### Survey with results summary

Please note; Any comments in italics represent the authors thoughts or interpretation

1. **Participant and software**

1.1. **Name of participant**

*Results intentionally not shown. This was used for contact purposes*

1.2. **Organization of participant**

*Results intentionally not shown. This was used for contact purposes*

1.3. **Name of the program used for the calculation**

| Name              | Number of occurrences |
|-------------------|-----------------------|
| ConquestTools     | 1                     |
| ContourScore      | 1                     |
| GoldenRule        | 1                     |
| MIM               | 1                     |
| Matlab 2021a      | 1                     |
| Python            | 3                     |
| Python+SimpleITK  | 1                     |
| ROIEval           | 1                     |
| WMatch            | 1                     |
| Djf_curveMetrics  | 1                     |
| Surface-distance  | 1                     |
1.4. Type of software

1.4 Type of software
13 responses

- Home made by me/my institution: 61.5%
- Free Off-the-shelf (downloaded from github etc): 30.8%
- Commercial package: 7.7%

1.5. Please provide the references if there are any academic publications relating to the implementation or definition of the tool

a. Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, van Elmp W. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology. 2020 Jan;1:13:1-6. 
   https://doi.org/10.1016/j.phro.2019.12.001

b. Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, Patel Y, Meyer C, Askham H, Romera-Paredes B, Kelly C. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430. 2018 Sep 12. 
   https://doi.org/10.48550/arXiv.1809.04430

c. Zhou R, Liao Z, Pan T, Milgrom SA, Pinnix CC, Shi A, Tang L, Yang J, Liu Y, Gomez D, Nguyen QN. Cardiac atlas development and validation for automatic segmentation of cardiac substructures. Radiotherapy and Oncology. 2017 Jan 1;122(1):66-71. 
   https://doi.org/10.1016/j.radonc.2016.11.016

d. Gooding MJ. On the Evaluation of Auto-Contouring in Radiotherapy. Auto-Segmentation for Radiation Oncology. 2021 Apr 18:217-52.

e. Heimann T, Van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, Bello F. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE transactions on medical imaging. 2009 Feb 10;28(8):1251-65. 
   https://doi.org/10.1109/TMI.2009.2013851
1.6. Please provide the references if there are any academic publications relating to the validation of the accuracy of the measure tool

*Note that these papers don’t actually relate to the validation of the measurement tool. The first provides motivation for a specific measurement, rather than validating accuracy of implementation. The second uses measurements for validating autocontouring.*

a. Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, van Elmpt W. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology. 2020 Jan 1;13:1-6. https://doi.org/10.1016/j.phro.2019.12.001

b. Huang K, Rhee DJ, Ger R, Layman R, Yang J, Cardenas CE, Court LE. Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. Journal of applied clinical medical physics. 2021 May;22(5):168-74.

1.7. Please provide links to implementation if the software is open source, or where a product can be bought/downloaded if generally available

| Link                                                                 | Occurrence in responses |
|----------------------------------------------------------------------|-------------------------|
| https://github.com/deepmind/surface-distance                        | 4                       |
| https://github.com/Auto-segmentation-in-Radiation-Oncology/Chapter-15 | 2                       |
| https://pypi.org/project/dicom2nifti/                                | 2                       |
| https://canislupusllc.com/portfolio-goldenrule/                       | 1                       |
| https://pydicom.github.io/                                           | 1                       |
| https://simpl elitk.org/                                             | 1                       |
| https://pypi.org/project/MedPy/                                      | 1                       |

2. General implementation

2.1. What is the primary representation used for the structure when calculating the measures?

2.1 What is the primary representation used for the structure when calculating the measures? 13 responses

![Diagram](https://example.com/diagram.png)

*Note that this is the primary representation. In some implementations multiple representations were used. For instance, an implementation may use a voxel-mask as its primary representation and use*
that to calculate DSC. That primary representation may then be used to create a mesh to calculate distance measures. This mesh could be viewed as a secondary or alternative representation required for some calculations. In this example the secondary representation is dependent on the accuracy of the first. The interpretation of this question was left to the participant, but as is seen in Q3.1 and Q4.1 multiple representations are used by some participants.

