Data-Driven Reliability Audit of nnU-Net Segmentation for BNCT Treatment Planning

Irina Baymuratova
Tashkent State Technical University named after Islam Karimov, University Str. 8, 100178 Tashkent, Uzbekistan
Duke University, Duke University Road 2080, 27708 Durham, North Carolina, United States
xerson2681@gmail.com

Keywords: Artificial Intelligence, nnU-Net, Medical Image Segmentation, BNCT, Data Analytics, Reliability Audit, Dice Coefficient, Treatment Planning.

Abstract: Artificial intelligence-based image segmentation can accelerate treatment-planning workflows, but geometric segmentation errors may propagate to downstream dosimetric calculations. This study presents a secondary computational analysis of published patient-level nnU-Net results for 16 Glioblastoma Multiforme cases evaluated in Boron Neutron Capture Therapy planning. Manual and neural-network-generated outputs were compared for tumour volume, irradiation time, and D98, D50, and D2 dose indicators. Mean absolute error, mean absolute percentage error, bias, Pearson correlation, and non-parametric comparisons were calculated. An exploratory retrospective audit additionally stratified cases using a Dice coefficient threshold of 0.80. The mean absolute percentage error was 73.71% for tumour volume, 1.88% for irradiation time, 7.21% for D98, 2.83% for D50, and 1.27% for D2. In cases with Dice coefficient >= 0.80, D98 percentage error was 1.64%, compared with 10.55% in lower-Dice cases. Higher segmentation overlap was associated with lower downstream computational error in the analysed cases. The Dice-informed stratification is proposed as a retrospective reliability-audit method rather than a prospective clinical decision rule, because Dice calculation requires a reference contour.

1 INTRODUCTION

Artificial intelligence has become an important component of medical image analysis, particularly in segmentation, image reconstruction, classification, and treatment-planning support [1]. Deep learning can automate high-dimensional imaging tasks, but the transition from model development to operational healthcare workflows requires attention to data quality, reliability, and downstream effects [2].

Automatic segmentation is especially important in radiation oncology and nuclear medicine because contour differences can affect subsequent computational planning. Convolutional neural networks have demonstrated strong capabilities in anatomical image segmentation [3], while modified U-Net architectures have improved tumour-focused processing of CT data [4]. However, geometric overlap alone does not fully describe whether an AI-generated contour is reliable for downstream dose calculations.

Pezzi et al. evaluated nnU-Net segmentation of Glioblastoma Multiforme (GBM) for Boron Neutron Capture Therapy (BNCT) treatment planning and reported patient-level segmentation and dosimetric outputs for 16 selected test cases [5]. Their data permit an additional data-analytics question: how strongly do segmentation discrepancies propagate to irradiation time and dose indicators, and can segmentation overlap retrospectively stratify cases by downstream computational error?

AI-assisted treatment planning increasingly combines target localisation, segmentation, dose prediction, and planning support [6]. This study therefore performs a secondary computational re-analysis of published nnU-Net outputs. The main contribution is a reproducible error-based reliability audit of tumour volume, irradiation time, D98, D50, and D2, followed by an exploratory Dice-informed stratification of lower- and higher-overlap cases. The analysis is positioned as applied AI and data analytics rather than as a clinical validation study.

2 METHODOLOGY

2.1 Data Source and Study Design

This study performed a secondary computational re-analysis of patient-level results reported by Pezzi et al. [5]. The source study used CT data from The Cancer Imaging Archive and evaluated nnU-Net-based automatic GBM segmentation for BNCT treatment planning. The present analysis used the 16 cases for which the source article reported Dice coefficient values, manually segmented tumour volumes, neural-network-generated tumour volumes, irradiation times, and D98, D50, and D2 dose indicators. The patient-level numerical values were extracted from the published source tables [5], organized into a structured case-level dataset, and independently recalculated before statistical analysis.

No new medical images were collected and the nnU-Net model was not retrained in the present study. Manual segmentation values reported in the source article were treated as reference outputs for the secondary computational comparison. The original patient identifiers were retained only as study-case labels during calculation; the reported analysis uses aggregate statistics.

