ecancermedicalscience

Benchmarking of radiobiological NTCP models in head and neck radiotherapy using independent computational pipelines: an institutional validation study with machine learning augmentation

Kalyan Mondal[1,2,3a], **Abhijit Mandal**[3b], **Anuj Vijay**[1c] and **Ganeshkumar Patel**[3d]

[1]Department of Physics, Institute of Applied Science & Humanities, GLA University, Mathura, Uttar Pradesh 281406, India
[2]North Bengal Medical College, Sushrutanagar, Darjeeling 734012, India
[3]Department of Radiotherapy and Radiation Medicine, Institute of Medical Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh 221005, India
[a] http://orcid.org/0000-0001-7685-7391
[b] http://orcid.org/0000-0001-5626-1072
[c] http://orcid.org/0000-0001-6610-3844
[d] http://orcid.org/0000-0002-1876-6069

Abstract

Background & purpose: Normal tissue complication probability (NTCP) models require institutional validation before clinical implementation. Traditional radiobiological models, such as the Lyman–Kutcher–Burman (LKB) and Equivalent Uniform Dose (EUD) models, provide mechanistic dose–response frameworks, while machine learning (ML) approaches offer exploratory, data-driven alternatives that remain inadequately characterised in South Asian populations.

Methods: This retrospective study included 51 head and neck cancer patients treated with definitive radiotherapy. Binary endpoints were Grade ≥2 xerostomia (n = 3), dysphagia (n = 5) and mucositis (n = 4), scored using Common Terminology Criteria for Adverse Events version 5.0. NTCP calculations were performed using two independent computational pipelines (MATLAB-based RBMODELv1 and a Python implementation), with agreement assessed using Bland–Altman analysis. Traditional NTCP models (LKB, EUD) were evaluated and compared with artificial neural networks and XGBoost in a hypothesis-generating framework using a stratified 70:30 train–test split. Model performance was assessed using the area under the receiver operating characteristic curve (area under the curve), accuracy and Spearman's rank correlation.

Results: Excellent agreement was observed between computational pipelines (mean bias 0.8%, 95% limits −1.9% to 3.5%). Traditional models demonstrated strong rank-order correlation with toxicity grades (ρ = 0.61–0.79, p < 0.001) and high accuracy (LKB: 90.0%–94.1%). Institution-specific parameters differed from quantitative analyses of normal tissue effects in the clinic values, including a lower parotid TD50 (34.1 versus 39.0 Gy). Exploratory ML analyses showed numerically higher discrimination for parallel organs but not for mixed-architecture structures; however, severe class imbalance (3–5 events per endpoint) limits statistical reliability.

Conclusion: Dual computational pipelines enable reproducible NTCP modeling for institutional use. Traditional radiobiological models perform acceptably after local calibration,

Research

Correspondence to: Ganeshkumar Patel
Email: ganeshgravity@gmail.com

ecancer 2026, **20**:2147
https://doi.org/10.3332/ecancer.2026.2147

Published: 16/06/2026
Received: 09/02/2026

Publication costs for this article were supported by *ecancer* (UK Charity number 1176307).

while exploratory ML findings suggest potential organ-architecture-dependent patterns that require validation in adequately powered multi-institutional cohorts.

Keywords: *normal tissue complication probability, radiobiological models, dose-volume histogram, machine learning, head and neck neoplasms*

Introduction

Head and neck cancer radiotherapy achieves high locoregional control rates exceeding 70% for locally advanced disease, but radiation-induced toxicities, including xerostomia, dysphagia and mucositis, significantly impact quality of life and functional outcomes [1–3]. Modern treatment planning relies predominantly on empirical dose-volume histogram (DVH) constraints that provide binary pass or fail assessments without quantifying actual complication risks [4]. Normal tissue complication probability (NTCP) models offer continuous risk estimates by integrating dose-volume data into mechanistic frameworks, enabling more nuanced treatment plan evaluation and optimisation [5–7].

The Lyman-Kutcher-Burman (LKB) model remains the most widely validated NTCP framework, employing a cumulative normal distribution with three tissue-specific parameters, TD50 (dose causing a 50% complication probability), m (slope of the dose-response curve) and n (volume effect parameter) [8, 9]. The Equivalent Uniform Dose (EUD) formalism reduces heterogeneous dose distributions to biologically effective uniform doses through the generalised EUD (gEUD) concept, with organ-specific volume parameter 'a', reflecting architectural tolerance patterns [10, 11]. The Relative Seriality (RS) model explicitly incorporates organ functional subunit architecture, distinguishing parallel organs (functional redundancy) from serial organs (sequential vulnerability) through the seriality parameter 's' [12, 13]. These models underpin the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) guidelines that established dose-volume constraints based on pooled literature analysis [14–17].

Despite extensive validation in Western populations, [15–17] several critical knowledge gaps remain. First, QUANTEC parameters were derived predominantly from European and North American cohorts, with limited validation in Asian populations, where genetic polymorphisms [18–20], lifestyle factors [21–23] and treatment protocols [24–27] may influence toxicity profiles. Second, computational implementation varies across software platforms, potentially affecting prediction consistency and clinical adoption [28–30]. Third, while machine learning (ML) approaches demonstrate promise in multi-factorial prediction [31–33], their performance relative to mechanistic models remains incompletely characterised, particularly across organs with distinct architectural properties [34]. Parallel organs such as parotid glands exhibit volume-dependent tolerance through functional subunit redundancy, whereas serial structures like pharyngeal constrictors demonstrate threshold-based dose-response patterns [12, 13]. Whether data-driven approaches offer advantages over mechanistic models in these distinct scenarios requires systematic investigation [34–36]. Furthermore, if population-specific factors alter toxicity risk, purely mechanistic models may be limited [37]. Data-driven ML approaches, capable of integrating dosimetric, clinical and potentially genomic variables, may offer a complementary path for personalised prediction in distinct cohorts [38]. Consequently, ML models are best viewed as exploratory, hypothesis-generating tools rather than replacements for mechanistic NTCP models at the current stage of evidence [38, 39].

This institutional validation study addresses these gaps through three primary objectives: (1) to validate traditional NTCP models (LKB, EUD and RS) using dual independent computational pipelines to ensure reproducible implementation; (2) to derive institution-specific model parameters reflecting local patient demographics, treatment protocols and toxicity assessment methods and (3) to conduct a hypothesis-generating exploratory comparison with contemporary ML algorithms (artificial neural networks (ANNs) and extreme gradient boosting) to explore potential organ-architecture-dependent performance patterns that can inform future multi-institutional validation. Within the constraints of an exploratory, single-institution cohort, this study establishes baseline NTCP model performance in an Indian population and identifies specific research questions requiring validation through multi-institutional collaboration.

Materials and methods

Study design and patient selection

This retrospective study was conducted at a tertiary cancer care institute in India between March 2024 and September 2024. A total of 63 consecutive patients with histologically confirmed head and neck squamous cell carcinoma treated with definitive radiotherapy were initially identified. Inclusion criteria comprised age ≥18 years, Karnofsky performance status ≥70, Stage I–IVB disease according to the American Joint Committee on Cancer 8th edition and completion of the prescribed radiotherapy course, with sufficient treatment planning data available to permit DVH analysis. Exclusion criteria included prior head and neck radiotherapy, distant metastases at diagnosis, loss to follow-up before the first toxicity assessment or incomplete DVH data precluding NTCP modeling. After applying these criteria, 12 patients were excluded (incomplete DVH data, n = 7; loss to follow-up, n = 3; baseline xerostomia before radiotherapy, n = 2), resulting in 51 analysable patients (81%).

