Victor Chukwu / Bio.Med.Pharm.Rev. 1(1) (2025) 10-17
https://doi.org/10.62587/BMPR.1.1.2025.10-17 ISSN: 3136-5248

Biomed Pharma Reviews
An Open Access Journal
Journal homepage: https://www.biomedpharmareviews.com

Review Article **Open Access**

Scoring above the Noise Floor: Overlap Metrics, Observer Variability and Dosimetric Impact in Deep Learning Auto-Segmentation for Radiotherapy

Victor Chukwu[1*]

[1]Independent Researcher, Lagos, Nigeria. E-mail: victor.chukwu210@gmail.com

Article Info

Volume 1, Issue 1, January 2025
Received : 13 September 2024
Accepted : 19 December 2024
Published : 25 January 2025
doi: 10.62587/BMPR.1.1.2025.10-17

Abstract

Background: Deep learning auto-segmentation is entering routine radiotherapy planning, and its performance is reported almost universally as a geometric overlap statistic. Whether that statistic establishes clinical readiness depends on how it relates to the reproducibility of the reference standard and to dose delivered. **Methods:** Systematic reviews and clinical evaluations of deep learning auto-segmentation in radiotherapy reporting evaluation methodology, overlap statistics or dosimetric outcomes were eligible. Reported metric usage was extracted. The fraction of a structure lying within a boundary shell of given thickness was derived analytically to quantify the sensitivity of volume-overlap statistics to boundary error. No pooling was performed. Analyses were performed in Python 3. **Results:** Across 117 reviewed studies, a geometric metric was used in 116 (99.1%) and the Dice similarity coefficient specifically in 113 (96.6%). Dosimetric assessment appeared in 27 (23.1%), qualitative or physician rating in 22 (18.8%), time saving in 18 (15.4%) and editing time in 11 (9.4%). Over 90 different names were used for geometric measures. A single manual contour served as ground truth in 65 studies (55.6%), and only 31 (26.5%) compared auto-contours with inter- or intra-observer variation, meaning 74% did not establish the reproducibility of their own reference standard. Reported expert reproducibility for a delineation task was a Dice of approximately 0.80 within observers and 0.67 between them; a deep learning model scored 0.76 against intra-observer 0.77 ($p = 0.307$) and 0.78 against inter-observer 0.67. In a commercial software evaluation with a mean Dice of 0.72, described as acceptable in clinical practice, the volume receiving 45 Gy in a rectal case was 10.4 cm^3 by manual contour and 289.4 cm^3 by automatic contour, a 27.8-fold difference. **Conclusions.** The metric that 97% of studies report is one the field states is not predictable of dosimetric impact, is measured against a reference standard whose own reproducibility is usually unestablished, and is structurally insensitive to boundary error because a 2 mm shell constitutes under a fifth of the volume of a 30 mm structure. Scores exceeding inter-observer agreement cannot be read as accuracy. Evaluation should pair overlap statistics with dosimetric assessment and physician rating, both of which appear in fewer than a quarter of published studies.

Keywords: *Auto-segmentation, Deep learning, Dice similarity coefficient, Inter-observer variability, Radiotherapy planning, Dosimetric evaluation*

1. Introduction

Delineating target volumes and organs at risk is among the most time-consuming steps in radiotherapy

* *Corresponding author: Victor Chukwu, Independent Researcher, Lagos, Nigeria. E-mail: victor.chukwu210@gmail.com*

planning and among the most variable. Deep learning auto-segmentation addresses both, and commercial implementations are now in routine clinical use.

Performance is reported, with striking consistency, as a geometric overlap statistic between the automatic contour and a manual reference—most often the Dice similarity coefficient (Dice, 1945). A value above some threshold, commonly 0.8, is described as clinically acceptable.

Three assumptions are embedded in that practice and each is checkable. The first is that the manual reference is a stable target: if two experts contouring the same structure agree only moderately with each other, then agreement with any one of them is bounded by that variability. The second is that geometric agreement predicts clinical consequence—that a contour scoring well will deliver dose similarly to the reference. The third is that a volume-overlap statistic is sensitive to the errors that matter, which in radiotherapy occur at boundaries where dose gradients are steepest.

None of these is a novel concern and the field has raised all three. What has been less examined is how consistently the practice persists despite them, and what the magnitude of the resulting ambiguity is.

This review therefore characterises metric usage across a large reviewed sample, sets reported scores against reported observer variability, and examines a case in which geometric and dosimetric assessments diverge sharply.

