Physics and Imaging in Radiation Oncology 24 (2022) 152–158

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy

Mark J. Gooding [a], Djamal Boukerroui [a], Eliana Vasquez Osorio [b], René Monshouwer [c], Ellen Brunenberg [c,*]

[a] *Mirada Medical Ltd, Oxford, United Kingdom*
[b] *University of Manchester, Manchester, United Kingdom*
[c] *Radboud University Medical Centre, Nijmegen, the Netherlands*

ARTICLE INFO

Keywords:
Radiotherapy
Contouring
Similarity measure
Dice Similarity Coefficient
Hausdorff Distance
Quality Assurance

ABSTRACT

Background and Purpose: A wide range of quantitative measures are available to facilitate clinical implementation of auto-contouring software, on-going Quality Assurance (QA) and interobserver contouring variation studies. This study aimed to assess the variation in output when applying different implementations of the measures to the same data in order to investigate how consistently such measures are defined and implemented in radiation oncology.

Materials and Methods: A survey was conducted to assess if there were any differences in definitions of contouring measures or their implementations that would lead to variation in reported results between institutions. This took two forms: a set of computed tomography (CT) image data with "Test" and "Reference" contours was distributed for participants to process using their preferred tools and report results, and a questionnaire regarding the definition of measures and their implementation was completed by the participants.

Results: Thirteen participants completed the survey and submitted results, with one commercial and twelve in-house solutions represented. Excluding outliers, variations of up to 50% in Dice Similarity Coefficient (DSC), 50% in 3D Hausdorff Distance (HD), and 200% in Average Distance (AD) were observed between the participant submitted results. Collaborative investigation with participants revealed a large number of bugs in implementation, confounding the understanding of intentional implementation choices.

Conclusion: Care must be taken when comparing quantitative results between different studies. There is a need for a dataset with clearly defined measures and ground truth for validation of such tools prior to their use.

1. Introduction

Contouring of both targets and organs-at-risk (OARs) plays an important role in radiotherapy planning. Consequently, the study of accuracy of such contouring is relatively commonplace, whether for assessing inter-observer variation [1,2] or the accuracy of auto-contouring [3,4], and quantitative measures of assessing contouring are required for such studies [5,6].

During the last few years, auto-contouring of OARs has evolved from a "nice-to-have" to a "must-have" for radiotherapy clinics. The development of deep learning-based algorithms, performing much better than previous methods such as atlas-based auto-segmentation, has pushed this technology towards the clinic [4]. In order to facilitate clinical implementation of auto-contouring software, a lot of recent work has been done on commissioning and pre-deployment quality assurance (QA) of such systems [7]. Possible tests comprise qualitative and quantitative measurements, where the latter can be subdivided into geometric, dosimetric, and time-saving measures [6,8], amongst others. Such approaches remain useful for on-going QA of auto-contouring post-deployment [9]. These measures are also relevant for studying inter-observer variation to investigate contouring differences (e.g. [10]) and help define guidelines (e.g. [11,12]).

While geometric measures such as the Dice Similarity Coefficient (DSC) [13] or Hausdorff Distance (HD) and Average Distance (AD) [14,15] might seem very simple, objective and deterministic, caution needs to be exercised [8]. The first limitation of these measures is that they are defined with respect to a "ground truth", most often manual contouring done by an expert clinical observer. This paradigm does not

* Corresponding author.
E-mail address: Ellen.Brunenberg@radboudumc.nl (E. Brunenberg).

https://doi.org/10.1016/j.phro.2022.11.009
Received 26 September 2022; Received in revised form 10 November 2022; Accepted 11 November 2022
Available online 15 November 2022

take into account intra- and inter-observer variation, and thus the possibility that for some instances, the test contour might be of higher quality than the reference.

In addition, geometric measures of overlap and distance are not necessarily clinically meaningful [5,6,16]. To assess the efficiency of auto-contouring, measures developed more recently such as Added Path Length (APL) [17] and Surface Dice [18] might be valuable because they have been shown to correlate with time-saving. Dosimetric evaluation can also add to clinical interpretation (e.g. [19]), although attention should be paid to the most relevant combination of contours and plans to examine [8].

Finally, if geometric measures are the desired evaluation method, it is important to use them in a consistent and accurate way. Results can be very sensitive to different choices made when implementing the measures, for instance whether calculations are performed in 2D or 3D, and what kind of contour representation is used. Furthermore, multiple definitions exist for some measures.