Sample size and statistical considerations

This study's retrospective design and limited sample size constrain the statistical strength of our findings. A post-hoc power analysis revealed that, with 51 patients and observed event rates of 5.9%–9.8%, the study was underpowered (30%–40%) to detect area under the curve (AUC) differences of 0.15 between models at α = 0.05 [40]. Furthermore, adhering to the 'events per variable' (EPV) guideline of ≥10 events per predictor, [41, 42] our endpoints provided only 3–5 events. This restricts traditional modeling to a maximum of 1–2 univariate predictors and renders any multivariable ML model suitable only for exploratory, hypothesis-generating analysis. Therefore, all multivariable analyses, especially those employing ML, must be interpreted as hypothesis-generating. The sample size is sufficient to identify general performance trends and generate organ-specific hypotheses, but is inadequate for definitive conclusions on model superiority. We explicitly acknowledge these constraints and contextualise our findings accordingly [43]. This approach is consistent with prior single-institution NTCP benchmarking studies with comparable event rates.

Treatment planning and delivery

Treatment plans were generated using the Eclipse Treatment Planning System (version 15.6; Varian Medical Systems, Palo Alto, CA, USA). Patients received either three-dimensional conformal radiotherapy (3DCRT, n = 23, 45%) or volumetric modulated arc therapy (VMAT, n = 28, 55%). The prescribed dose was 66–70 Gy in 2-Gy fractions (median 68 Gy) delivered over 6.5–7 weeks. Organs at risk were delineated following international consensus guidelines [44]. Concurrent chemotherapy (weekly cisplatin 40 mg/m^2 or cetuximab loading 400 mg/m^2 followed by weekly 250 mg/m^2) was administered to 21 patients (41.2%) per institutional protocol.

Toxicity assessment

Toxicity was assessed using Common Terminology Criteria for Adverse Events version 5.0 (CTCAE v5.0) at baseline, weekly during treatment, and at 3, 6, 9 and 12 months post-treatment. Toxicity grading was primarily based on patient-reported symptoms captured using validated EORTC QLQ-C30 and QLQ-H&N35 instruments, with CTCAE v5.0 grades assigned accordingly. The primary binary endpoint was Grade ≥2 toxicity, aligning with QUANTEC guidelines and most comparative literature. This threshold represents clinically meaningful impairment requiring intervention: Grade ≥2 xerostomia indicates moderate symptoms limiting self-care, Grade ≥2 dysphagia requires dietary modification and Grade ≥2 mucositis necessitates analgesic support.

DVH extraction and processing

Cumulative DVHs were extracted from archived treatment plans using the Eclipse scripting interface. DVH files contained dose-volume data at 0.1-Gy resolution, exported in text format with two columns: absolute dose (cGy) and absolute volume (cm^3). Analysed organs included

bilateral parotid glands, larynx and oral cavity. For bilateral parotid analysis, the ipsilateral and contralateral glands were combined as recommended by QUANTEC [15].

NTCP model implementations

NTCP calculations were performed using two independent computational pipelines to ensure reproducibility and implementation robustness.

Pipeline 1 – RBMODELv1 (MATLAB R2021a): A MATLAB-based NTCP calculation platform developed and previously validated at our institution [45]. The software incorporates Niemierko's EUD formalism with documented corrections to published code implementations and supports LKB, gEUD and RS models.

Pipeline 2 – Python Implementation (Python 3.9): An independent custom implementation developed using NumPy (v1.21) and SciPy (v1.7), publicly available via a GitHub repository. All algorithms were implemented using identical mathematical formulations and numerical conventions to those used in RBMODELv1.

All NTCP formulations followed canonical published definitions for the LKB, gEUD and RS models [8–13]. The gEUD formalism provides a unifying framework, of which the LKB EUD and Niemierko EUD formulations represent specific parameterisations.

For the LKB model, NTCP was calculated using the Lyman probit formulation with the Kutcher–Burman dose–volume reduction method [9]:

$$NTCP_{LKB} = \Phi(t)$$

where,

$$t = (EUD - TD50)/(m \times TD50)$$

and the EUD for the LKB model was computed as follows:

$$EUD = (\Sigma_i\, v_i\, D_i^{(1/n)})^n.$$

Here, $\Phi(t)$ denotes the cumulative distribution function of the standard normal distribution, v_i represents the fractional volume receiving dose D_i, TD50 is the dose associated with 50% complication probability for uniform irradiation, 'm' is the slope parameter and 'n' characterises the volume effect.

The gEUD formulation was expressed as follows [11]:

$$gEUD = (\Sigma_i\, v_i\, D_i^{a})^{(1/a)}$$

where the parameter 'a' characterises organ architecture, with large negative values corresponding to parallel organs and large positive values corresponding to serial organs.

The RS model was implemented according to published formulations [12]:

$$NTCP_{RS} = 1 - \prod_i \left[1 - P(D_i)^s\right]^{(v_i/s)}$$

where,

$$P(D_i) = 1/[1 + (D50/D_i)^k].$$

Here, $P(D_i)$ is the probability that a small fractional volume of an organ will develop a complication if it receives dose D_i, 's' denotes the RS parameter ($s \to 0$ for parallel organs, $s \to 1$ for serial organs), D50 represents the dose associated with 50% complication probability for uniform whole-organ irradiation and 'k' defines the slope of the dose–response relationship.

Initial model parameters were derived from published literature and QUANTEC recommendations, [8, 9, 14–17] followed by institutional calibration. Final parameter values used in this study were: parotid (TD50 = 34.1 Gy, m = 0.11, n = 1.0), larynx (TD50 = 43.6 Gy, m = 0.16, n = 0.45) and oral cavity (TD50 = 48.5 Gy, m = 0.18, n = 1.0).

ML implementation

Given the limited number of Grade ≥2 toxicity events per endpoint (3–5), ML analyses were undertaken for exploratory, hypothesis-generating purposes rather than for clinical model development or deployment. The available event counts are substantially below commonly recommended EPV thresholds for stable multivariable prediction modeling, [41, 42] and all ML results should therefore be interpreted as illustrative of relative trends within this dataset.

Feature engineering: A total of seventeen DVH–derived features were extracted for each organ, including mean dose, maximum dose, minimum dose, median dose, V10–V70 (volumes receiving ≥10–70 Gy in 10-Gy increments) and D2cc (minimum dose to the hottest 2 cm^3). In addition, three clinical variables (age, sex and concurrent chemotherapy) were included, resulting in a total of 20 candidate predictors per endpoint.

Model architecture: ANN models consisted of a single hidden layer with ten nodes and ReLU activation, trained using binary cross-entropy loss and the Adam optimiser (learning rate 0.001), with a maximum of 100 epochs and early stopping (patience = 10 epochs). Extreme gradient boosting (XGBoost) models employed an ensemble of 100 decision trees with a maximum depth of 3, learning rate of 0.1 and L2 regularisation (λ = 1.0) [46].

Class imbalance handling: To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) [47] was applied exclusively to the training data, generating synthetic minority-class samples to achieve an approximate 1:2 class ratio (original ratios ranged from 1:10 to 1:16). This approach was applied conservatively, recognising that with very small minority-class sizes (n = 3–5), synthetic samples may incompletely reflect the underlying data distribution [48].

Hyperparameter optimisation: Hyperparameter tuning was performed using grid search with five-fold cross-validation on the training set. Owing to the low number of events, individual folds frequently contained zero or one event, limiting the stability of cross-validated optimisation and reinforcing the exploratory nature of these analyses [49].