2.2. How are distance measures calculated?

2.3. How are measures reported?
3. Measurements available

3.1. Please select all similarity measures calculated by the tool and the representation used

Without providing sufficient options for the participants, some of these responses resulted in ambiguity that had to be resolved through discussion with the participant. It was unclear what a participant should select when a voxel mask was used as an intermediate step to create mesh. In this instance some participants selected the mesh as the representation from which the measurement was made, some selected the voxel mask as the input they needed to feed into the measurement code, and some selected both.

| Measure                      | 2D Contour-based | 2D Pixel/Mask-based | 3D Mesh-based | 3D Voxel/mask-based |
|------------------------------|------------------|---------------------|---------------|---------------------|
| True Positive Volume         | 1                | 1                   | 1             | 1                   |
| False Positive Volume        | 1                | 1                   | 1             | 1                   |
| True Negative Volume         | 1                | 1                   | 1             | 1                   |
| False Negative Volume        | 1                | 1                   | 1             | 1                   |
| True Positive Rate           | 1                | 1                   | 2             |                     |
| False Positive Rate          | 1                | 1                   | 2             |                     |
| Sensitivity                  | 1                |                     | 1             |                     |
| Specificity                  |                 |                     |               | 1                   |
| Dice Similarity Coefficient  | 1                | 1                   | 12            |                     |
| Jaccard Index                |                 |                     |               | 3                   |
| Hausdorff Distance           | 1                | 2                   | 5             | 9                   |
| 95% Hausdorff Distance       | 1                | 1                   | 4             | 6                   |
| Mean Surface Distance / Average distance | 1 | 1 | 5 | 8 |
| Median Surface Distance      | 1                |                     | 1             |                     |
| Added Path Length            | 2                |                     |               | 1                   |
| Surface DSC                  | 2                |                     | 2             |                     |
| Volume Difference            | 1                |                     | 1             |                     |
| Volume Ratio                 |                 |                     | 1             | 2                   |
| Centre of mass difference    | 1                |                     |               |                     |
3.2. Please list any other measures calculated by the tool, and the nature of the shape representation used

The full list of measurements is given in Supplementary Material 3.

4. Specific Implementation Questions

4.1. If you use a 3D mesh-based representation, how do you convert from RTSS to the mesh?

Some implementations using the Surface-distance code selected “Not applicable”. The surface-distance code creates a mesh, from an input of a voxel array for the purposes of weighting the distances by “surfel” area (the term “surfel” was used by the authors of the Surface DSC work to describe a surface element i.e. a triangle of the mesh), although not actually used to compute distance. Therefore, it can be interpreted that these participants are using that library as a black box.

4.2. If you use a mask or other voxel-based representation, how do you define whether a voxel is “inside” the structure when converting from RTSS?

[Diagrams showing distribution of responses]

- Majority (by proportion) of voxel inside the RTSS contour
- Any part of the voxel inside the RTSS contour
- Centre (or Center, depending how you want to spell it) of the voxel inside the RTSS contour
- Not applicable
- (Other) I think it’s center enclosed voxels by I will check again
- (Other) Not sure how it was implemented. I think it is majority, but I can be wrong
4.3. How do you define the tolerance for the Surface Dice?

4.4. How do you define the tolerance for the Added Path Length?

5. Distance measure implementation

5.1. For distance-based measures, how do you sample the surface/contour?
5.2. How do you define the non-symmetric average distance?

5.2 How do you define the non-symmetric average distance?
13 responses

- 36.5%: Equation a. The mean of \( d(r,s) \) over all points \( r \) on \( R \), based on number of sample points
- 30.0%: Equation b. The weighted mean of \( d(r,s) \) over all points \( r \) on \( R \) (explain weighting in Implementation PDF)
- 15.4%: Not applicable (N.B. This might not be a measure that you report, but may be used in the definition of the symmetric distance below - if so please still define it here)
- 15.4%: (Other) I compute differences for volume elements, not just surfaces. I do this because RT and DVHs are inherently volume-based, not surface
- 7.7%: (Other) Eq a, with the caveat that distances are sampled from distance maps rather than calculating distances between points

5.3. How do you define the symmetric average distance?

5.3 How do you define the symmetric average distance?
13 responses

- 36.5%: Equation a. The average of each one-way distance
- 23.1%: Equation b. The maximum of each one-way distance
- 15.4%: Equation c. The mean of one-way distances computed over all points on both surfaces
- 7.7%: Equation d. The weighted mean of weighted one-way distances over all points on both surfaces
- 7.7%: Not Applicable
- 7.7%: (Other) I implemented both Eq. a and c
- 7.7%: (Other) I computer differences for volume elements, not just surfaces. I do this because RT and DVHs are inherently volume-based, not surface

5.4. How do you define the symmetric Hausdorff Distance?

5.4 How do you define the symmetric Hausdorff Distance?
13 responses

- 92.3%: Equation a. Maximum of one-way Hausdorff distances
- 7.7%: Equation b. Average of one-way Hausdorff distances
- 7.7%: Not applicable
5.5. How do you define the symmetric 95% Hausdorff distance?