In this study, possible differences in software used in radiotherapy research and by radiotherapy institutes were investigated, using a questionnaire and benchmarking dataset, building on a provisional survey conducted prior to the ESTRO 2022 conference [20], to establish the level of variation existing in the field of radiotherapy when calculating measures for comparing contours.

2. Material and methods

A survey was conducted to assess what contouring measures were used and if there were any differences in their implementation that lead to variation in results reported between institutions. This survey took two forms: (1) a questionnaire regarding the definition of metrics and their implementation was completed by each of the participants, and (2) a set of CT image data with “Test” and “Reference” contours was distributed for participants to process and report results using their preferred tools.

Participation was encouraged through personal invitation (n = 14), word-of-mouth, and informal promotion within appropriate society networks, such as the Dutch medical physics, deep learning in radiotherapy communities and the ESTRO physics workshop groups. The survey was further advertised at ESTRO 2022 during a presentation of the results of a preliminary survey [20]. The survey was open to participation for six months from 30 Jan 2022 to 30 Jun 2022.

2.1. Questionnaire

The questionnaire focused on expected areas of variation in definition and implementation of measures following a preliminary study that considered the results of 6 participants from 6 different centers [20]. The questionnaire composed 6 sections; (1) the participant, institution, and software, (2) high-level implementation and internal shape representation of the structures, (3) the range of measurements available, (4) implementation of shape representation conversion and tolerances, (5) distance measure definitions and implementation, (6) results upload and additional information.

This questionnaire was intended to facilitate the analysis of whether reported differences in definitions and analysis existed within radiation oncology and to assist in understanding of the quantitative results submitted, rather than for performing statistical analysis (e.g. the prevalence of each measure). The full questionnaire is provided in Supplementary Material 1.

2.2. Dataset

The dataset consisted of two studies: a synthetic set with analytical geometric shapes and a clinical set with real anatomical shapes. Both studies were in DICOM format, with contours stored as Radiotherapy Structure Sets (RTSS), the format used in clinical practice.

The first study set contained a synthetic CT image of 0.977 mm × 0.977 mm in-plane resolution and 2 mm slice spacing, with 126 slices of 512 × 512 voxels. The synthetic CT was filled with random uniform noise from −1024 to 2975 HU, since the comparison of the contours should not depend on the image content. The analytical shapes consisted of cuboids (4 pairs, shapes A-D), spheres and octahedrons (2 pairs, shapes E and F). The main variation in the data was in how the control points of the polygons were sampled in the RTSS. The table in Supplementary Material 2 provides a detailed description of the shapes defined. The geometric nature of the shapes was intended to simplify the interpretation of any definition and implementation differences between contouring assessment tools. The synthetic data was created in python (v3.8.10), writing DICOM with the pydicom (v2.1.2) library.

The second study was taken from the AAPM 2017 Thoracic Auto-Segmentation Challenge dataset [3,21], case LTCSC-Test-S1-201. This data consists of a thoracic CT image, in radiotherapy treatment position, with reference contours for lungs, esophagus, heart and spinal cord. The “ground truth” contours available with this data were used as “Reference”, and contours from a deep learning-based auto-contouring system (DLCExpert with model Thorax_CT_NL007_MO, Workflow Box 2.6, Mirada Medical Ltd, Oxford, UK) were used as the “Test”. These shapes were intended to evaluate the impact that any variation in definition or implementation of measurements could have in a scenario of clinical commissioning or assessment of auto-contouring. Note, the purpose was not to assess the accuracy of the auto-contouring system.

The dataset has been made publicly available [22].

2.3. Analysis and participant follow-up

The quantitative results submitted, based on the common dataset, were reviewed by plotting variation against the median result for each measure and structure. The median result was chosen since the survey was conducted to determine the presence and causes of variation in results, rather than to assess tools against a predefined “right” answer. Analysis of the reported values was only made for measures for which five or more participants submitted results. The causes of variations in results were investigated by considering the questionnaire responses. Follow-up discussions were held with participants individually, during the analysis of the results, to fully understand any differences in implementation observed in the results that could not be explained by the questionnaire answers.

3. Results

Twenty-one individuals from 18 institutions registered interest in participation. Of these, 13 participants from ten different institutions subsequently completed the full survey and submitted results. Of the 14 personal invitations to participate, 11 registered interest and six participated. One registered individual indicated that they felt unable to complete the questionnaire on account of insufficient knowledge of the methods used in implementation.

3.1. Questionnaire findings

The full results of the questionnaire are available in the Supplementary Material 1.