2.2 Quantitative Error Analysis

For each parameter, the neural-network output was compared with the corresponding manually derived reference value. Absolute error, mean absolute error (MAE), mean absolute percentage error (MAPE), and mean bias were calculated. For case i, the absolute percentage error was defined as:

$$APE_i = |x_{AI,i} - x_{ref,i}| / |x_{ref,i}| \times 100\%. \quad (1)$$

MAPE was calculated as the arithmetic mean of APE across the 16 cases. Pearson correlation coefficients were used to assess linear association between manual and AI-generated outputs. Because of the limited sample size, paired differences were additionally evaluated using the two-sided Wilcoxon signed-rank test. The significance level was set at 0.05. Data-driven analysis of biomedical variables can reveal relationships that are not apparent when individual parameters are considered separately [7].

2.3 Dice-Informed Retrospective Reliability Stratification

To investigate whether segmentation overlap was associated with downstream computational error, an exploratory retrospective stratification was performed using a Dice coefficient (DC) threshold of 0.80. Cases were divided into DC < 0.80 and DC >= 0.80 groups. Mean absolute percentage errors were compared between the groups using the two-sided exact Mann-Whitney U test. The 0.80 threshold was used only for analytical stratification and is not proposed as a clinical acceptance criterion. Because Dice requires a reference contour, it is treated here as a retrospective audit variable rather than an inference-time deployment signal. Five parameters were examined in a small secondary dataset; therefore, p-values are reported descriptively without multiplicity correction.

2.4 Reliability Audit Workflow

The proposed retrospective reliability audit follows a sequential multi-stage analytical workflow. First, manually segmented and AI-generated outputs are aligned at the case level. Second, segmentation overlap is characterized using the Dice coefficient reported for each case. Third, downstream computational deviations are quantified for tumour volume, irradiation time, D98, D50, and D2 using absolute and percentage-based error measures. Fourth, association and paired-comparison statistics are calculated to determine whether systematic differences are present between manually and automatically derived outputs. Finally, cases are stratified by segmentation overlap to examine whether lower geometric agreement is associated with increased downstream computational error.

The workflow is intended to separate geometric segmentation performance from computational consequence. A low Dice coefficient does not automatically imply an unacceptable treatment-planning result, while high geometric overlap does not independently guarantee stability of all dosimetric parameters. Therefore, the proposed audit evaluates both segmentation similarity and error propagation across downstream variables.

In practical AI validation, the same workflow can be used as an offline model-audit procedure. Segmentation models can first be evaluated on a labelled validation dataset containing expert reference contours. Dice scores and downstream planning deviations are then calculated for each case. The resulting relationship between model confidence, segmentation overlap, and computational error may subsequently be used to calibrate a prospective uncertainty indicator for deployment. This separates retrospective model auditing from inference-time decision support and avoids the incorrect use of Dice

as a directly available real-time variable. The sequence of the proposed reliability-audit procedure is illustrated in Figure 1.

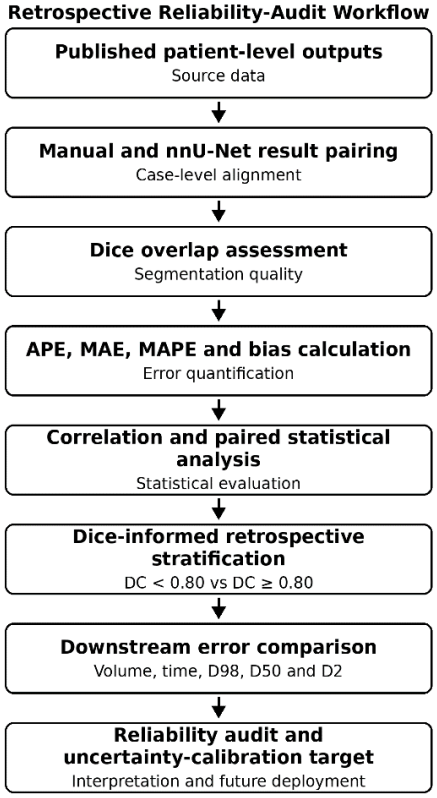

Figure 1: Proposed retrospective reliability-audit workflow for AI-generated segmentation outputs.

3 RESULTS

3.1 Quantitative Comparison of Manual and AI-Generated Outputs

The overall agreement and error characteristics of manual and AI-generated computational outputs are summarized in Table 1.