2. Derivation of boundary sensitivity

The Dice similarity coefficient is twice the intersection volume divided by the sum of the two volumes, and is therefore dominated by the interior of a structure. To quantify how much of a structure a boundary error can affect, we computed the fraction of a spherical structure of radius r lying within a shell of thickness t as one minus ((r – t)/r) cubed. A sphere is an idealisation and real structures are irregular, with higher surface-to-volume ratios, so this calculation is conservative: it understates the proportion of a real structure that a boundary error would touch, and correspondingly understates the problem it illustrates.

2.1. Statistical analysis

Reported metric usage counts and overlap statistics were extracted as published. Derived quantities are deterministic functions of the stated inputs. No pooling, heterogeneity estimation or significance testing was performed. Analyses were performed in Python 3 using NumPy (Taha and Hanbury, 2015).

3. Results

Seven sources met the inclusion criteria and six contributed data: a systematic review of evaluation methodology

Table 1: Characteristics of the contributing sources

Source	Type	Scope	Principal contribution
Evaluation methodology review	Systematic review	117 studies	Metric usage; reference standard practice
Nodal level study	Clinical evaluation	20 nodal levels	Model against observer variability
Cervical cancer evaluation	Clinical evaluation	Independent cohort	Structure-by-structure Dice
Commercial software study	Implementation study	Multiple sites	Geometric and dosimetric together
Glioblastoma target study	Clinical evaluation	100 cases	Target volume Dice; editing time
Breast radiotherapy study	Clinical evaluation	Clinical vs pilot	Qualitative acceptability
Brain organ-at-risk study	Clinical evaluation	22 structures	Dose impact of geometric variation

Source: Overlap in the structures and systems evaluated. Figures are as reported and were not recomputed

Table 1: Characteristics of the contributing sources

Source	Type	Scope	Principal contribution
Evaluation methodology review	Systematic review	117 studies	Metric usage; reference standard practice
Nodal level study	Clinical evaluation	20 nodal levels	Model against observer variability
Cervical cancer evaluation	Clinical evaluation	Independent cohort	Structure-by-structure Dice
Commercial software study	Implementation study	Multiple sites	Geometric and dosimetric together
Glioblastoma target study	Clinical evaluation	100 cases	Target volume Dice; editing time
Breast radiotherapy study	Clinical evaluation	Clinical vs pilot	Qualitative acceptability
Brain organ-at-risk study	Clinical evaluation	22 structures	Dose impact of geometric variation

Source: Overlap in the structures and systems evaluated. Figures are as reported and were not recomputed

across 117 studies (Radiother Oncol., 2021); a head and neck nodal level study reporting model performance against observer variability (arXiv preprint 2208.13224, 2022); a cervical cancer evaluation reporting structure-by-structure overlap (Sci Rep., 2022); a commercial software implementation reporting geometric and dosimetric outcomes together (PMC9782080); a glioblastoma target volume evaluation (Radiother Oncol., 2025); and a breast radiotherapy study comparing clinical use against a pilot setting (PMC11332522). Characteristics are in Table 1.

3.1. The metric monoculture

Across 117 reviewed studies, a geometric metric was used in 116 (99.1%) and the Dice similarity coefficient specifically in 113 (96.6%). Metrics with a direct clinical interpretation were far rarer: dosimetric assessment in 27 studies (23.1%), qualitative or physician rating in 22 (18.8%), time saving in 18 (15.4%) and editing time—arguably the quantity most relevant to whether a system helps a clinician—in 11 (9.4%). This is shown in Figure 2.

Heterogeneity within categories compounds the imbalance. Over 90 different names were used for geometric measures, and qualitative assessment methods differed in all but two papers, so even the minority of studies reporting clinically interpretable outcomes rarely reported comparable ones.