The majority of participants, 12 out of 13, reported using “homemade” or open-source software or a combination of the two. Only one participant reported using commercial software. Although seven of the participants reported academic publications relating to the measures [8,17,18] or their use [23–25], no participants reported studies which actually demonstrated the accuracy of the implementation used.

A range of choices in implementation design were found, as shown in Fig. 1. Three different methods of conversion from RTSS to a voxel mask were used by ten participants, and a further two responses expressed uncertainty in the implementation details. Seven different approaches

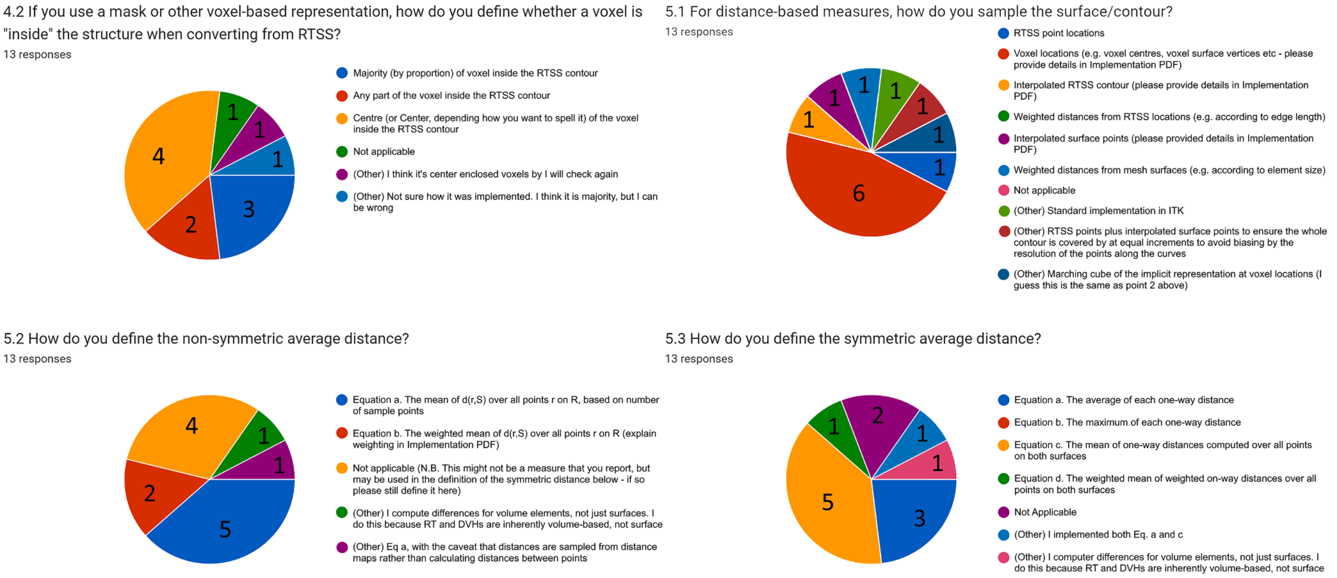

Fig. 1. Survey questions showing the greatest variation in implementation details (4.2, 5.1) and measurement definitions (5.2, 5.3). Note for question 4.2 the use of “I think”.

were taken to the method of sampling the surface when calculating distance measures. Four definitions of the one-way average distance were used and five definitions for symmetric average distance were reported. Several participants made use of the “other” option when filling out responses to implementation choice questions, including as a method of expressing uncertainty as to their implementation’s details.

Of the twelve participants using a voxel-based representation for at least one measure, all but two used the CT image resolution as the basis for the voxel grid of the mask. One participant calculated the measures on an arbitrarily chosen resolution of 0.5 mm isotropic volume, and a second used 1 mm isotropic sampling.

3.2. *Comparison of quantitative results*

Twenty different measures were reported as available by more than one respondent, with the frequency of each measure shown in Fig. 2. A further 47 different measures were reported by only one respondent. The full list of measures is available in the Supplementary Material 3. Only four measures, Dice Similarity Coefficient (DSC), 3D Hausdorff Distance (HD), 3D 95% Hausdorff Distance (95% HD), Average Distance (AD) (sometimes reported as “Mean Distance to Agreement”), were reported by more than five participants. While five results were reported for Surface DSC by four participants (one participant submitted results at two different tolerances), the different acceptance tolerances used meant that comparison was not performed. A spreadsheet of submitted