As shown in Table 1, tumour-volume estimates exhibited the largest relative deviation, with a MAPE of 73.71%. In contrast, downstream computational parameters showed smaller relative errors: 1.88% for irradiation time, 7.21% for D98, 2.83% for D50, and 1.27% for D2. Pearson correlations were 0.925, 0.954, and 0.967 for D98, D50, and D2, respectively, indicating strong linear association between manually and AI-derived dosimetric outputs. The Wilcoxon signed-rank test identified a paired shift for tumour volume ($p = 0.039$), whereas irradiation time and the three dose indicators did not show statistically significant paired differences ($p >= 0.258$).

Table 1: Quantitative comparison of manual and AI-generated outputs.

Parameter	MAE	MAPE (%)	Mean bias	Pearson *r*
Tumour volume	37.570 cm³	73.71	+22.001 cm³	0.712
Irradiation time	0.635 min	1.88	+0.306 min	0.944
D98	1.356 Gy	7.21	-0.531 Gy	0.925
D50	0.713 Gy	2.83	-0.175 Gy	0.954
D2	0.363 Gy	1.27	+0.163 Gy	0.967

To assess the association between segmentation overlap and downstream error, the cases were stratified by Dice coefficient, as presented in Table 2.

Table 2: Mean absolute percentage error by Dice coefficient group.

Parameter	DC < 0.80 (n=10)	DC >= 0.80 (n=6)	p-value
Tumour volume	112.94%	8.32%	0.022
Irradiation time	2.61%	0.66%	0.147
D98	10.55%	1.64%	0.002
D50	3.58%	1.60%	0.031
D2	1.71%	0.55%	0.428

Table 2 shows a marked reduction in computational error among cases with higher segmentation overlap. The mean D98 percentage error decreased from 10.55% in the DC < 0.80 group to 1.64% in the DC >= 0.80 group. Tumour-volume error decreased from 112.94% to 8.32%, while D50 error decreased from 3.58% to 1.60%. Exploratory Mann-Whitney analysis indicated group differences for tumour-volume error ($p = 0.022$), D98 error ($p = 0.002$), and D50 error ($p = 0.031$). The irradiation-time and D2 comparisons were not statistically significant.

3.2 Error Propagation Pattern

The combined results of Tables 1 and 2 indicate that segmentation disagreement did not propagate uniformly across computational outputs. Tumour volume showed the largest relative sensitivity to contour differences, whereas irradiation time and D2 remained comparatively stable. D98 demonstrated an

intermediate but pronounced computational sensitivity, with substantially larger percentage errors in the lower-Dice group.

This pattern suggests that the influence of segmentation error depends on the downstream variable being evaluated. Geometric contour deviations can produce large volumetric discrepancies without causing proportional changes in every calculated treatment parameter. Therefore, a reliability audit based only on tumour-volume agreement may overestimate the instability of some downstream outputs, while evaluation based only on irradiation time or maximum dose may underestimate the effect of segmentation disagreement on coverage-related indicators such as D98.

The observed parameter-specific behaviour supports a multi-indicator audit approach. Segmentation overlap, volumetric deviation, and dosimetric error should be interpreted jointly rather than as isolated performance metrics.

4 DISCUSSION

The results reveal an important distinction between geometric segmentation error and downstream computational stability. AI-generated tumour volumes showed substantial relative deviations in several low-Dice cases, while irradiation time and central or maximum dose indicators remained comparatively stable. The high tumour-volume MAPE partly reflects pronounced over-contouring in several low-overlap cases and the amplification of percentage error when reference tumour volumes are small. Among the dosimetric indicators, D98 was the most sensitive to segmentation disagreement.

The Dice-based grouping is useful as a retrospective reliability audit because it shows that poorer geometric overlap is associated with larger D98 and tumour-volume errors in the analysed cases. It must not, however, be presented as a prospective automated quality-control rule: Dice cannot be calculated at inference time without a reference contour. A practical AI workflow should instead calibrate an inference-time uncertainty or confidence indicator against Dice and downstream-error behaviour on validation data, and then use the calibrated indicator to trigger expert review.

From an applied IT perspective, the present contribution is a transparent data-analytics layer for auditing the reliability of AI-generated segmentation outputs. AI-assisted medical workflows remain vulnerable to bias, model error, and inappropriate deployment [8], while methodological weaknesses and inadequate validation can produce models that are unsuitable for clinical use [9]. Interpretable data-driven prioritization has also been applied in other safety-critical engineering workflows, where quantitative indicators are transformed into transparent review priorities [10]. The same decision-support principle motivates future calibration of segmentation uncertainty as a review-priority signal in BNCT planning.