Statistical analysis

Performance metrics: Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC), with 95% confidence intervals estimated using DeLong's method for correlated ROC curves [50]. Given the very low event counts per endpoint (n = 3–5), Brier scores were considered unreliable for calibration assessment and are therefore not reported; this limitation is explicitly acknowledged. Accuracy is reported for completeness; however, given the binary endpoints and marked class imbalance, it primarily reflects the prediction of the majority class and was not used as a primary performance metric [51].

Correlation analysis: Because NTCP outputs are continuous probability estimates and toxicity grades are ordinal categorical variables, associations were evaluated using Spearman's rank correlation coefficient (ρ) rather than Pearson's correlation or linear regression. Squared Spearman correlations (ρ^2) are reported to indicate the proportion of variance in toxicity grade rankings explained by NTCP-based rankings. This approach is more appropriate than linear regression–based R^2, which assumes continuous, normally distributed outcomes [52].

Model comparison: Differences in AUC between models were assessed using DeLong's test for correlated ROC curves [50]. To account for multiple comparisons across endpoints and model pairs, the Bonferroni correction was applied (adjusted significance threshold α = 0.05/9 = 0.0056). Results were interpreted in the context of a limited sample size and event counts [53].

Computational agreement: Agreement between NTCP values generated by RBMODELv1 and the independent Python implementation was evaluated using Bland–Altman analysis [54]. Mean bias and 95% limits of agreement were calculated as bias ± 1.96 × SD of the differences. Acceptable agreement was predefined as a mean bias <3% and 95% limits of agreement within ±5%.

Confidence intervals: Bootstrap resampling (1,000 iterations, stratified by outcome) was used to generate additional 95% confidence intervals for AUC estimates. Given the low number of events per endpoint (3–5), bootstrap-based estimates may exhibit increased variability [55]. Accordingly, both DeLong-based and bootstrap confidence intervals are reported for completeness, with results interpreted cautiously.

Software: Statistical analyses were performed using R version 4.2.0 (packages: *pROC*, *caret*, *SMOTE*), Python 3.9 (*scikit-learn*, *XGBoost*), MATLAB R2021a (RBMODELv1) and SPSS version 26.

All statistical tests were two-tailed. In light of the limited sample size and class imbalance, emphasis was placed on effect sizes and confidence intervals rather than sole reliance on *p*-values, recognising that statistical significance does not necessarily imply clinical relevance or adequate statistical power [56, 57].

The schematic diagram of the computational framework and workflow is shown in Figure 1.

Schematic overview of the dual computational pipeline used for NTCP modeling. The workflow shows: (A) Patient selection and DVH extraction, (B) Dual-pipeline NTCP calculation (RBMODV1(i.e., RBMODELv1) in MATLAB and Python implementation), (C) Traditional radiobiological modeling (LKB, EUD, RS), (D) Exploratory ML analysis (ANN, XGBoost with SMOTE) and (E) Statistical validation including Bland–Altman analysis for computational agreement and DeLong test for model comparison.

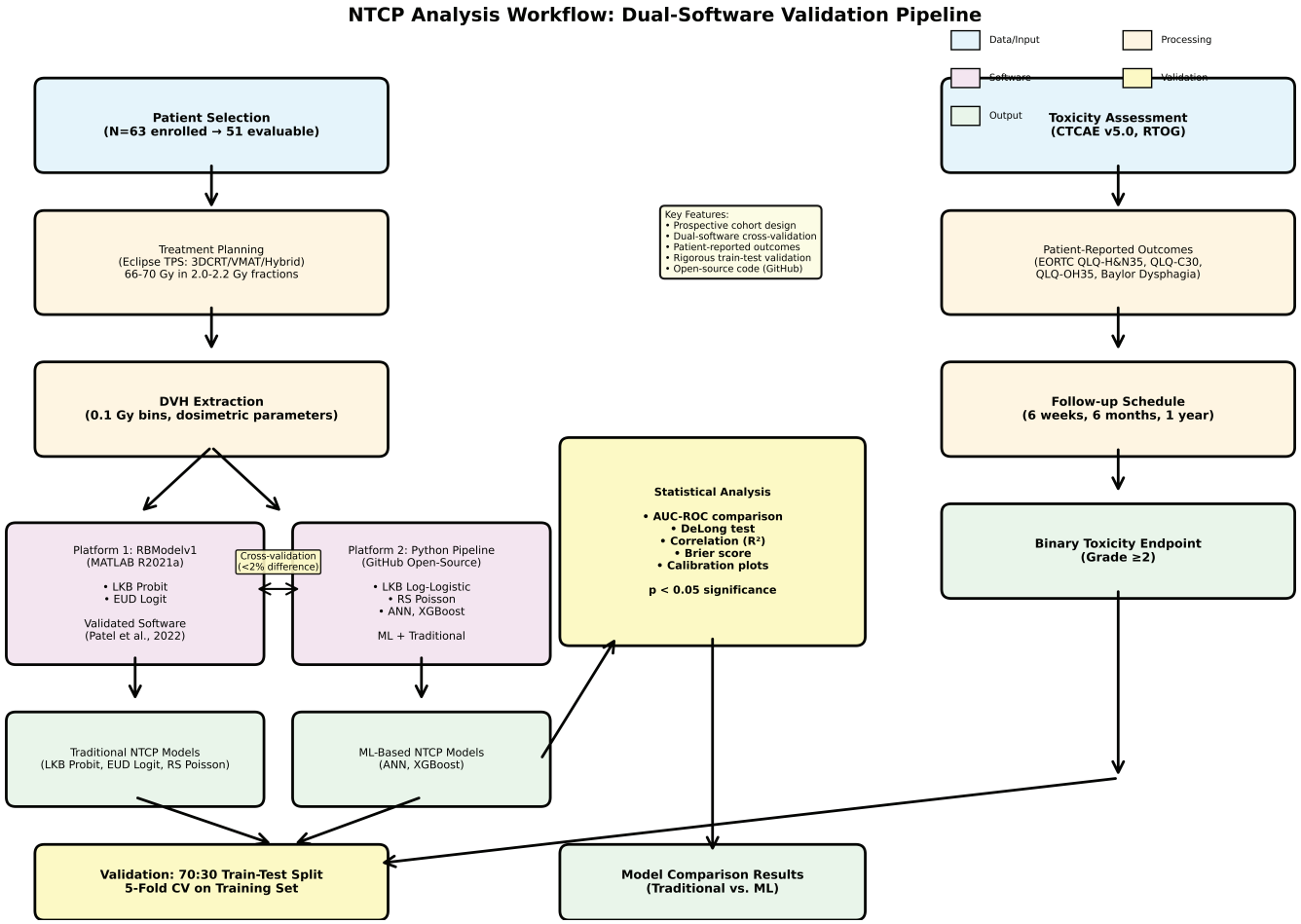

Figure 1. Study flowchart and computational framework.