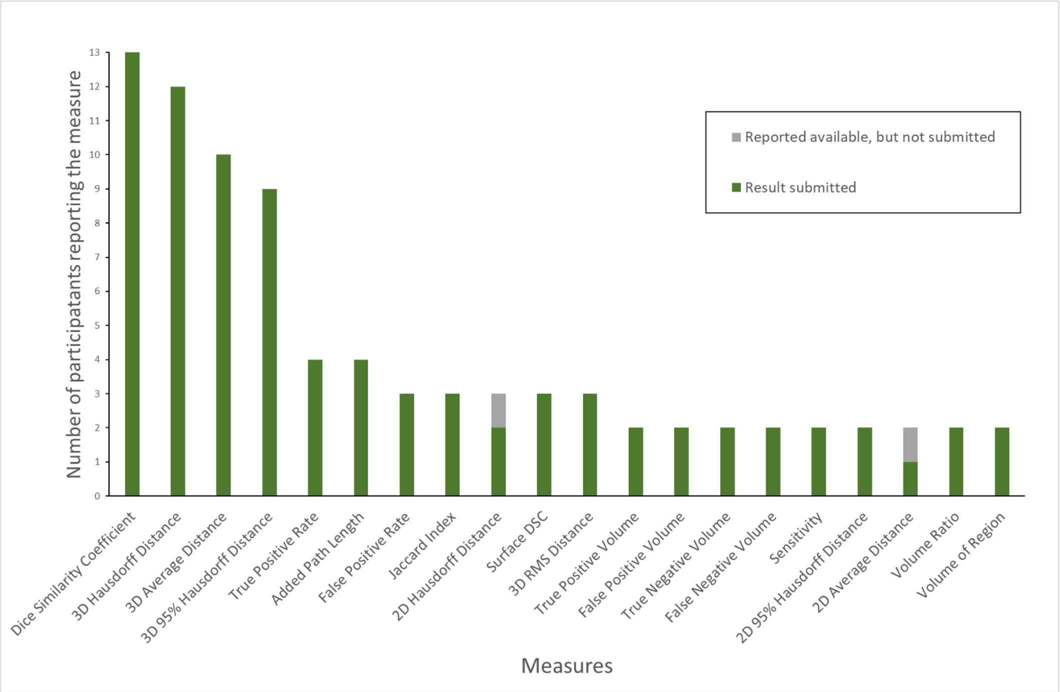

Fig. 2. Frequency of measures reported as available by participants. A further 47 measures were reported by only one participant.

values is available in Supplementary Material 4.

On the synthetic data, minor variations in DSC results, of around 1% of the median, were observed. However, as can be seen in Fig. 3, larger variation can be observed in the real data for the long thin structures of the spinal cord and esophagus, with a maximum variation of −9% with respect to the median result being seen for the spinal cord. All but two respondents reported using a voxel mask representation. However, no clear correlation was observed between the reported method of voxelization and the variation in DSC.

The 3D HD was reported by all but one of the respondents. Wide variation was observed in the reported HD values, with several outliers. The largest outlier was over 3000% greater than the median, occurred for synthetic shape B, with the same implementation having a large outlier of around 350% greater than the median for synthetic shape C. A further implementation also reported large values of around 350% greater than the median for synthetic shapes C & D. Fig. 4 shows the range of HD reported as a percentage of the median values. Two subfigures vary the range of the y-axis to illustrate the range of variation given the significant outliers found. The following-up with participants found a number of 'bugs' in implementation. These are indicated in the figure by circles. Notwithstanding implementation bugs, the observed variation was up to 40%, as shown in the lower plot of Fig. 4.

Participants where a bug was found in the analysis of HD were excluded from the analysis of 95% HD if the bug was likely to impact the accuracy of the result. This is reflected in the upper plot of Fig. 5, with the maximum variation from the median being about 40%. The variation in the AD, excluding likely bugs, is shown in the lower plot of Fig. 5. For the synthetic shapes A and B, variation of up to 50%, is seen. Larger percentage variation is observed in the clinical structures, with variations larger than 50% with respect to the median value for three participants.

3.3. Participant follow-up

Investigation of outliers was performed with seven of the 13 participants. One participant had submitted full code, so investigation could be performed without direct follow-up discussion. Bugs impacting DSC, HD, 95% HD and AD were found for one, five, two, one implementations respectively.

4. Discussion

The results show that there can be substantial variation in values reported for the most commonly used measures. It would be reasonable to assume that such variation also extends to those less commonly reported. This variation has implications both for the comparison of values reported in different studies where different implementations may have been used to assess contouring performance, but also in the clinic for commissioning an auto-contouring system, for example. That there are differences is concerning, but it is necessary to understand from where these differences arose to be able to address this challenge.