The present study is limited by the small sample size and by reliance on published patient-level results from a single source study. Manual contours were treated as reference outputs, although inter-observer variability is known to affect medical segmentation. The DC threshold of 0.80 was chosen for exploratory stratification and must not be interpreted as a clinical decision threshold. Moreover, the five group comparisons were not adjusted for multiple testing. Future work should evaluate the audit framework on larger independent datasets and calibrate prospective inference-time uncertainty indicators against observed downstream dosimetric error.

5 CONCLUSIONS

This study presented a secondary computational reliability audit of nnU-Net-generated tumour segmentation outputs for BNCT treatment planning. Quantitative comparison showed that segmentation-volume errors can be substantially larger than the corresponding deviations in irradiation time and dosimetric indicators. Among the analysed dose parameters, D98 was the most sensitive to segmentation disagreement.

Exploratory Dice-based stratification demonstrated that cases with DC >= 0.80 had considerably lower tumour-volume, D98, and D50 percentage errors: specifically, D98 error decreased from 10.55% in the lower-Dice group to 1.64% in the higher-Dice group, and tumour-volume error decreased from 112.94% to 8.32% (Table 2). The main contribution is a transparent retrospective data-analytics approach for auditing how segmentation overlap relates to downstream computational error.

The proposed analysis represents an applied AI and data-analytics contribution rather than a clinical validation study. Dice itself is not proposed as an inference-time acceptance rule. Future work should calibrate deployable uncertainty or confidence

indicators against segmentation quality and dosimetric error on larger independent datasets.

6 ACKNOWLEDGMENTS

The author acknowledges the members of the innovative project of the Ministry of Innovative Development of the Republic of Uzbekistan and the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan, together with the Institute of Nuclear Physics of the Academy of Sciences of the Republic of Uzbekistan, project No. AM-PZ 2019062031, 'Creation of multimedia textbooks for bachelor's and master's degree students in Nuclear Energy, Nuclear Medicine and Technology, and Radiation Medicine and Technology.

REFERENCES

[1] L. Xing, M. L. Giger, and J. K. Min, Artificial Intelligence in Medicine: Technical Basis and Clinical Applications. London, UK: Academic Press, 2020.

[2] A. Zhang, L. Xing, J. Zou, and J. C. Wu, “Shifting machine learning for healthcare from development to deployment and from models to data,” Nature Biomedical Engineering, vol. 6, pp. 1330-1345, 2022.

[3] B. Ibragimov and L. Xing, “Segmentation of organs-at-risk in head and neck CT images using convolutional neural networks,” Medical Physics, vol. 44, pp. 547-557, 2017.

[4] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, “Modified U-Net (mU-Net) with incorporation of object-dependent high-level features for improved liver and liver-tumor segmentation in CT images,” IEEE Transactions on Medical Imaging, vol. 39, pp. 1316-1325, 2020.

[5] C. Pezzi, F. Morosato, B. Marcaccio, et al., “Dosimetric comparison of the BNCT treatment planning performances when using a nnU-Net to automatically segment Glioblastoma Multiforme,” Health and Technology, vol. 15, pp. 811-822, 2025, [Online]. Available: https://doi.org/10.1007/s12553-025-00996-2.

[6] T. J. Netherton, C. E. Cardenas, D. J. Rhee, L. E. Court, and B. M. Beadle, “The emergence of artificial intelligence within radiation oncology treatment planning,” Oncology, vol. 99, pp. 124-134, 2021.

[7] M. Islam and L. Xing, “A data-driven dimensionality-reduction algorithm for the exploration of patterns in biomedical data,” Nature Biomedical Engineering, vol. 5, pp. 624-635, 2021.

[8] R. Challen et al., “Artificial intelligence, bias and clinical safety,” BMJ Quality & Safety, vol. 28, pp. 231-237, 2019.

[9] M. Roberts et al., “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” Nature Machine Intelligence, vol. 3, pp. 199-217, 2021.

[10] R. Salmorbekova and M. Talipov, “Digital risk matrix and safety management workflow for airport infrastructure in developing countries: a data-driven prioritization approach,” Vibroengineering Procedia, vol. 62, pp. 653-661, Jun. 2026, [Online]. Available: https://doi.org/10.21595/vp.2026.26121.