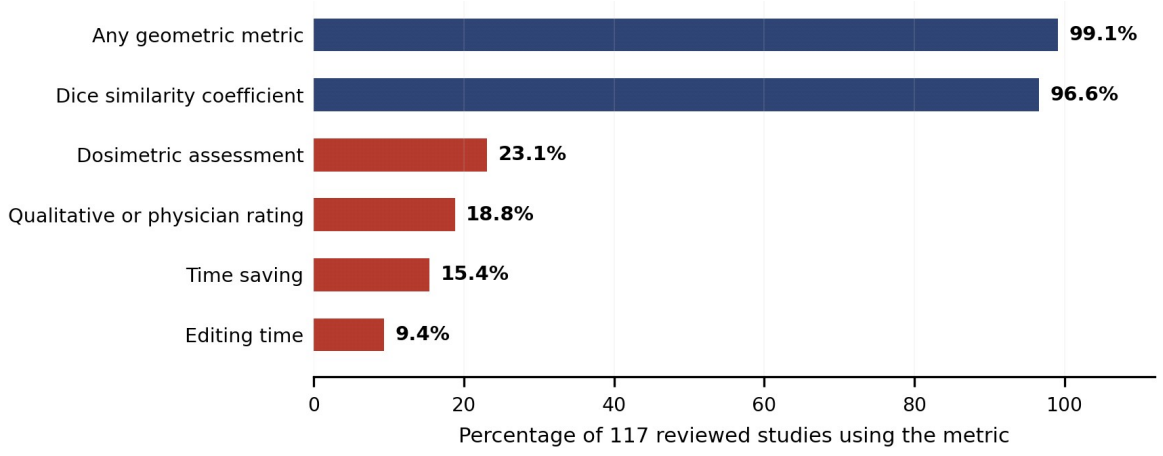

Figure 1: Metric usage across 117 reviewed evaluations. Blue denotes geometric overlap measures and red denotes metrics with a direct clinical interpretation. Editing time, arguably the quantity most relevant to whether a system helps a clinician, was reported in under a tenth of studies

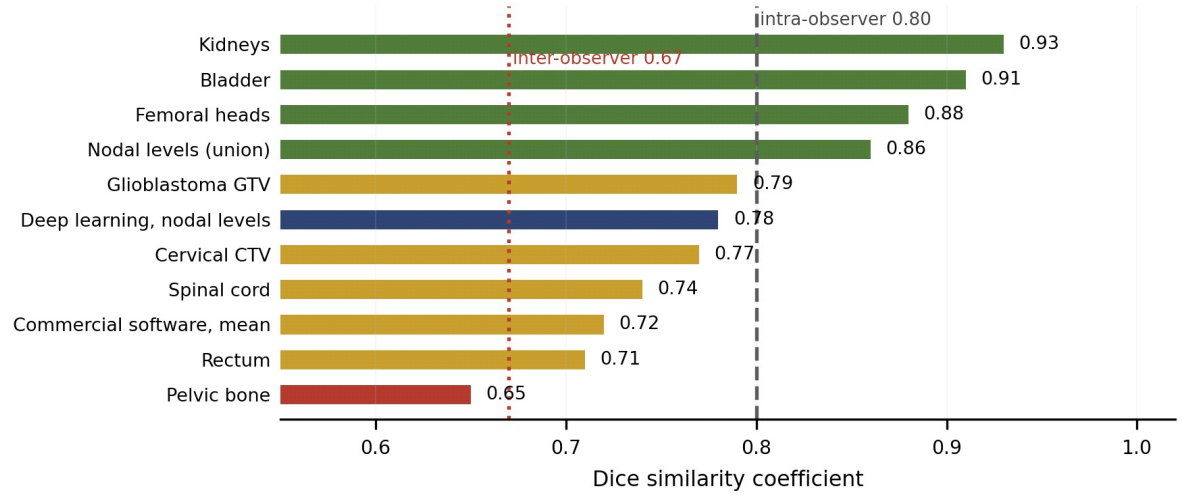

Figure 2: Reported Dice values by structure and system, against the reported intra-observer and inter-observer agreement for expert manual delineation. Values above the inter-observer line exceed the reproducibility of the standard being compared against, and that excess is not interpretable as accuracy

The review that assembled these figures drew the conclusion directly: many of the most commonly used geometric indices, including the Dice similarity coefficient, are not well correlated with clinically meaningful endpoints (Sherer *et al.*, 2021). The metric that 97% of studies report is one the field has stated does not predict what it is used to predict.

3.2. The reference standard and its reproducibility

A single manual contour served as ground truth in 65 of 117 studies (55.6%). Only 31 (26.5%) compared auto-contours against inter- or intra-observer variation, meaning that 74% of studies did not establish how reproducible their own reference standard was.

Where it has been measured, the answer constrains interpretation considerably. For expert manual delineation of head and neck nodal levels, intra-observer agreement was a Dice of approximately 0.80 and inter-observer agreement approximately 0.67. Against that background, a deep learning model scored 0.76 against intra-observer 0.77, a difference that was not statistically significant ($p = 0.307$), and 0.78 against an inter-observer figure of 0.67.