Results

Patient cohort and toxicity events

Patient and treatment characteristics are summarised in Table 1. The cohort was representative of a typical head and neck cancer population, with a median age of 56 years (range, 28–74), a predominance of male patients (66.7%), a high proportion of advanced-stage disease (70.6% Stage III–IV) and a mix of treatment techniques (3DCRT 45%, VMAT 55%). Grade ≥ 2 toxicity events were infrequent, with xerostomia observed in three patients (5.9%), dysphagia in five patients (9.8%) and mucositis in four patients (7.8%). While these low event rates are consistent with contemporary radiotherapy techniques and adherence to established dose constraints, they resulted in pronounced class imbalance across endpoints, with event-to-non-event ratios ranging from approximately 1:9 to 1:16. This imbalance limits the statistical power available for model discrimination analyses, particularly for multivariable and ML approaches [58]. Consistent with this, post-hoc power estimation indicated that the available sample size provided approximately 30%–40% power to detect an AUC difference of 0.15 at a two-sided $\alpha = 0.05$ [40]

DVH characteristics

Mean organ doses across the cohort were 42.8 ± 14.8 Gy for the parotid glands, 48.1 ± 11.9 Gy for the larynx and 48.7 ± 12.0 Gy for the oral cavity. Patients who developed Grade ≥ 2 toxicities generally received higher mean organ doses compared with those without such events, although differences did not consistently reach statistical significance. Specifically, mean parotid dose was 53.5 Gy versus 42.1 Gy ($p = 0.204$), mean laryngeal dose was 51.7 Gy versus 47.7 Gy ($p = 0.527$) and mean oral cavity dose was 60.8 Gy versus 47.7 Gy ($p = 0.030$) for patients with and without Grade ≥ 2 toxicity, respectively. These comparisons should be interpreted cautiously, given the limited number of events per endpoint. Only six patients (11.8%) achieved the QUANTEC-recommended constraint of a mean parotid dose ≤ 25 Gy, [6] reflecting the challenge of meeting stringent dose–volume objectives in routine clinical practice for locally advanced head and neck cancers.

3DCRT, three-dimensional conformal radiotherapy; VMAT, volumetric modulated arc therapy. Grade ≥2 toxicity defined per CTCAE v5.0 criteria

Dual-pipeline computational validation

Bland–Altman analysis demonstrated excellent agreement between the MATLAB-based and Python-based pipelines across all organs for the LKB model (Figure 2). The mean bias was 0.8%, with 95% limits of agreement ranging from −1.9% to +3.5%, meeting the predefined acceptability criteria (mean bias <3% and limits within ±5%). The maximum observed individual deviation was 2.4% and occurred at higher NTCP values (>80%), where small absolute differences result in larger relative percentage differences. Separate Bland–Altman analyses for the EUD and RS model outputs yielded comparable agreement bounds (mean bias <1.0%, 95% LoA within ±4.0% for all evaluated models); however, detailed panels are not presented as LKB served as the primary validation model for this study. Overall, these findings indicate that, when identical mathematical formulations and numerical conventions are applied, NTCP calculations are reproducible and computationally consistent across software platforms.

Bland–Altman plots demonstrating agreement between MATLAB-based RBMODELv1 and independent Python pipeline NTCP calculations for the LKB model across all three organs: (A) Parotid LKB model, (B) Larynx LKB model, (C) Oral cavity LKB model. *X*-axis shows mean NTCP between two implementations; *Y*-axis shows difference (Python - MATLAB) as percentage. Solid horizontal line indicates mean bias (0.8%); dashed lines show 95% limits of agreement (−1.9% to +3.5%). Agreement meets pre-defined acceptability criteria (bias <3%, limits within ±5%), confirming computational reproducibility. Maximum deviation of 2.4% occurred at high NTCP values (>80%). Detailed Bland–Altman panels for EUD and RS model outputs are not separately shown; EUD and RS agreement was confirmed to fall within comparable acceptability bounds (mean bias <1.0%, 95% LoA within ±4.0%).

Table 1. Patient and treatment characteristics (*N* = 51).

Characteristic	Value
Age, median (range), years	56 (28–74)
Sex: Male/Female, *n* (%)	34 (66.7)/17 (33.3)
Primary site, *n* (%)	
Larynx	17 (33.3)
Tongue	14 (27.5)
Buccal mucosa	9 (17.6)
Others	11 (21.6)
Disease stage: I–II/III–IV, *n* (%)	15 (29.4)/36 (70.6)
Treatment technique, *n* (%)	
3DCRT	23 (45.1)
VMAT	28 (54.9)
Concurrent chemotherapy, *n* (%)	21 (41.2)
Total dose, mean ± SD, Gy	67.7 ± 1.2
Toxicity outcomes (Grade ≥2)	
Xerostomia, *n* (%)	3 (5.9)
Dysphagia, *n* (%)	5 (9.8)
Mucositis, *n* (%)	4 (7.8)
Mean organ doses, Gy	
Parotid (mean ± SD)	42.8 ± 14.8
Larynx (mean ± SD)	48.1 ± 11.9
Oral cavity (mean ± SD)	48.7 ± 12.0

Traditional NTCP model performance

Benchmarking results for traditional NTCP models are summarised in Table 2. The LKB probit model demonstrated the highest discrimination across all three endpoints, with AUC values of 0.89 for parotid toxicity, 0.88 for laryngeal toxicity and 0.91 for oral cavity toxicity, although confidence intervals were wide owing to the limited number of events. The EUD-based model achieved comparable discrimination for dysphagia (AUC 0.83) and xerostomia (AUC 0.82), but showed lower performance for mucositis (AUC 0.64). The RS model demonstrated lower discrimination for parotid toxicity (AUC 0.59, 95% CI 0.46–0.72), consistent with the known limitations of serial-organ modeling when applied to predominantly parallel organ architectures such as the parotid glands.

Rank-order associations between calculated NTCP values and observed toxicity grades were moderate to strong across endpoints (ρ^2 = 0.37–0.63, all $p < 0.001$), indicating robust relative risk stratification despite potential uncertainty in absolute probability calibration (Figure 3). Reported accuracy values (85.0%–94.1%) largely reflected correct classification of the majority class (absence of Grade ≥ 2 toxicity) and were therefore not interpreted as indicators of clinical utility in the context of severe class imbalance [51].

Institution-specific parameter estimation revealed systematic differences compared with QUANTEC-recommended values. In particular, the parotid TD50 was estimated at 34.1 Gy, compared with 39.0 Gy reported in QUANTEC, with a corresponding slope parameter m of 0.11 versus 0.18 as shown in the institute-specific dose-response curves (Figure 4). The detailed comparison of parameters is given in Table 3. These differences may reflect a combination of population-specific factors, differences in toxicity assessment methodology and contemporary treatment practices, including a higher proportion of VMAT use in the present cohort (55%) relative to earlier QUANTEC-era studies [1, 15, 59–62].