5.5 How do you define the symmetric 95% Hausdorff distance?
13 responses

- **Equation a.** The distance for which only 5% of points on the combined set of points from both surfaces exceed that value
- **Equation b.** The maximum of the one-way 95% Hausdorff distances
- **Equation c.** The average of the one-way 95% Hausdorff distances
- **Not applicable**

6. File uploads

6.1. Please attach the results excel spreadsheet

A summary of the results for commonly reported measures is given in Supplementary Material 4

6.2. Please include a PDF detailing the technical implementation of each of the measures (that you have implemented as stated in section 3). This should include; equations defining the measure, the representation used or any resampling done. i.e. any steps in the implementation that would someone to reproduce your measurements exactly. Where elements are common between measures (e.g. method of conversion to a mask), it is fine to include this once and refer back to it.

These files contain various levels of detail, from copies of code (most useful!) to commercial package training manuals (least useful!). In the end it was necessary to talk directly with most participants to fully understand the results and detect implementation errors.
7. One last question

7.1. What do you think are the most meaningful measures for assessing contouring in radiotherapy?
Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy
Mark J Gooding, Djamal Boukerroui, Eliana Vasquez Osorio, René Monshouwer, Ellen Brunenberg

7.2. Please comment on the value of any other measures not mentioned above.

This quotes are reproduced verbatim from the answers given.

“The measures above are purely geometric. For radiotherapy it makes sense to (additionally) determine more clinically meaningful metrics, like dosimetric endpoints and editing times.”

“I believe the combination of multiple measures is meaningful”

“I did a distance error-weighted metric in a commercial product (StructSure, published in the Red Journal: Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys. 2012 Jan 1;82(1):368-78). It was very sensitive and somewhat specific, and is better for RT-based analyses because of the types of dose gradients we see in VMAT and proton therapy. However, it’s so sensitive that it freaks people out (e.g., they get low scores for what they think are accurate contours, but which have lots of errors at distance). It’d be better to use for AI and human accuracy, in my opinion, but in my experience, our industry is too thin-skinned and cowardly to do things like this correctly.”

“I think none can be used separately. The usefulness of some metrics is important for the target volume (any volume related measure + center of mass), others for contouring OARs such as APL (as correlated to editing time).”

“Robust measures should be preferred (median, 95HD) on sensitive measure (HD).”

“Mathematical sound measure, for example, Jaccard as is a metric, should normally preferred to Dice (not a metric).”
### Supplementary Material 2

#### Synthetic shapes used for evaluation

*Table 2-1: Synthetic shapes used for evaluation of the measurement tools. The principal variation was in the sampling of the contour points.*

| Shape ID | Reference shape | Test shape | Reference shape sampling | Test shape sampling |
|----------|-----------------|------------|--------------------------|-------------------|
| **Object A** | Cuboid 100 x 100 x 50 mm | Cuboid 103 x 101 x 50 mm | Sampled at in-plane corners of the cuboid | Sampled at in-plane corners of the cuboid |
| **Object B** | Cuboid 100 x 100 x 50 mm | Cuboid 103 x 101 x 50 mm | Sampled at voxel boundary intersections | Sampled at in-plane corners of the cuboid |
| **Object C** | Cuboid 100 x 100 x 50 mm | Cuboid 110 x 90 x 70 mm | Sampled at in-plane corners of the cuboid | Sampled at voxel boundary intersections |
| **Object D** | Cuboid 100 x 100 x 50 mm | Cuboid 110 x 90 x 70 mm | Sampled at voxel boundary intersections | Sampled at voxel boundary intersections |
| **Object E** | Sphere, radius 100 mm | Octahedron, $|x| + |y| + |z| = 100$ | Sampled every 15° | Sampled at voxel boundary intersections |
| **Object F** | Sphere, radius 100 mm | Octahedron, $|x| + |y| + |z| = 100$ | Sampled every 15° | Sampled at voxel boundary intersections |

These synthetic shapes, along with the clinical contour dataset are available [here](https://doi.org/10.17632/3jsdmmc3xr.2).
Supplementary Material 3

List of measures reported

Question 3.2 of the survey requested participants reported the measures they used. Table 3-1 gives the full list of measures, and the number of participants reporting each measure.