That second comparison is the one that requires care. A model scoring 0.78 where two experts agree at 0.67 is scoring above the reproducibility of the standard it is being measured against. The excess cannot be read as accuracy, because there is no single correct contour against which 0.78 is closer than 0.67 would be; there is a distribution of expert opinions, and the model sits within it. Reported scores against those two reference lines are shown in Figure 3.

The structure-by-structure spread reinforces the point. Reported Dice values ranged from 0.93 for kidneys to 0.65 for pelvic bone, and a single acceptability threshold applied across that range would classify well-delineated large organs and poorly delineated irregular ones by the same criterion.

3.3. Geometric agreement and dosimetric consequence

The clearest illustration of the disconnect comes from a commercial software implementation reporting both together, shown in Figure 4. Average Dice across structures was 0.72, which the report describes as acceptable in clinical practice. In a rectal cancer case from the same evaluation, the mean volume receiving 45 Gy was 10.4 cm^3 with the manual contour and 289.4 cm^3 with the automatic contour—a 27.8-fold difference, and an absolute difference of 279 cm^3.

The mechanism was a definitional disagreement rather than a modelling failure: the software treated the bowel structure as the entire abdominal cavity while the manual contour bounded it by the presence of

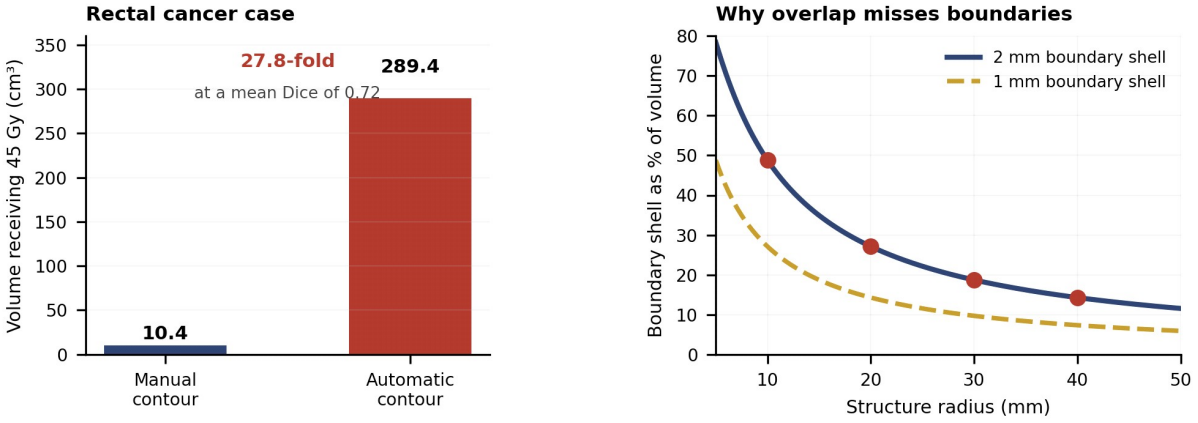

Figure 3: Left: volume receiving 45 Gy by manual and automatic contour in a rectal case from an evaluation hose mean Dice of 0.72 was described as clinically acceptable. Right: the fraction of a spherical structure lying within a boundary shell, showing why volume-overlap statistics are least sensitive where dose gradients are steepest

intestinal loops. That is a large volumetric difference in a region where the two definitions overlap substantially, so overlap statistics register it weakly while dose calculation registers it fully.

Contributing sources state the general relationship explicitly: the relationship between geometrical performance metrics and dosimetric impact cannot be predicted, and even where geometric differences are small the impact on the final dose distribution may still be clinically relevant (Front Oncol., 2021). A brain organ-at-risk study reached the opposite conclusion in its own setting—geometric variations had minimal impact on dose distribution (PMC10276287)—which is consistent with the same point: the relationship is not fixed, so it must be measured rather than assumed in either direction.

3.4. Why overlap statistics miss boundary error

There is a structural reason why Dice is insensitive to the errors radiotherapy cares about. Dose gradients are steepest at structure boundaries, and a volume-overlap statistic weights every voxel equally, so it is dominated by the interior.

Quantifying this for a spherical structure, a 2 mm boundary shell constitutes 48.8% of the volume at 10 mm radius, 27.1% at 20 mm, 18.7% at 30 mm and 14.3% at 40 mm. A systematic 2 mm boundary error on a 30 mm structure can therefore move less than a fifth of its volume, and the Dice penalty is correspondingly modest, while the dosimetric consequence at a steep gradient may not be.