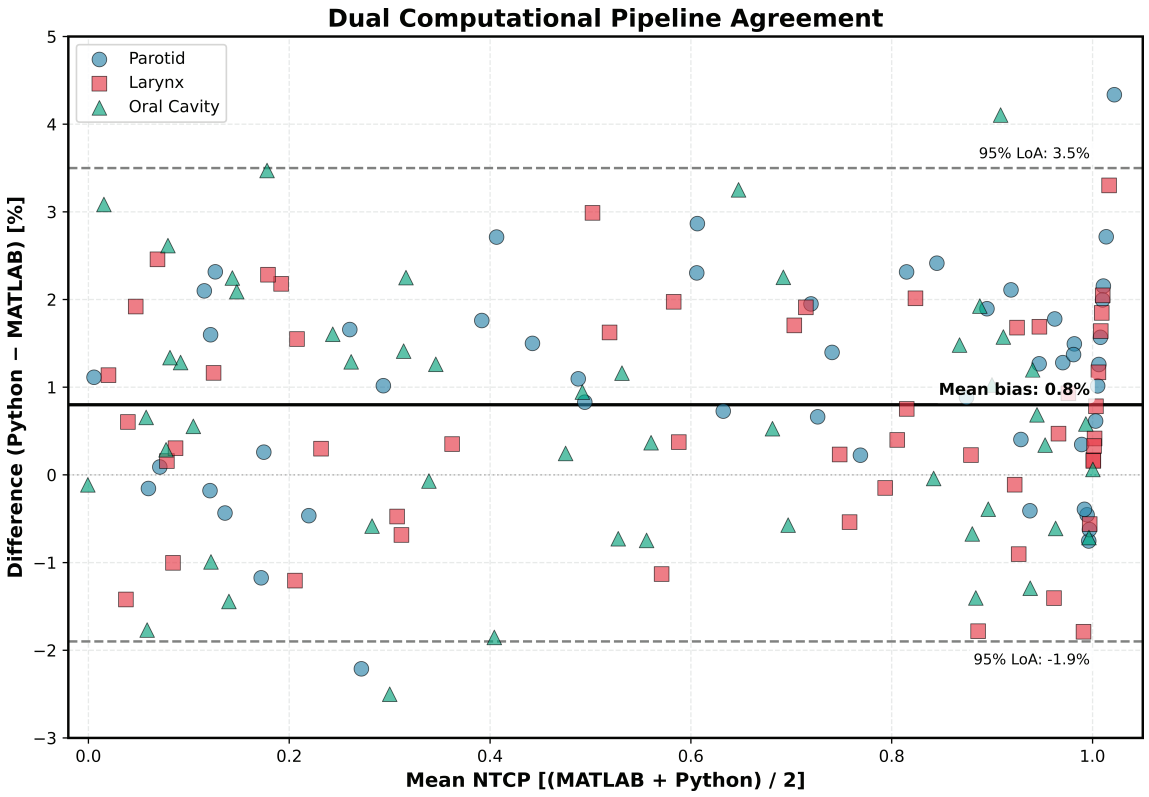

Figure 2. Bland-Altman analysis for computational validation.

Table 2. Traditional radiobiological model performance benchmarking.

Organ (Toxicity)	Model	AUC (95% CI)	Accuracy (%)	ρ^2
Parotid (Xerostomia)				
	LKB Probit	0.89 (0.82–0.96)	94.1	0.61
	EUD Logit	0.82 (0.74–0.90)	88.2	0.50
	RS Poisson	0.59 (0.46–0.72)	85.0	0.47
Larynx (Dysphagia)				
	LKB Probit	0.88 (0.79–0.95)	90.9	0.37
	EUD Logit	0.83 (0.73–0.92)	86.4	0.37
Oral cavity (Mucositis)				
	LKB Probit	0.91 (0.85–0.97)	90.0	0.63
	EUD Logit	0.64 (0.52–0.76)	85.0	0.09

Note: AUC, area under the curve; CI, confidence interval; LKB, Lyman-Kutcher-Burman; EUD, equivalent uniform dose; RS, relative seriality; ρ^2, squared Spearman rank correlation

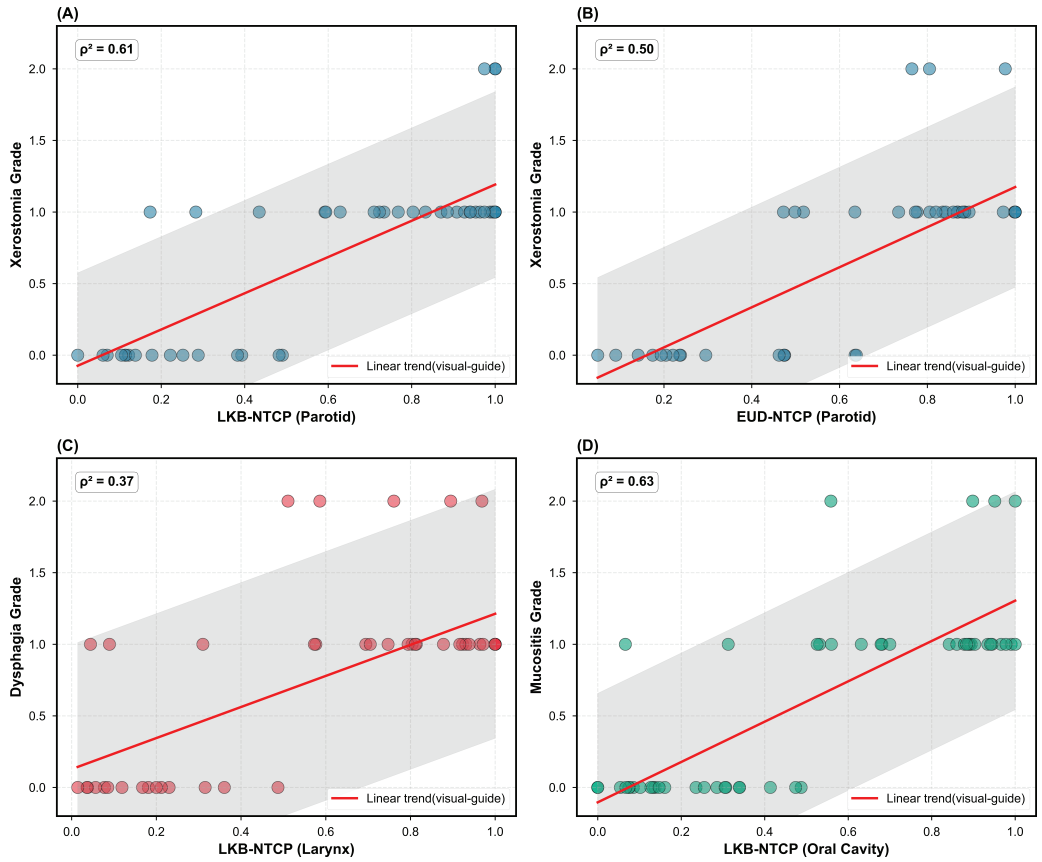

Figure 3. Correlation between NTCP and observed toxicity grades.

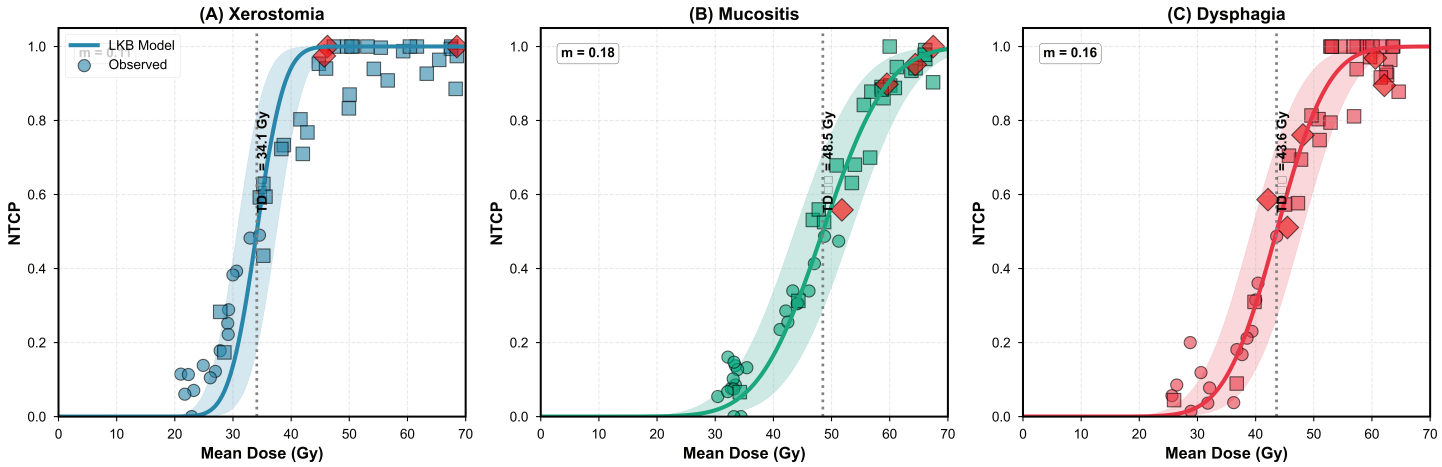

Figure 4. Institution-specific dose-response curves.