Table 3-1: Measures reported in survey question 3.2 and submitted with results in question 6.1 by participants

| Measures                              | Result submitted | Reported as available, but not submitted |
|---------------------------------------|------------------|-----------------------------------------|
| Dice Similarity Coefficient           | 13               | 0                                        |
| 3D Hausdorff Distance                 | 12               | 0                                        |
| 3D Average Distance                   | 10               | 0                                        |
| 3D 95% Hausdorff Distance             | 9                | 0                                        |
| True Positive Rate                    | 4                | 0                                        |
| Surface DSC                           | 4                | 0                                        |
| Added Path Length                     | 3                | 0                                        |
| False Positive Rate                   | 3                | 0                                        |
| Jaccard Index                         | 3                | 0                                        |
| 3D RMS Distance                       | 3                | 0                                        |
| 2D Hausdorff Distance                 | 2                | 1                                        |
| True Positive Volume                  | 2                | 0                                        |
| False Positive Volume                 | 2                | 0                                        |
| True Negative Volume                  | 2                | 0                                        |
| False Negative Volume                 | 2                | 0                                        |
| Sensitivity                           | 2                | 0                                        |
| 2D 95% Hausdorff Distance             | 2                | 0                                        |
| 2D Average Distance                   | 1                | 1                                        |
| Volume Ratio                          | 2                | 0                                        |
| Volume of Region                      | 2                | 0                                        |
| Specificity                           | 1                | 0                                        |
| 3D Median Distance                    | 0                | 1                                        |
| 2D Median Distance                    | 1                | 0                                        |
| Volume Difference                     | 1                | 0                                        |
| Centre of Mass Difference             | 1                | 0                                        |
| Centroid location                     | 1                | 0                                        |
| Length of Contour                     | 1                | 0                                        |
| Volume of Overlap                     | 1                | 0                                        |
| Missing Volume                        | 1                | 0                                        |
| Extra Volume                          | 1                | 0                                        |
| Max Difference [-] (mm)               | 1                | 0                                        |
| Max Difference [+] (mm)               | 1                | 0                                        |
| Mean Difference (mm)                  | 1                | 0                                        |
| Portion Diff Missing (%)              | 1                | 0                                        |
| Measure                              | Value 1 | Value 2 |
|-------------------------------------|---------|---------|
| Portion Diff Extra (%)              | 1       | 0       |
| Portion Diff within 1 mm (%)        | 1       | 0       |
| Portion Diff within 2 mm (%)        | 1       | 0       |
| Portion Diff within 3 mm (%)        | 1       | 0       |
| Portion Diff within 4 mm (%)        | 1       | 0       |
| Portion Diff within 5 mm (%)        | 1       | 0       |
| Portion Diff within 6 mm (%)        | 1       | 0       |
| Portion Diff within 7 mm (%)        | 1       | 0       |
| Portion Diff within 8 mm (%)        | 1       | 0       |
| Portion Diff within 9 mm (%)        | 1       | 0       |
| Portion Diff within 10 mm (%)       | 1       | 0       |
| Max Difference X [-] (mm)           | 1       | 0       |
| Max Difference X [+] (mm)           | 1       | 0       |
| Mean Difference X (mm)              | 1       | 0       |
| Portion of Diff -X (%)              | 1       | 0       |
| Portion of Diff +X (%)              | 1       | 0       |
| Max Difference Y [-] (mm)           | 1       | 0       |
| Max Difference Y [+] (mm)           | 1       | 0       |
| Mean Difference Y (mm)              | 1       | 0       |
| Mean Difference Y (mm)              | 1       | 0       |
| Portion of Diff -Y (%)              | 1       | 0       |
| Portion of Diff +Y (%)              | 1       | 0       |
| Max Difference Z [-] (mm)           | 1       | 0       |
| Max Difference Z [+] (mm)           | 1       | 0       |
| Mean Difference Z (mm)              | 1       | 0       |
| Portion of Diff -Z (%)              | 1       | 0       |
| Portion of Diff +Z (%)              | 1       | 0       |
| Signed Average Distance             | 1       | 0       |
| Recall                              | 1       | 0       |
| Precision                           | 1       | 0       |
| True Negative Rate                  | 1       | 0       |
Summary of quantitative results

Values displayed in red indicate values that are suspected as resulting from a bug in the implementation that would impact the presented value.