This calculation is conservative. Real anatomical structures are irregular with higher surface-to-volume ratios than spheres, so a real boundary shell is a larger fraction of volume than computed here and the sensitivity is somewhat better than stated. The qualitative conclusion is unaffected: overlap statistics are least informative precisely where dose is most sensitive, which is why surface-based measures such as surface Dice and 95th percentile Hausdorff distance were developed (Nikolov *et al.*, 2021) and are used by several contributing studies.

4. Discussion

97% of published evaluations of deep learning auto-segmentation report a Dice similarity coefficient. Fewer than a quarter report a dosimetric outcome, fewer than a fifth a physician rating, and fewer than a tenth the time required to edit the output. The metric that dominates is one the field itself has concluded is not well correlated with clinically meaningful endpoints.

Two structural problems make that dominance more consequential than a mere reporting preference. The first concerns what the score is measured against. 56% of studies used a single manual contour as ground truth and only a quarter established observer variability at all. Where it has been measured, two experts delineating the same nodal levels agreed at a Dice of about 0.67. A model scoring 0.78 against that background

Table 2: Evaluation practice and its consequences

Quantity	Value	Note
Studies using any geometric metric	116 of 117 (99.1%)	Reviewed sample
Studies using Dice specifically	113 of 117 (96.6%)	—
Studies with dosimetric assessment	27 of 117 (23.1%)	—
Studies with qualitative rating	22 of 117 (18.8%)	—
Studies reporting editing time	11 of 117 (9.4%)	—
Single manual contour as ground truth	65 of 117 (55.6%)	—
Compared against observer variation	31 of 117 (26.5%)	74% did not
Distinct names for geometric measures	over 90	—
Expert intra-observer Dice	~0.80	Nodal level delineation
Expert inter-observer Dice	~0.67	Same task
Model vs intra-observer	0.76 vs 0.77	$p = 0.307$
Model vs inter-observer	0.78 vs 0.67	Exceeds the reference reproducibility
45 Gy volume, manual vs automatic	10.4 vs 289.4 cm^3	27.8-fold, at mean Dice 0.72
2 mm shell of a 30 mm sphere	18.7% of volume	Derived

Note: Percentages are of the 117 studies in the contributing methodology review. Derived quantities are deterministic and carry no independent uncertainty.

is not 0.78 accurate; it is sitting inside a distribution of expert opinions whose width exceeds the difference being reported. The natural reading of a high score—that the model is close to correct—requires a correct answer to exist, and for many structures the field has not established one.

The second concerns what the score is sensitive to. Dice weights all voxels equally and is dominated by structure interiors, while dose gradients are steepest at boundaries. For a 30 mm structure a 2 mm boundary shell is under a fifth of the volume, so a systematic boundary error produces a modest Dice penalty and a potentially large dosimetric one. Surface-based measures address this and are used by some contributing studies, but they remain a minority practice.

The rectal case makes the consequence concrete. A mean Dice of 0.72 described as clinically acceptable accompanied a 27.8-fold difference in the volume receiving 45 Gy. The cause was a definitional disagreement about the extent of a bowel structure rather than a modelling failure—which is exactly the kind of discrepancy an overlap statistic registers weakly and a dose calculation registers fully.

It is worth stating what this does not show. A brain organ-at-risk study in the same body of evidence found that geometric variations had minimal effect on dose distribution, and that finding is equally valid. The relationship between geometry and dose is not fixed; it depends on the structure, its position relative to the target, and the plan. That is precisely why it has to be measured rather than inferred from an overlap statistic in either direction.

The recommendation follows from the numbers and is not new, though it remains largely unimplemented. Evaluations should pair a geometric measure with a dosimetric assessment and a physician rating, should report the observer variability of the reference standard alongside the model score, and should report editing time (Vaassen *et al.*, 2020). Every one of these is already collected by some studies; none requires new technology; and the two that matter most for clinical readiness appear in under a quarter of the literature.