A major part of the questionnaire focused on the definition of measurements, as a result of prior observation that there is some variation. For instance, the open-source Plastimatch [26] defines the symmetric 95% HD as the mean of one-way 95% HDs, whereas the open-source EvaluateSegmentation [14] defines it as the maximum of the one-way 95% HDs. Further variation was seen within this survey, with some participants calculating the percentile or mean over the set one-way distances and then combining these results by taking the maximum or mean respectively, while other participants performed calculations after combining two sets of one-way distances into a single set. Although, these different definitions did not have a major influence on the values reported in this study, their impact will be larger in circumstances where the sets of one-way differences for the reference and test shapes have different sizes and there are differences in the tails of the distributions of one-way distances.

In addition to the variation in definitions, a range of implementation choices were evident. The definition of the DSC was consistent between all participants, yet some variation is still observed in the results. The main factor that could affect the DSC was how implementations

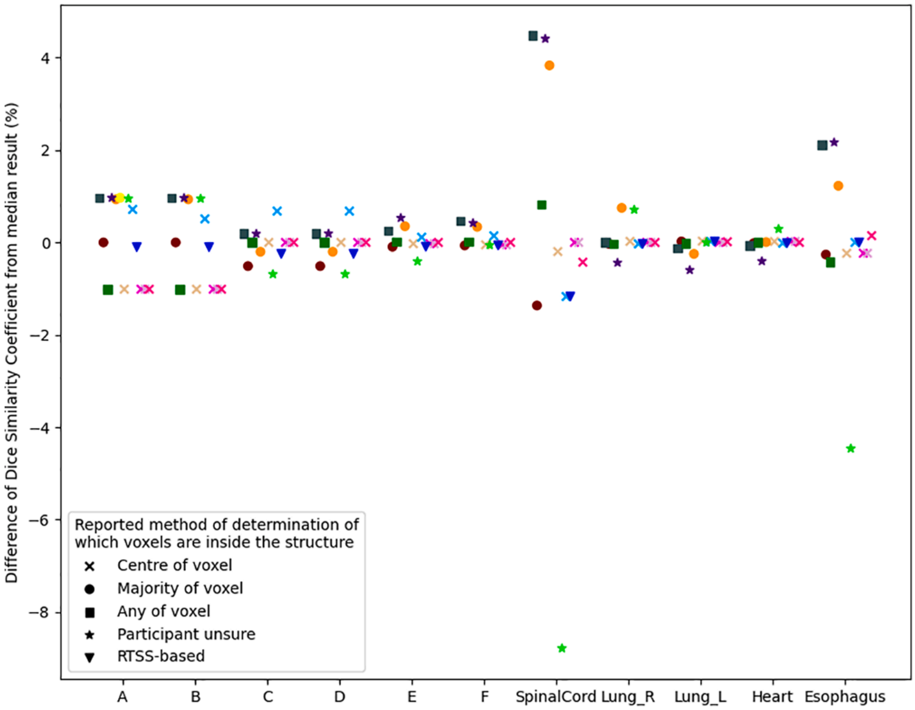

Fig. 3. Variation in Dice Similarity Coefficient as a percentage of the median value of participants' submissions for each structure. The largest variation appears for long thin structures. The different point shapes represent the different method of conversion of the RTSS to a voxel-array (or not), prior to calculation of DSC, as reported by the participants. Corresponding colors in Figures 2-4 indicate the same participant.