1. Dice Similarity Coefficient

| Dataset | Object       | 1  | 2  | 3   | 4  | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | Median |
|---------|--------------|----|----|-----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| 1       | A            | 0.9905 | 0.9810 | 0.9711 | 0.9905 | 0.9905 | 0.9711 | 0.9903 | 0.9880 | 0.9800 | 0.9711 | 0.9711 | 0.9711 | 0.9810 |
| 1       | B            | 0.9904 | 0.9810 | 0.9711 | 0.9904 | 0.9902 | 0.9711 | 0.9903 | 0.9860 | 0.9800 | 0.9711 | 0.9711 | 0.9711 | 0.9810 |
| 1       | C            | 0.7573 | 0.7520 | 0.7559 | 0.7573 | 0.7543 | 0.7573 | 0.7559 | 0.7507 | 0.7610 | 0.7540 | 0.7559 | 0.7559 | 0.7559 |
| 1       | D            | 0.7573 | 0.7520 | 0.7559 | 0.7573 | 0.7543 | 0.7573 | 0.7559 | 0.7507 | 0.7610 | 0.7540 | 0.7559 | 0.7559 | 0.7559 |
| 1       | E            | 0.4887 | 0.4870 | 0.4876 | 0.4900 | 0.4892 | 0.4887 | 0.4873 | 0.4855 | 0.4880 | 0.4870 | 0.4873 | 0.4873 | 0.4870 | 0.4875 |
| 1       | F            | 0.4856 | 0.4830 | 0.4834 | 0.4853 | 0.4849 | 0.4855 | 0.4831 | 0.4830 | 0.4840 | 0.4830 | 0.4831 | 0.4831 | 0.4833 |
| 2       | SpinalCord   | 0.5402 | 0.5100 | 0.5213 | 0.5399 | 0.5369 | 0.5369 | 0.5160 | 0.4716 | 0.5110 | 0.5110 | 0.5110 | 0.5171 | 0.5148 | 0.5171 |
| 2       | Lung_R       | 0.9593 | 0.9590 | 0.9590 | 0.9665 | 0.9595 | 0.9595 | 0.9661 | 0.9590 | 0.9590 | 0.9590 | 0.9590 | 0.9593 | 0.9593 | 0.9593 |
| 2       | Lung_L       | 0.9745 | 0.9760 | 0.9755 | 0.9700 | 0.9734 | 0.9747 | 0.9761 | 0.9758 | 0.9760 | 0.9760 | 0.9758 | 0.9759 | 0.9759 | 0.9758 |
| 2       | Heart        | 0.9516 | 0.9520 | 0.9522 | 0.9483 | 0.9522 | 0.9516 | 0.9523 | 0.9549 | 0.9520 | 0.9520 | 0.9523 | 0.9523 | 0.9521 | 0.9521 |
| 2       | Esophagus    | 0.7791 | 0.7610 | 0.7597 | 0.7795 | 0.7724 | 0.7796 | 0.7612 | 0.7290 | 0.7630 | 0.7630 | 0.7612 | 0.7612 | 0.7641 | 0.7630 |
Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy
Mark J Gooding, Djamal Boukerroui, Eliana Vasquez Osorio, René Monshouwer, Ellen Brunenberg

2. 3D Hausdorff distance

| Dataset | Object | 1     | 2     | 3      | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    | Median |
|---------|--------|-------|-------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 1       | A      | 0.9770| 1.5450| 4.1176 | 0.9770| 1.5811| 0.9770| 1.9540| 1.6135| 1.5000| 2.1846| 2.2361| 1.9916| 1.5973|        |
| 1       | B      | 1.3817| 1.5450| 4.1176 | 1.3817| 1.5811| 1.3817| 49.8366| 1.6174| 1.5100| 2.1846| 2.2361| 1.9916| 1.5993|        |
| 1       | C      | 12.0000| 11.3520| 11.1294| 11.1294| 11.1974| 11.1294| 51.0548| 12.0883| 50.9900|11.1294| 7.0711|11.7307|11.2747|        |
| 1       | D      | 12.0000| 11.3520| 11.1294| 11.1294| 11.1974| 11.1294| 11.1294| 12.0862| 50.9900|11.1294| 7.0711|11.7307|11.1634|        |
| 1       | E      | 42.0614| 42.1380| 37.9549| 42.0614| 42.2726| 42.0614| 42.0841| 43.2857| 42.1500|42.0614|30.5614|42.3487|42.0728|        |
| 1       | F      | 42.6362| 42.1380| 37.9549| 42.6362| 42.3085| 42.6362| 42.0614| 43.8761| 42.1500|42.6362|30.6757|42.3803|42.3444|        |
| 2       | SpinalCord | 8.8431| 9.0530| 27.0000| 9.0000| 8.9164| 9.0000| 10.2392| 9.4827| 10.1500| 9.0528| 9.0000| 9.0512| 9.0520|        |
| 2       | Lung_R | 26.2217| 18.8720| 16.8278| 18.9094| 18.9778| 18.3722| 28.0822| 18.2185| 30.3100|19.0601|14.1774|20.4419|18.9436|        |
| 2       | Lung_L | 13.4410| 13.4410| 16.6016| 13.4410| 13.1360| 13.4410| 21.8570| 12.4141| 16.9000|13.9288|12.6886|14.2081|13.4410|        |
| 2       | Heart  | 10.6501| 10.5260| 13.8105| 10.4238| 10.5833| 10.6501| 31.3110| 8.0804| 17.1200|10.5602| 6.7082|10.9210|10.6167|        |
| 2       | Esophagus | 18.5529| 17.8790| 15.0318| 17.6304| 17.9343| 17.6304| 17.8187| 19.0283| 17.8800|18.0051|11.3578|19.1166|17.8795|        |