5. Limitations

- The metric usage figures derive from a single systematic review of 117 studies with its own inclusion criteria, and a differently constituted sample might give different proportions.
- Observer variability figures come from one delineation task in one anatomical region. Inter-observer agreement varies substantially by structure and is higher for well-defined organs than for nodal levels or clinical target volumes.
- The dosimetric example is a single case from one evaluation, and its 27.8-fold discrepancy arose from a definitional disagreement rather than from segmentation error as such. It illustrates the disconnect rather than quantifying its typical magnitude.
- The boundary shell calculation assumes a spherical structure. Real structures are irregular with higher surface-to-volume ratios, so the calculation understates overlap sensitivity and is conservative in the direction of the argument being made.
- Reported Dice values were extracted across different structures, imaging modalities, institutions and reference standards and are not comparable with one another; they are presented to show spread rather than to rank systems.
- A model scoring above inter-observer agreement is not thereby shown to be inaccurate, only that the score does not establish accuracy. Distinguishing the two requires an outcome-anchored reference that this literature does not provide.
- Editing time, the quantity most relevant to clinical utility, was reported too rarely and too heterogeneously to characterise.
- Systems, training data and anatomical sites differ across contributing studies, and no attempt was made to attribute performance differences to any of these.
- Downstream clinical outcomes—whether auto-segmentation changes toxicity or tumour control—were not reported by any contributing source and are the endpoint that would settle the question.
- No pooling was performed and no formal quality appraisal instrument was applied to the contributing evaluations.
- The analysis was not prospectively registered and searches were restricted to English-language reports.

6. Conclusion

Deep learning auto-segmentation for radiotherapy is evaluated almost exclusively by geometric overlap: a geometric metric in 99.1% of 117 reviewed studies and the Dice coefficient in 96.6%, against dosimetric assessment in 23.1%, physician rating in 18.8% and editing time in 9.4%. A single manual contour served as ground truth in 55.6% of studies and only 26.5% established observer variability.

That matters because the reference standard is not stable—expert inter-observer Dice for one delineation task was 0.67, below the 0.78 achieved by a model—and because overlap statistics are structurally insensitive to boundary error, a 2 mm shell being under a fifth of a 30 mm structure. In one evaluation a mean Dice of 0.72 described as clinically acceptable accompanied a 27.8-fold difference in the volume receiving 45 Gy. Geometric agreement should be reported with, not instead of, dosimetric assessment, physician rating and the observer variability of the reference standard.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Clinical note

This review concerns how auto-segmentation performance is reported and does not evaluate the safety or accuracy of any specific commercial system. Clinical deployment should follow local quality assurance procedures with physician review of every contour.

References

[Author list to be completed from source record] (2021). A preliminary experience of implementing deep-learning based auto-segmentation in head and neck cancer. *Front Oncol,* 11: 638197. doi: 10.3389/fonc.2021.638197

[Author list to be completed from source record] (2021). Clinical evaluation of deep learning and atlas-based auto-segmentation for radiotherapy: A systematic review of evaluation methodology. *Radiother Oncol.* doi: 10.1016/j.radonc.2021.09.019

[Author list to be completed from source record]. Comparison of the use of a clinically implemented deep learning segmentation model with the simulated study setting for breast cancer patients receiving radiotherapy. [PMC11332522].

[Author list to be completed from source record] (2022). Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy. *arXiv preprint 2208.13224.*

[Author list to be completed from source record]. Deep-learning magnetic resonance imaging-based automatic segmentation for organs-at-risk in the brain: Accuracy and impact on dose distribution. [PMC10276287].

[Author list to be completed from source record] (2022). Evaluation of auto-segmentation for EBRT planning structures using deep learning-based workflow on cervical cancer. *Sci Rep*, 12: 13650. doi: 10.1038/s41598-022-18084-0

[Author list to be completed from source record] (2025). Evaluation of compartmentalized automatic segmentation for definition of the GTV in glioblastoma radiotherapy. *Radiother Oncol.*

[Author list to be completed from source record]. Implementation of a commercial deep learning-based auto segmentation software in radiotherapy: Evaluation of effectiveness and impact on workflow. [PMC9782080].

Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology,* 26(3): 297-302.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J. *et al.* (2021). Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J Med Internet Res,* 23(7): e26151.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. *et al.* (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: n71.

Sherer, M.V., Lin, D., Elguindi, S., Duke, S., Tan, L.T., Cacicedo, J. *et al.* (2021). Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol*, 160: 185-191.

Taha, A.A. and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*, 15: 29.

Vaassen, F., Hazelaar, C., Vaniqui, A., Gooding, M., van der Heyden, B., Canters, R. and van Elmpt, W. (2020). Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*, 13: 1-6.

Cite this article as: Victor Chukwu (2025). Scoring above the Noise Floor: Overlap Metrics, Observer Variability and Dosimetric Impact in Deep Learning Auto-Segmentation for Radiotherapy. *Biomed Pharma Reviews.* 1(1), 10-17. doi: 10.62587/BMPR.1.1.2025.10-17.