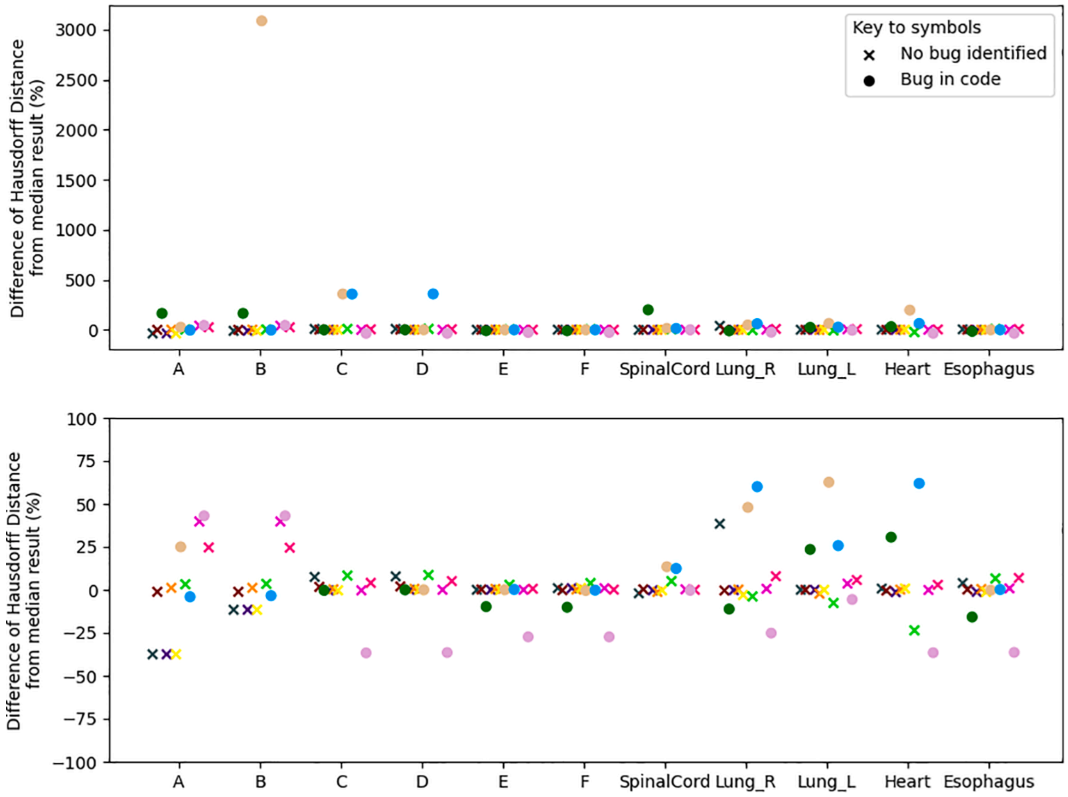

Fig. 4. Variation in Hausdorff Distance as a percentage of the median Hausdorff Distance reported. (Upper) Full range of reported values, dominated by one outlier. (Lower) Rescaled y-axis to remove outliers >100%. Circles indicate implementations where bugs were discovered during the analysis and through discussion with the participant. Corresponding colors in Figures 2-4 indicate the same participant.

