface distance[26,53,57]	$\frac{\sum_{a\in\partial A} d(a,\partial B)+\sum_{b\in\partial B} d(b,\partial A)}{\lvert\partial A\rvert+\lvert\partial B\rvert}$
ASD^{max}	Average surface distance, maximum[35,64,66,72,75,76,98,99]	$\max\left\{\frac{\sum_{a\in\partial A} d(a,\partial B)}{\lvert\partial A\rvert}, \frac{\sum_{b\in\partial B} d(b,\partial A)}{\lvert\partial B\rvert}\right\}$
ASD^{mid}	Average surface distance, mid-value[24,32,40,55,56,61,81]	$\frac{1}{2}\left(\frac{\sum_{a\in\partial A} d(a,\partial B)}{\lvert\partial A\rvert}+\frac{\sum_{b\in\partial B} d(b,\partial A)}{\lvert\partial B\rvert}\right)$
$ASD^{n/a}$	Average surface distance, unspecified[39,58,75,79,88]	<unspecified>
DTA^{avg}	Average distance to agreement[27,42,68,77,84,87]	$\frac{\sum_{b\in\partial B} d(b,\partial A)}{\lvert\partial B\rvert}$

Vrtovec 2020

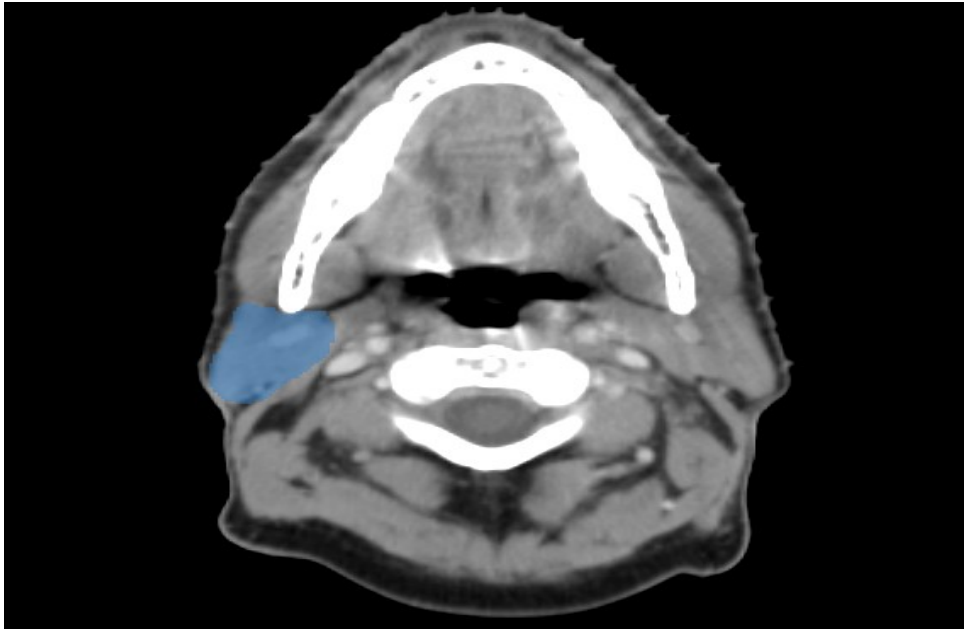
Quis custodiet ipsos custodes?

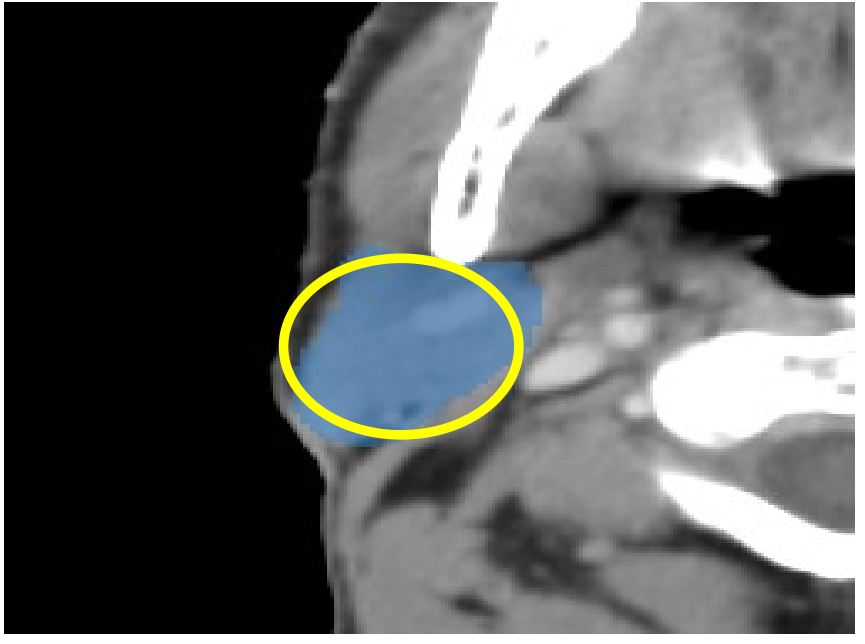
Test case	Max H.D. Software A (mm)	Max H.D. Software B (mm)	95% H.D. Software A (mm)	95% H.D. Software B (mm)
A	4.94	4.12	1.96	3.74
B	4.64	4.12	2.48	3.74
C	6.57	5.48	3.27	5.39
D	3.58	3.00	1.18	2.83
E	4.64	3.32	2.08	3.16
F	46.22	29.27	19.16	28.74