3. 95% 3D Hausdorff distance

| Dataset | Object | 1     | 2     | 3      | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    | Median |
|---------|--------|-------|-------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 1       | A      | 0.9770| 4.0000| 0.9770| 1.5739| 0.9770| 1.9540| 1.5008| 1.9540| 1.4942|1.5008|1.4942|        |        |        |
| 1       | B      | 0.9770| 4.0000| 0.9770| 1.5739| 0.9770| 39.0922| 1.5007| 1.9540| 1.4942|1.5007|1.4942|        |        |        |
| 1       | C      | 10.0000| 10.0476| 10.0000|10.7099| 10.0000| 40.3538| 11.0136|10.0000|10.8576|10.0476|10.0000|        |        |        |
| 1       | D      | 10.0000| 10.0476| 10.0000| 10.7099| 10.0000| 9.3735| 11.0138|10.0000|10.8576|10.0000|10.0000|        |        |        |
| 1       | E      | 39.7947| 33.8772| 39.3242| 40.7162| 39.7947| 40.3826| 40.5953| 39.7947| 40.7864|39.7947|39.7947|        |        |        |
| 1       | F      | 40.3781| 34.4757| 40.3781| 41.1816| 40.3781| 40.9343| 41.3220| 40.3781| 41.2628|40.3781|40.3781|        |        |        |
| 2       | SpinalCord | 4.8828| 12.5059| 4.8828| 5.2280| 4.8828| 4.8828| 5.1626| 4.8828| 5.0624|4.8828|4.8828|        |        |        |
| 2       | Lung_R | 9.0000| 8.7891| 9.0000| 11.5878| 9.0000| 4.9253| 5.5773| 9.0000| 9.2611|9.0000|9.0000|        |        |        |
| 2       | Lung_L | 2.1837| 9.2457| 2.1837| 2.5747| 2.1837| 2.1837| 2.2266| 2.1837| 2.1563|2.1837|2.1837|        |        |        |
| 2       | Heart  | 5.2590| 2.9297| 5.2984| 6.1797| 5.2590| 5.7308| 4.5029| 5.2590| 5.6678|5.2590|5.2590|        |        |        |
| 2       | Esophagus | 6.0000| 8.4007| 6.0000| 5.7532| 6.0000| 4.6258| 5.4328| 6.0000| 5.0436|6.0000|6.0000|        |        |        |
## 4. Average 3D distance

| Dataset | Object | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | Median | Median excluding bugs |
|---------|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--------|------------------------|
| 1       | A      | 0.2377 | 0.5960 | 2.2519 | 0.2377 | 0.8162 | 0.2377 | 0.8762 | 0.7074 | 0.4967 | 0.6487 | 0.6223 | 0.5960 | 0.6223 |
| 1       | B      | 0.2387 | 0.5960 | 2.2519 | 0.2388 | 0.8162 | 0.2388 | 0.8774 | 0.7074 | 0.4967 | 0.6487 | 0.6223 | 0.5960 | 0.6223 |
| 1       | C      | 6.7155 | 6.6440 | 5.4251 | 6.7428 | 6.7164 | 6.8974 | 6.3647 | 7.9575 | 8.1022 | 6.7293 | 6.7424 | 6.7293 | 6.7424 |
| 1       | D      | 6.7158 | 6.6440 | 5.4207 | 6.7428 | 6.7150 | 6.8901 | 6.3647 | 7.9575 | 8.1022 | 6.7293 | 6.7424 | 6.7293 | 6.7424 |
| 1       | E      | 28.4776 | 29.2780 | 21.8825 | 28.2802 | 29.7766 | 28.4813 | 29.3807 | 29.4752 | 28.9726 | 28.7720 | 28.9726 | 28.7720 | 28.9726 |
| 1       | F      | 28.8682 | 29.6740 | 22.1356 | 28.8584 | 30.1597 | 28.8751 | 29.7829 | 28.8349 | 29.4207 | 29.1479 | 29.4207 | 29.1479 | 29.4207 |
| 2       | SpinalCord | 2.4010 | 2.5510 | 3.8294 | 2.4910 | 2.9437 | 2.4271 | 2.7839 | 2.5768 | 2.4616 | 2.5639 | 2.5510 | 2.5639 | 2.5510 |
| 2       | Lung_R | 0.9781 | 1.0220 | 2.3886 | 1.0440 | 1.7042 | 0.9602 | 1.3248 | 1.1230 | 1.4153 | 1.2239 | 1.1230 | 1.2239 | 1.1230 |
| 2       | Lung_L | 0.3711 | 0.4890 | 1.5520 | 0.4363 | 0.8341 | 0.3679 | 0.9879 | 0.3214 | 0.5252 | 0.5071 | 0.4890 | 0.5071 | 0.4890 |
| 2       | Heart  | 1.2635 | 1.3530 | 1.4665 | 1.3538 | 1.8018 | 1.2629 | 1.7910 | 1.3121 | 1.4135 | 1.3836 | 1.3538 | 1.3836 | 1.3538 |
| 2       | Esophagus | 1.0704 | 1.2550 | 2.1148 | 1.0986 | 1.6131 | 1.0686 | 1.7674 | 0.6977 | 1.8273 | 1.4340 | 1.2550 | 1.4340 | 1.2550 |