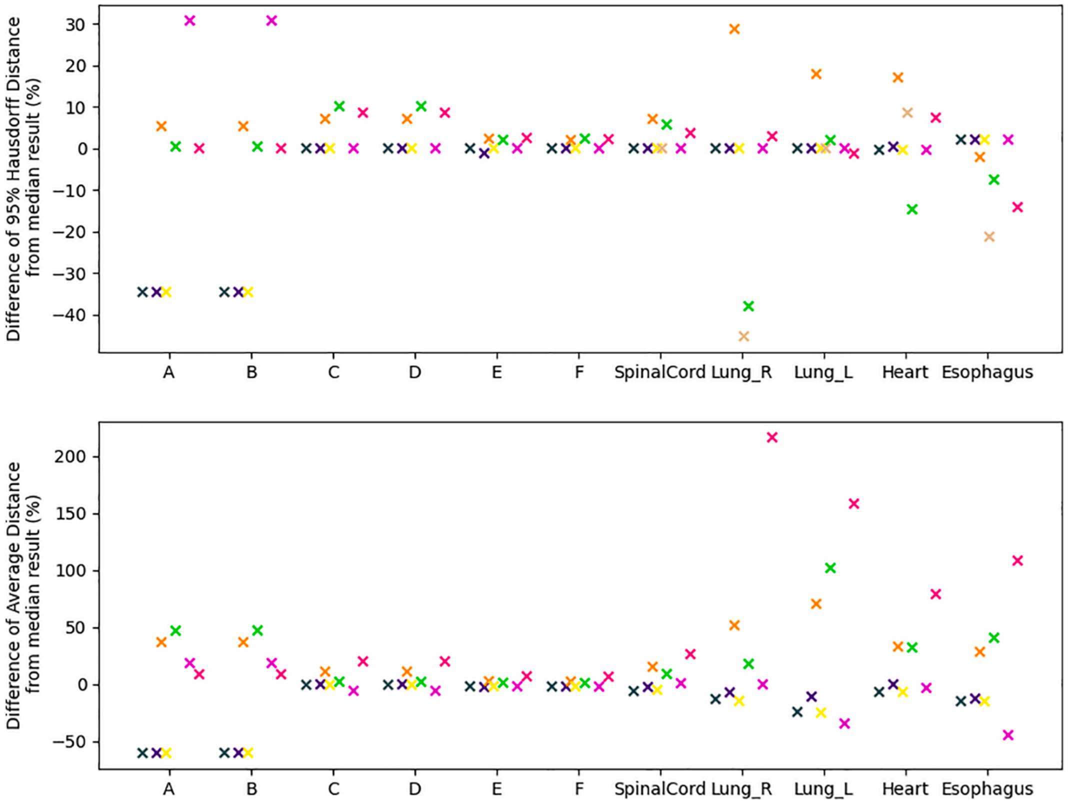

Fig. 5. Variation in distance measures from the median results. Results of participants where a bug was found in the 100% Hausdorff Distance have been excluded if the bug was likely to impact the accuracy of the plotted distance result. (Upper) Variation in 95% Hausdorff Distance as a percentage of the median of results. (Lower) Variation in Average Distance as a percentage of the median result reported. Corresponding colors in Figures 2-4 indicate the same participant.

interpreted the RTSS polygon when rasterizing to a voxel array. While the self-reported methods of conversion did not correlate with the scoring, as observed in Fig. 3, many of the values could be reproduced by varying this choice and by applying rounding when converting from real world coordinates to voxel indices (see Supplementary Material 5 for details). The lack of apparent correspondence between results reported and the method reported by the participant, may indicate the participants were not fully aware of how the libraries they used operated.

For distance-based measures the implementation choices expand considerably. 3D distance calculations require converting to a 3D shape representation, either voxel-based (whether a binary mask or implicit function) or mesh-based, where creation of a voxel-representation was a prior step in creating a mesh representation, thus inheriting any inaccuracy in the rasterization process. Calculation of sets of distances from these can be performed in a number of ways; directly from mesh vertices or voxel locations or via distance transforms from either the meshes or the voxel arrays. Distances must then be sampled to the other surface, leading to choices on the density and method of sampling. The majority of the residual variation in HD observed, after bugs were excluded, can be explained by variations induced by the choice of voxelization method, suggesting other factors have negligible impact on this single extreme value. However, the quantization induced by rasterization to a binary mask, together with choice of surface sampling method, can have a large impact on AD and 95% HD. When using a binary mask-based representation, the quantization can result in up to one voxel error (see Supplementary Material 6 for details).

The investigation of outliers revealed bugs in five of the 13 implementations. Bugs related to an incorrect assumption of dense RTSS control point spacing, failure to correctly account for image resolution (during rasterization), and incorrect rasterization of the surface. While medical device regulations should ensure adequate testing of clinical software, higher risk exists where home-made tools are being used for research. Importantly, accuracy of such tools must be ensured where they are used in the commissioning of medical devices.

Only one participant used commercial software and therefore no assessment can be made as to the degree of variation in definitions or implementations of contouring measure between commercial products. However, these findings still reveal the challenges and problems for assessing contouring with off-the-shelf tooling. It was notable that one participant felt unable to complete the survey as they were not aware of the implementation choices in the application used and several of the responses also indicated uncertainty about the exact behavior of the software used. The lack of detailed implementation information available for commercial software may have been a cause of the under-representation of such systems in this survey.

The informal invitation and voluntary participation approach taken for this study is a limitation and may therefore not reflect the wider experience with quantitative measures in the radiotherapy field. This self-selecting nature may have biased the results to those who have implemented the measures for themselves (as reflected in the survey) and those who are particularly interested in the evaluation of contouring from the perspective of contouring research publication or development and commissioning of auto-contouring.

This study has highlighted the variation and potential inaccuracy of implementation in common quantitative measurements for assessing contouring in radiotherapy. While measures should be robust to implementation with precise definition, this should not form the basis of choice of measurements used; rather measure should be selected on the basis of their ability to provide useful information. Variation in definition for measures means that care must be taken in reporting the results of studies evaluating contouring. Such studies should clearly state the definition of any measurement used. However, even where definitions agree, variation in algorithmic implementation choices between different tools can result in substantial variations in the output measurements. Therefore, care must be taken when comparing results from different studies – even where the same approach appears to have been used. Furthermore, this study helped in identifying bugs in implementations having an impact on the quantitative measurements. This points to a need for careful testing of such tools. Provision of a DICOM dataset with known ground truth answers would assist in the validation of such tools and ensure greater consistency when assessing contouring.

Declaration of Competing Interest

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.

Acknowledgements

We thank the participants for their efforts in running the evaluation of the shapes and completing the survey. Participants (excluding authors) were: Femke Vaassen, Maastro Clinic; Rita Simões, NKI; Jinzhong Yang, MD Anderson Cancer Centre; Prerak Mody, Leiden University Medical Center; Sara Poeta, Institut Jules Bordet; Adam Szmul, University College London; Benjamin Nelms, Canus Lupus; Edward Henderson, University of Manchester; Poppy Nikou, University College London. We also thank Jan Derks van de Ven for his work on the preliminary survey and abstract that led to this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2022.11.009.