Unpublished

Expected results

- What is the Dice coefficient when matching a parotid gland with a sphere?

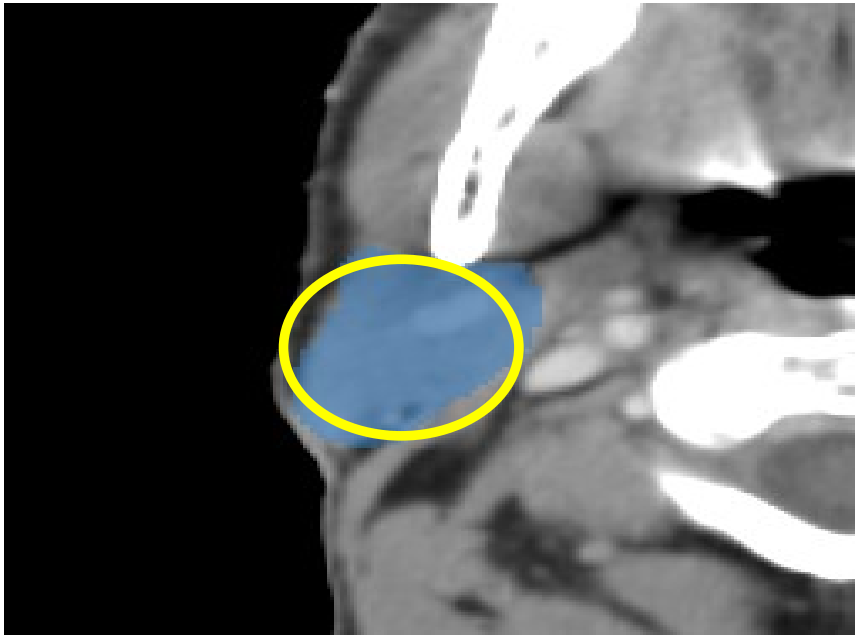




Expected results

- What is the Dice coefficient when matching a parotid gland with a sphere?

Dice coefficient = 0.74



Expected results: Male pelvis

	DSC	Hausdorff	Av Surf Dist (mm)
Prostate	0.85-0.90	5.5-7.5	1.5-3.0
Bladder	0.85-0.95	4.5-7.0	1.0-2.5
Rectum	0.85-0.90	6.0-9.0	1.5-4.0

**** WORK IN PROGRESS ****

Expected results: Thorax

	DSC	Hausdorff 95 (mm)	Av Surf Dist (mm)
Lungs	0.95-0.99	2.0-4.0	0.5-1.5
Heart	0.85-0.95	5.0-8.0	1.5-2.5
Esophagus	0.70-0.75	7.0-9.0	1.5-2.5
Spinal cord	0.85-0.90	1.5-2.5	0.5-1.5

**** WORK IN PROGRESS ****

Expected results: Head and neck

	DSC	Hausdorff 95 (mm)	Av Surf Dist (mm)
Brain Stem	0.80-0.90	1.5-2.5	1.0-1.5
Optic Chiasm	0.50-0.75	1.5-3.5	1.0-1.5
Parotid Glands	0.85-0.90	2.0-3.5	1.0-2.0
S-m Glands	0.75-0.85	2.0-4.0	1.5-2.0

**** WORK IN PROGRESS ****

Summary

- Metrics are hard

Summary

- Metrics are hard
- All classes of quantitative metric have problems
- Correlation with application objectives is weak