Table 3. Comparison of institution-specific NTCP model parameters with QUANTEC-recommended values.

Organ	Model	Parameter	This study	QUANTEC [Ref]	Diff.	Possible explanation
LKB model parameters						
Parotid	LKB	TD_{50} (Gy)	34.1	39.0 [15]	−4.9	Population genetics; PRO versus physician-rated toxicity
		m	0.11	0.18	−0.07	Steeper dose–response in this cohort
		n	1.0	1.0	0.0	—
Larynx	LKB	TD_{50} (Gy)	43.6	46.0 [16]	−2.4	Treatment technique differences (VMAT 55% versus ~30%)
		m	0.16	0.16	0.0	—
		n	0.45	0.45	0.0	—
Oral Cavity	LKB	TD_{50} (Gy)	48.5	50.0 [6,14]	−1.5	Similar to parotid; PRO detection earlier
		m	0.18	0.18	0.0	—
		n	1.0	1.0	0.0	—
gEUD volume-effect parameter						
Parotid	gEUD	a	−1	−1 [15]	0.0	Parallel architecture; consistent with QUANTEC classification
Larynx	gEUD	a	+3	+3 [16]	0.0	Mixed architecture (serial-dominant); consistent with QUANTEC
Oral Cavity	gEUD	a	−1	−1 [6,14]	0.0	Parallel architecture designation
RS model parameters (Parotid only)						
Parotid	RS	D_{50} (Gy)	34.1	39.9 [15]	−5.8	Same trend as LKB TD_{50}; population-specific shift
		k	2.1	2.11	−0.01	—
		s	0.05	0.07	−0.02	Lower seriality reflects parallel organ nature

Scatter plots with regression lines showing correlation between calculated NTCP and observed toxicity grades (0, 1, 2). (A) Xerostomia vs. LKB-NTCP for parotid (ρ^2 = 0.61, ρ = 0.78, p < 0.001); (B) Xerostomia versus EUD-NTCP for parotid (ρ^2 = 0.50, ρ = 0.71, p < 0.001); (C) Dysphagia versus LKB-NTCP for larynx (ρ^2 = 0.37, ρ = 0.61, p < 0.001); (D) Mucositis versus LKB-NTCP for oral cavity (ρ^2 = 0.63, ρ = 0.79, p < 0.001). Solid lines represent ordinary least-squares (OLS) linear trend lines shown for visual reference only; statistical association was quantified by Spearman's rank correlation coefficient (ρ). Dashed lines indicate 95% confidence bands of the OLS trend line. Strong rank-order correlations indicate robust relative risk stratification despite imperfect absolute probability calibration.

Sigmoid dose-response curves fitted with institution-specific LKB parameters. (A) Parotid xerostomia: TD50 = 34.1 Gy, m = 0.11, n = 1.0 (versus QUANTEC TD50 = 39.0 Gy); (B) Oral cavity mucositis: TD50 = 48.5 Gy, m = 0.18, n = 1.0 (versus QUANTEC TD50 = 50.0 Gy); (C) Larynx dysphagia: TD50 = 43.6 Gy, m = 0.16, n = 0.45 (versus QUANTEC TD50 = 46.0 Gy). Observed data points shown as circles (Grade 0), squares (Grade 1), and rhombus (Grade 2). Systematically lower TD50 values reflect population-specific factors and methodological differences from QUANTEC cohorts.

All correlations p < 0.001. **Accuracy values** (85.0%–94.1%) primarily reflect correct prediction of the majority class (Grade 0–1 toxicity) and should NOT be interpreted as indicators of clinical utility given severe class imbalance (events: non-events ratios approximately 1:9 to 1:16, consistent with Table 1 event counts). **AUC confidence intervals** are wide due to limited events (n = 3–5), indicating substantial statistical uncertainty in discrimination estimates. **ρ^2 values** represent squared Spearman rank correlations, NOT linear regression R^2, reflecting the proportion of variance in toxicity grade rankings explained by NTCP rankings. This is more appropriate for ordinal outcomes. **RS model's poor performance** for parotid (AUC 0.59) likely reflects architectural mismatch – applying a serial-organ model to an inherently parallel structure.

The RS model was applied only to the parotid gland for two reasons: (1) historical precedent – QUANTEC parameters [15] enabled direct comparison in Table 3; and (2) hypothesis testing – the poor performance empirically confirms that serial models are unsuitable for parallel organs. The RS model was not evaluated for the larynx (mixed architecture, no established parameters) or oral cavity (complex mucosa, serial formulation theoretically inappropriate). This selective application is a transparent limitation: parotid RS results serve primarily as historical reference and architectural validation, not as a clinically validated model. **Performance metrics** based on 70:30 train-test split with stratification. Cross-validation metrics not reported due to unreliability with <10 events per fold. **Statistical power** is inadequate (30%–40%) to detect clinically meaningful AUC differences of 0.15 at $\alpha = 0.05$.

Notes: TD_{50}, dose causing 50% complication probability; *m*, slope parameter; *n*, volume effect parameter; a, gEUD volume-effect parameter (negative = parallel, positive = serial); D_{50} (RS), dose causing 50% response in the RS model; *k*, dose–response steepness; *s*, seriality parameter (0 = purely parallel, 1 = purely serial); PRO, patient-reported outcomes; VMAT, volumetric modulated arc therapy

Systematically lower TD_{50} values in this study likely reflect: (1) genetic polymorphisms in South Asian populations affecting radiation sensitivity, (2) EORTC PRO instruments detecting toxicity earlier than LENT–SOMA physician ratings and (3) different spatial dose distributions with modern VMAT techniques.

RS model applicability: The RS model was applied only to the parotid gland for two reasons: (1) historical precedent – QUANTEC parameters [15] enabled direct comparison; and (2) hypothesis testing – the poor discrimination performance (AUC 0.59 in Table 2) empirically confirms that serial models are unsuitable for parallel organs. RS was not evaluated for larynx (mixed architecture, no established RS parameters) or oral cavity (complex mucosa, serial formulation theoretically inappropriate). These parameters are reported for completeness and transparency; their inclusion does not imply clinical validation of the RS model for parotid applications.

QUANTEC model parameters were taken from organ-specific QUANTEC publications: parotid glands [15], larynx/pharynx [16] and oral mucosa based on pooled QUANTEC consensus summaries [6, 14].

Exploratory ML analysis

Exploratory ML analyses were performed to examine whether data-driven approaches exhibit organ-dependent performance patterns when applied to the same dosimetric inputs used in traditional NTCP modeling. Given the limited number of Grade ≥ 2 toxicity events per endpoint in this cohort, these analyses were conducted for hypothesis-generating purposes rather than to support claims of clinical superiority. The small number of observed events (3–5 per endpoint) limits the statistical stability of multivariable models, and test-set performance is therefore sensitive to minor variations in event distribution [41, 42, 55]. In addition, although class-imbalance mitigation using SMOTE and internal cross-validation was applied exclusively within training data, very low event counts may restrict the representativeness of synthetic samples and internal validation estimates [48]. No external validation cohort was available, and statistical significance alone should not be interpreted as evidence of clinical utility or generalisability. Within these constraints, Table 4 and Figure 5 summarise exploratory performance trends. On the independent hold-out test set, ANN models were compared against the best-performing traditional model for each organ (LKB for all three). For the parotid glands, ANN (AUC 0.93) showed a modest numerical improvement over LKB (AUC 0.89; ΔAUC = +0.04; $p = 0.420$). For the oral cavity, ANN (AUC 0.95) showed a similar modest advantage over LKB (AUC 0.91; ΔAUC = +0.04; $p = 0.380$). For the mixed-architecture larynx, performance differences were similarly small and non-significant (ANN AUC 0.91 versus LKB 0.88; ΔAUC = +0.03; $p = 0.150$).