## 5. Surface DSC

| Dataset | Object | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 11   | 12   | 13   | Median |
|---------|--------|------|------|------|------|------|------|------|------|------|------|------|------|------|--------|
| 2       | TOLERANCE | 1 mm | 3mm | 1 mm | 2 mm | 1 mm |      |      |      |      |      |      |      |      | 1.0000 |
| 1       | A      | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.7419 |      |      |      |      |      |      |      |      | 1.0000 |
| 1       | B      | 0.9996 | 1.0000 | 0.9996 | 1.0000 | 0.1400 |      |      |      |      |      |      |      |      | 0.9996 |
| 1       | C      | 0.0355 | 0.0888 | 0.0355 | 0.0491 | 0.0000 |      |      |      |      |      |      |      |      | 0.0355 |
| 1       | D      | 0.0355 | 0.0889 | 0.0356 | 0.0491 | 0.0243 |      |      |      |      |      |      |      |      | 0.0356 |
| 1       | E      | 0.0007 | 0.0040 | 0.0012 | 0.0003 | 0.0004 |      |      |      |      |      |      |      |      | 0.0004 |
| 1       | F      | 0.0008 | 0.0040 | 0.0007 | 0.0002 | 0.0003 |      |      |      |      |      |      |      |      | 0.0003 |
| 2       | SpinalCord | 0.1532 | 0.7378 | 0.1520 | 0.1786 | 0.0640 |      |      |      |      |      |      |      |      | 0.1520 |
| 2       | Lung_R | 0.8756 | 0.9235 | 0.8631 | 0.8658 | 0.9180 |      |      |      |      |      |      |      |      | 0.8658 |
| 2       | Lung_L | 0.9304 | 0.9775 | 0.9153 | 0.9436 | 0.9490 |      |      |      |      |      |      |      |      | 0.9436 |
| 2       | Heart  | 0.6426 | 0.8959 | 0.6079 | 0.6917 | 0.5206 |      |      |      |      |      |      |      |      | 0.6079 |
| 2       | Esophagus | 0.7207 | 0.9164 | 0.7211 | 0.7533 | 0.6130 |      |      |      |      |      |      |      |      | 0.7211 |
# 6. Added Path Length

| Dataset | Object | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | Median |
|---------|--------|---|---|---|---|---|---|---|---|---|----|----|----|----|--------|
| TOLERANCE | 1 mm | 1 voxel | 1 mm | |
| 1 | A | 4950.0 | 610.6 | 4950.0 | 4950.0 |
| 1 | B | 4950.0 | 317.5 | 4950.0 | 4950.0 |
| 1 | C | 9800.0 | 14088.3 | 9800.0 | 9800.0 |
| 1 | D | 9789.9 | 13990.6 | 9789.9 | 9789.9 |
| 1 | E | 49222.5 | 18060.8 | 49222.5 | 49222.5 |
| 1 | F | 49362.9 | 37612.5 | 49362.9 | 49362.9 |
| 2 | SpinalCord | 5183.3 | 3636.7 | 5183.3 | 5183.3 |
| 2 | Lung_R | 5177.9 | 20244.1 | 5177.9 | 5177.9 |
| 2 | Lung_L | 3423.2 | 16346.7 | 3423.2 | 3423.2 |
| 2 | Heart | 6249.6 | 9887.7 | 6249.6 | 6249.6 |
| 2 | Esophagus | 1417.4 | 3341.8 | 1417.4 | 1417.4 |
Understanding impact of rasterization method on Dice similarity coefficient

The main factor appearing to affect the Dice Similarity coefficient was how software interpreted the RTSS polygon when rastering to a voxel array. While the self-reported methods of conversion did not correlate with the scoring, many of the values could be reproduced by varying this choice. The code snippet below, provided by one participant in their detailed information, converts RTSS point locations in real-world coordinates to voxel-based coordinates. The scikit python library is then used to raster this polygon into a voxel mask, based on whether the center of each voxel is included in the polygon.

