Evaluation Methods for Auto-Segmentation and Expected Results

Greg Sharp
July 13, 2022
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?

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

know about our segmentation?
## Geometric Methods

<table>
<thead>
<tr>
<th>Key examples</th>
<th>Overlap</th>
<th>Boundary Distance</th>
<th>Path Length</th>
<th>Surface Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volumetric DSC, JSC</td>
<td><img src="overlap.png" alt="Overlap" /></td>
<td><img src="bd.png" alt="Boundary Distance" /></td>
<td><img src="path.png" alt="Path Length" /></td>
<td><img src="surf.png" alt="Surface DSC" /></td>
</tr>
</tbody>
</table>

## 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

<table>
<thead>
<tr>
<th>Dosimetric</th>
<th>Qualitative Scoring</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose</td>
<td><img src="qual.png" alt="Qualitative Scoring" /></td>
<td><img src="time.png" alt="Time" /></td>
</tr>
<tr>
<td>Volume</td>
<td><img src="1.png" alt="1" /> <img src="2.png" alt="2" /> <img src="3.png" alt="3" /></td>
<td></td>
</tr>
</tbody>
</table>

## 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
<table>
<thead>
<tr>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>La Maccia 2012</td>
<td>Simmat 2012</td>
</tr>
<tr>
<td>Daisne 2013</td>
<td>Thomson 2014</td>
</tr>
<tr>
<td>Langmack 2014</td>
<td></td>
</tr>
<tr>
<td>Walker 2014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manual (min)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Levendag (2017)</td>
<td>180</td>
</tr>
<tr>
<td>Lustberg (2018)</td>
<td>20</td>
</tr>
<tr>
<td>Van der Veen (2019)</td>
<td>34</td>
</tr>
<tr>
<td>Cha (2021)</td>
<td>40</td>
</tr>
</tbody>
</table>

** 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

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

\[ J = \frac{|X \cap Y|}{|X \cup Y|} \]

\[ J = \frac{DSC}{2 - DSC} \]
Hausdorff distance

Avg Surface distance

(max) Hausdorff

95% Hausdorff
The Hausdorff distance is sensitive to outliers.
Surface Dice / Added path length

Tolerance
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

**ASD\(^{\text{max}}\)**
- Average surface distance, maximum

**ASD\(^{\text{mid}}\)**
- Average surface distance, mid-value

**ASD\(^{\text{n/a}}\)**
- Average surface distance, unspecified

**DTA\(^{\text{avg}}\)**
- Average distance to agreement

\[
\sum_{a \in \partial A} d(a, \partial B) + \sum_{b \in \partial B} d(b, \partial A) \\
\frac{|\partial A| + |\partial B|}{2} \\
\max \left\{ \frac{\sum_{a \in \partial A} d(a, \partial B)}{|\partial A|}, \frac{\sum_{b \in \partial B} d(b, \partial A)}{|\partial B|} \right\} \\
\frac{1}{2} \left( \frac{\sum_{a \in \partial A} d(a, \partial B)}{|\partial A|} + \frac{\sum_{b \in \partial B} d(b, \partial A)}{|\partial B|} \right) \\
<\text{unspecified}> \\
\sum_{b \in \partial B} d(b, \partial A)
\]

Vrtovec 2020
Quis custodiet ipsos custodes?

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

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

<table>
<thead>
<tr>
<th></th>
<th>DSC</th>
<th>Hausdorff</th>
<th>Av Surf Dist (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>0.85-0.90</td>
<td>5.5-7.5</td>
<td>1.5-3.0</td>
</tr>
<tr>
<td>Bladder</td>
<td>0.85-0.95</td>
<td>4.5-7.0</td>
<td>1.0-2.5</td>
</tr>
<tr>
<td>Rectum</td>
<td>0.85-0.90</td>
<td>6.0-9.0</td>
<td>1.5-4.0</td>
</tr>
</tbody>
</table>

**** WORK IN PROGRESS ****
Expected results: Thorax

<table>
<thead>
<tr>
<th></th>
<th>DSC</th>
<th>Hausdorff 95 (mm)</th>
<th>Av Surf Dist (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lungs</td>
<td>0.95-0.99</td>
<td>2.0-4.0</td>
<td>0.5-1.5</td>
</tr>
<tr>
<td>Heart</td>
<td>0.85-0.95</td>
<td>5.0-8.0</td>
<td>1.5-2.5</td>
</tr>
<tr>
<td>Esophagus</td>
<td>0.70-0.75</td>
<td>7.0-9.0</td>
<td>1.5-2.5</td>
</tr>
<tr>
<td>Spinal cord</td>
<td>0.85-0.90</td>
<td>1.5-2.5</td>
<td>0.5-1.5</td>
</tr>
</tbody>
</table>

**** WORK IN PROGRESS ****
Expected results: Head and neck

<table>
<thead>
<tr>
<th></th>
<th>DSC</th>
<th>Hausdorff 95 (mm)</th>
<th>Av Surf Dist (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain Stem</td>
<td>0.80-0.90</td>
<td>1.5-2.5</td>
<td>1.0-1.5</td>
</tr>
<tr>
<td>Optic Chiasm</td>
<td>0.50-0.75</td>
<td>1.5-3.5</td>
<td>1.0-1.5</td>
</tr>
<tr>
<td>Parotid Glands</td>
<td>0.85-0.90</td>
<td>2.0-3.5</td>
<td>1.0-2.0</td>
</tr>
<tr>
<td>S-m Glands</td>
<td>0.75-0.85</td>
<td>2.0-4.0</td>
<td>1.5-2.0</td>
</tr>
</tbody>
</table>

**** WORK IN PROGRESS ****
Summary

- Metrics are hard
Summary

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