Table 4. Exploratory ML model performance comparison (hold-out test set).

Organ	Best traditional (AUC)	ANN (AUC)	ΔAUC	*p*-value†
Parotid	LKB: 0.89	0.93	+0.04	0.420
Larynx	LKB: 0.88	0.91	+0.03	0.150
Oral cavity	LKB: 0.91	0.95	+0.04	0.380

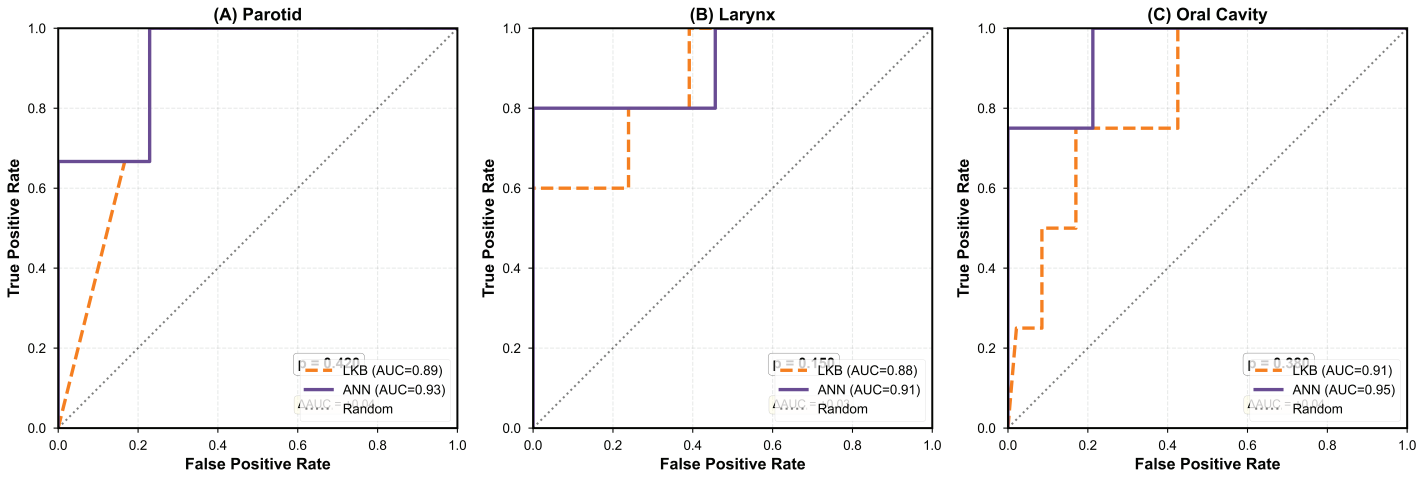

Figure 5. Exploratory ROC curve comparison: traditional versus ML models.

These findings support a hypothesis warranting further investigation in adequately powered, multi-institutional cohorts: ML approaches may offer relative advantages for parallel-architecture organs characterised by complex spatial dose distributions and functional subunit redundancy, whereas mechanistic radiobiological models may remain sufficient for organs with more straightforward dose–response relationships [42, 63].

ROC curves comparing traditional models and exploratory ML on hold-out test sets, using the best-performing traditional model (LKB) for each organ as a comparator. (A) Parotid xerostomia: ANN (AUC = 0.93, 95% CI 0.79–1.00) versus best traditional LKB model (AUC = 0.89, 95% CI 0.82–0.96), p = 0.420 by DeLong test; (B) Larynx dysphagia: ANN (AUC = 0.91, 95% CI 0.82–1.00) versus best traditional LKB model (AUC = 0.88, 95% CI 0.79–0.95), p = 0.150; (C) Oral cavity mucositis: ANN (AUC = 0.95, 95% CI 0.87–1.00) versus best traditional LKB model (AUC = 0.91, 95% CI 0.85–0.97), p = 0.380. *Note:* This is an exploratory ML model performance across organs. Results are hypothesis-generating only due to limited event counts (3–5 per endpoint). No statistically significant differences were observed between ANN and the best traditional model (LKB) for any organ. Statistical significance does not indicate clinical utility or generalisability.

Extreme gradient boosting (XGBoost) models showed lower generalisation performance compared with ANN across all endpoints, with evidence of overfitting despite regularisation (training AUC 0.98–1.00; test AUC 0.67–0.82). This overfitting pattern constitutes a methodologically important negative finding that further reinforces the cautionary narrative regarding ML in low-event settings. To maintain clarity and focus on the primary comparative patterns observed, detailed XGBoost results are provided as Supplementary Table S1.

No results were statistically significant (p < 0.05, DeLong test). †DeLong test for correlated ROC curves. Best traditional model for all organs: LKB (the highest performing traditional model per Table 2). ANN, artificial neural network; AUC, area under the curve; ΔAUC, difference in AUC (ANN minus best traditional model: LKB); LKB, Lyman–Kutcher–Burman

Notes: ***Severe class imbalance:*** Test sets contained only 1–2 positive events, rendering AUC estimates statistically unstable and confidence intervals unreliable. ***Violated sample size requirements:*** With 3–5 events total, the ML analysis violates established EPV guidelines (≥10 events per predictor). The 17–20 feature models represent extreme overfitting risk. ***No external validation:*** All results from the single-institution dataset; generalisability unknown. External validation in independent cohorts is absolutely essential. ***SMOTE limitations:*** Synthetic oversampling with n = 3–5 training events creates artificial data that may not represent the true minority class distribution. This is a fundamental methodological concern. ***Cross-validation unreliability:*** With 3–5 events, cross-validation folds contained 0–1 events, rendering hyperparameter optimisation unreliable. ***Statistical instability:*** DeLong test p-values are questionable with extreme class imbalance. Bootstrap confidence intervals similarly unreliable. Statistical Significance ≠ Clinical Utility: p < 0.05 does NOT indicate clinical readiness, generalisability or

superiority over traditional models. ***Appropriate interpretation:*** These exploratory findings suggest a hypothesis – ML may offer advantages for parallel-architecture organs – that requires validation in multi-institutional cohorts with ≥50 events per endpoint before any clinical consideration.

Feature importance analysis

Feature importance was explored using SHapley Additive exPlanations (SHAP) analysis applied to the ANN models. Across all organs, mean dose emerged as the most influential predictor (mean SHAP value 0.28), followed by intermediate dose–volume parameters such as V30 (0.17) and patient age (0.13) (Figure 6). DVH metrics reflecting mid-to-high dose regions (e.g., V30 and V50) consistently ranked higher than low-dose volume parameters.

These patterns are consistent with established radiobiological understanding and prior NTCP literature, which emphasise the relevance of mean dose and intermediate dose–volume exposure in normal tissue toxicity prediction. Given the exploratory nature of the ML analyses and limited event counts, feature importance rankings should be interpreted as indicative of relative trends rather than definitive evidence of causal relationships.

Bar plot showing mean SHAP values across all three organ models, indicating the relative importance of features in ANN predictions. Top predictors: mean dose (0.28), V30 (0.17), patient age (0.13), V50 (0.11), maximum dose (0.09), gEUD (0.08), V20 (0.06), sex (0.05), concurrent chemotherapy (0.03). DVH metrics capturing mid-to-high dose regions (V30, V50) consistently ranked higher than low-dose volumes, supporting QUANTEC emphasis on mean dose and intermediate dose-volume parameters for toxicity prediction. However, given severe class imbalance (3–5 events), feature importance estimates should be interpreted as exploratory observations only.