References

[1] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imaging Radiat Oncol 2016;60:393–406. https://doi.org/10.1111/1754-9485.12462.

[2] Cacicedo J, Navarro-Martin A, Gonzalez-Larragan S, De Bari B, Salem A, Dahele M. Systematic review of educational interventions to improve contouring in radiotherapy. Radiother Oncol 2020;144:86–92. https://doi.org/10.1016/j.radonc.2019.11.004.

[3] Yang J, Veeraraghavan H, Armato SG, Farahani K, Kirby JS, Kalpathy-Kramer J, et al. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med Phys 2018;45:4568–81. https://doi.org/10.1002/mp.13141.

[4] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. Semin Radiat Oncol 2019;29:185–97. https://doi.org/10.1016/j.semradonc.2019.02.001.

[5] Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlentherapie Und Onkol 2012;188:160–7. https://doi.org/10.1007/s00066-011-0027-6.

[6] Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol 2021;160:185–91. https://doi.org/10.1016/j.radonc.2021.05.003.

[7] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.

[8] Gooding MJ. On the Evaluation of Auto-Contouring in Radiotherapy. In: Yang J, Sharp GC, Gooding MJ, editors. Auto-Segmentation Radiat. Oncol., CRC Press; 2021, p. 217–42.

[9] Claessens M, Oria CS, Brouwer C, Ziemer BP, Scholey JE, Lin H, et al. Quality Assurance for AI-Based Applications in Radiation Therapy. Semin Radiat Oncol 2022. https://doi.org/10.1016/j.semradonc.2022.06.011.

[10] Brouwer CL, Steenbakkers RJHM, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D Variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7:32. https://doi.org/10.1186/1748-717X-7-32.

[11] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, 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. https://doi.org/10.1016/j.radonc.2015.07.041.

[12] Duane F, Aznar MC, Bartlett F, Cutter DJ, Darby SC, Jagsi R, et al. A cardiac contouring atlas for radiotherapy. Radiother Oncol 2017;122:416–22. https://doi.org/10.1016/j.radonc.2017.01.008.

[13] Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.

[14] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 2015:15. https://doi.org/10.1186/s12880-015-0068-x.

[15] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993;15:850–63. https://doi.org/10.1109/34.232073.

[16] Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. J Med Imaging Radiat Oncol 2010;54:401–10. https://doi.org/10.1111/j.1754-9485.2010.02192.x.

[17] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.

[18] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, de Fauw J, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. J Med Internet Res 2021:23. https://doi.org/10.2196/26151.

[19] Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiother Oncol 2011;98:373–7. https://doi.org/10.1016/j.radonc.2010.11.017.

[20] Brunenberg E, Derks van de Ven J, Gooding MJ, Boukerroui D, Gan Y, Henderson E, et al. PD-0064 Multicenter comparison of measures for quantitative evaluation of automatic contouring. Radiother Oncol 2022;170:S37–8. https://doi.org/10.1016/S0167-8140(22)02734-7.

[21] Yang J, Veeraraghavan H, van Elmpt W, Dekker A, Gooding M, Sharp G. CT images with expert manual contours of thoracic cancer for benchmarking auto-segmentation accuracy. Med Phys 2020;47:3250–5. https://doi.org/10.1002/mp.14107.

[22] Gooding MJ, Derks van de Ven J, Boukerroui D, Brunenberg E. Dataset for multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy. Mendeley Data 2022. https://doi.org/10.17632/3jsdmmc3xr.2.

[23] Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 2009;28:1251–65. https://doi.org/10.1109/TMI.2009.2013851.

[24] Zhou R, Liao Z, Pan T, Milgrom SA, Pinnix CC, Shi A, et al. Cardiac atlas development and validation for automatic segmentation of cardiac substructures. Radiother Oncol 2017;122:66–71. https://doi.org/10.1016/j.radonc.2016.11.016.

[25] Huang K, Rhee DJ, Ger R, Layman R, Yang J, Cardenas CE, et al. Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. J Appl Clin Med Phys 2021;22:168–74. https://doi.org/10.1002/acm2.13207.

[26] Sharp GC, Li R, Wolfgang J, Chen GTY, Peroni M, Spadea MF, et al. Plastimatch – An Open Source Software Suite for Radiotherapy Image Processing. Proc. XVI'th Int. Conf. use Comput. Radiother., Amsterdam, NL. 2010.