Code snippet 1:

```python
# conversion of real-world co-ordinates to in-plane voxel co-ordinates
rows = ((rtss_coords[:, 1] - pos_r) / spacing_r)
cols = ((rtss_coords[:, 0] - pos_c) / spacing_c)
# rastering polygon performed with python library Scikit
rr, cc = skimage.draw.polygon(rows, cols)
voxel_mask_data[cc, rr, z_index] = class_id
```

This code was modified to investigate the impact of the rasterization method. First, the code was modified to round the voxel coordinates of the polygon to integer values, to evaluate the variation in implementation that can occur when converting between real-world and voxel coordinates. Second, the real-world polygon was jittered in all directions by half a voxel in-plane. Voxels that were included in every jittered polygon (without rounding of voxel coordinates) were taken as those where the whole voxel would be included in the original polygon. Voxels that were included in any jittered polygon were taken as those where any of the voxel would be included in the original polygon.

Table 5-2 shows the impact that different rasterization approaches have on the DSC, where all other aspects of the code (data loading, DSC calculation etc) are kept consistent, together with the range of values submitted by participants. This indicates that the rasterization process is the greatest source of variation in calculation of the DSC, and plausibly explains the participants submitted values.

| Object     | Participants’ maximum DSC | Participants’ minimum DSC | Center of a voxel using RTSS points in the real-world coordinate system | Center of a voxel with rounding RTSS points to the voxel coordinate system | Any part of a voxel | The whole of a voxel |
|------------|---------------------------|---------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------|--------------------|----------------------|
| SpinalCord | 0.540                     | 0.472                     | 0.521                                                                    | 0.540                                                                    | 0.545              | 0.460                |
| Lung_R     | 0.966                     | 0.955                     | 0.959                                                                    | 0.955                                                                    | 0.960              | 0.957                |
| Lung_L     | 0.976                     | 0.970                     | 0.976                                                                    | 0.970                                                                    | 0.975              | 0.975                |
| Heart      | 0.955                     | 0.948                     | 0.952                                                                    | 0.951                                                                    | 0.955              | 0.951                |
| Esophagus  | 0.780                     | 0.760                     | 0.760                                                                    | 0.780                                                                    | 0.778              | 0.743                |
Supplementary Material 6

Impact of sampling on distance measurement statistics

1. Use of binary masks

For the sake of simplicity, the impact of binary masks representation on the Average Distance is analyzed in 2D. Consider a reference shape that is a square 9.99 x 9.99 mm and a test shape, also square of 10.01 x 10.01 mm. These shapes have the same centroid such that there is a 0.02 mm gap between them on all faces. These shapes are placed on a pixel grid on 1 mm isotropic spacing, such that the centers of the pixels fall in the 0.02 mm gap between the two pixels. Although the average distance between the two shape is 0.02 mm (ignoring the corners), the pixel-based measurement is 1 pixel, or 1 mm, as illustrated in Figure 6-1.

![Figure 6-1: Impact of mask-based quantization on distance statistics (Left) Actual RTSS on pixel grid, with exaggerated gap for visibility, has an average distance of 0.02 mm (Right) Resulting quantized mask with 1-pixel error on a 1×1mm pixel grid has an average distance of about 1mm.](image)

2. Non-isotropic volumes

The previous example is extended to a non-isotropic volume, as shown in Figure 6-2, with 2mm pixel resolution in one direction. In this new scenario, the error in the coarsest direction is 2 mm, the pixel resolution. When sampling the boundary once per voxel, this is proportionally more densely sampled, with 9 pixels across (again ignoring corners) compared to 4 pixels down. Thus, the average distance would be approximately 1.7mm.

![Figure 6-2: Impact of resolution. (Left) Actual RTSS on pixel grid, with exaggerated gap for visibility, has an average distance of 0.02 mm (Right) Resulting quantized mask with 1-pixel error on a 1×2 mm pixel grid has an average distance of about 1.7 mm, since the 2 mm vertical error is more densely sampled than the 1 mm horizontal error.](image)