Sensitivity analysis with grade ≥1 endpoint

To evaluate the stability of NTCP model performance under conditions of higher event frequency and to explore the impact of severe class imbalance observed in the primary analysis, a sensitivity analysis was performed using a lower toxicity threshold (Grade ≥1). This redefinition resulted in substantially higher event rates across endpoints: xerostomia (n = 41, 80.4%), dysphagia (n = 38, 74.5%) and mucositis (n = 39, 76.5%). Using this alternative endpoint, traditional NTCP models demonstrated discrimination patterns broadly consistent with the primary analysis. LKB model performance remained moderate, with AUC values of 0.82 (95% CI: 0.72–0.92) for parotid toxicity, 0.79 (95% CI: 0.67–0.91) for laryngeal toxicity and 0.84 (95% CI: 0.74–0.94) for oral cavity toxicity. The preservation of rank-order discrimination across endpoints suggests that the underlying dose–response relationship is detectable even when less severe toxicity definitions are applied. However, Grade ≥1 toxicity generally corresponds to mild or transient symptoms that do not require clinical intervention, limiting the direct clinical relevance of this endpoint compared with Grade ≥2 toxicity. Accordingly, this sensitivity analysis should be interpreted as supportive evidence for biological plausibility and model stability rather than as a substitute for the primary, clinically meaningful endpoint. Together with the primary analysis, these findings indicate that while Grade ≥2 toxicity remains the appropriate clinical endpoint, lower-grade toxicity analyses can provide complementary insight into dose–response behaviour in small cohorts where event rates are limited.

Discussion

This institutional validation study establishes a reproducible computational framework for NTCP modeling in head and neck radiotherapy through dual-pipeline verification and derives population-specific model parameters for an Indian patient cohort. Traditional radiobiological models demonstrated acceptable discrimination following local calibration; however, the low number of Grade ≥ 2 toxicity events (3–5 per endpoint) imposes important constraints on statistical stability and precludes definitive conclusions regarding the added value of ML–based augmentation. Accordingly, all ML findings are interpreted within a hypothesis-generating framework.

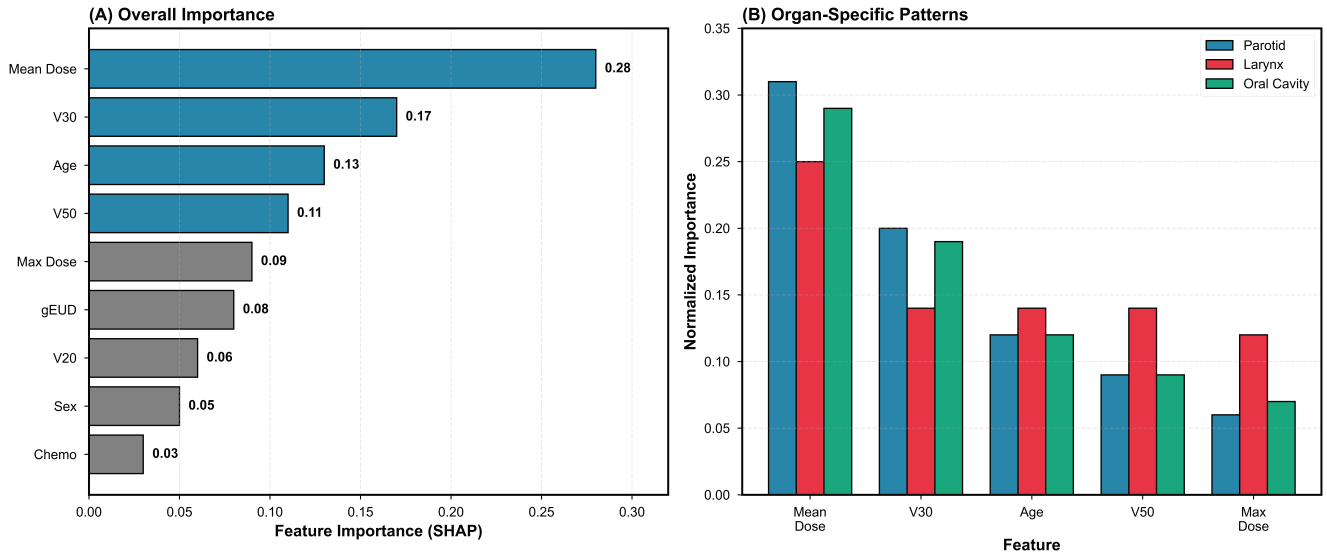

Figure 6. Feature importance from SHAP analysis.

Computational validation and reproducibility

A key contribution of this work is the demonstration that independent computational implementations – MATLAB-based RBMODELv1 and a custom Python-based pipeline – produce highly concordant NTCP estimates, with a mean bias of 0.8% and narrow 95% limits of agreement (−1.9% to +3.5%). This finding directly addresses long-standing concerns regarding inter-software variability that have limited broader clinical adoption of NTCP models [64, 65]. Prior reports have documented discrepancies of up to 10%–15% in NTCP or EUD calculations across platforms, attributable to differences in dose–volume interpolation, numerical integration schemes and rounding conventions [28, 30, 65]. Our results indicate that when identical mathematical formulations and numerical conventions are implemented with sufficient precision, computational consistency is achievable across software environments. Such reproducibility is essential for (1) independent validation of published NTCP models, (2) quality assurance of clinical decision-support tools and (3) multi-institutional collaborative studies requiring standardised calculations across participating centers [66].

RBMODELv1, developed at our institution, provides a validated, transparent platform for NTCP estimation using established radiobiological models [45]. Cross-verification against an independent Python implementation based on widely used scientific computing libraries (NumPy, SciPy) further enhances transparency and auditability, which are increasingly important for regulatory-compliant analyses and prospective clinical studies. On this basis, we recommend the adoption of similar dual-validation strategies by institutions seeking to implement NTCP-based treatment plan evaluation in routine practice.

Population-specific parameter calibration

Institution-specific LKB parameter estimates differed systematically from QUANTEC-recommended values across organs, with lower TD50 estimates observed for the parotid glands (34.1 Gy versus QUANTEC 39.0 Gy), oral cavity (48.5 Gy versus 50.0 Gy) and larynx (43.6 Gy versus 46.0 Gy). Differences of approximately 4–5 Gy are potentially clinically relevant and likely reflect the combined influence of population-specific, methodological and treatment-related factors rather than a single causal mechanism.

One contributing factor may be inter-population biological variability. Genetic polymorphisms in DNA damage response and repair pathways have been reported at higher prevalence in South Asian populations and have been associated with increased radiation sensitivity in prior studies [18, 19, 67]. In addition, methodological differences in toxicity assessment may influence estimated dose–response parameters. In the present study, toxicity grading was informed by patient-reported outcomes using EORTC quality-of-life instruments, whereas QUANTEC parameters were largely derived from physician-reported LENT–SOMA scales. Patient-reported measures have been shown to capture symptom onset earlier and at lower dose levels, which may contribute to lower apparent TD50 estimates [68]. Differences in treatment technique and dose distribution also warrant consideration. A higher proportion of patients in the present cohort were treated with VMAT (55%) compared with cohorts contributing to QUANTEC analyses (approximately 30%), potentially resulting in altered spatial dose patterns and volume-effect relationships that are not fully captured by historical parameters [69].

Consistent with these observations, population-specific NTCP parameter variation has been reported in other Asian cohorts. Single-institution Japanese IMRT studies have demonstrated parotid TD50 estimates lower than those commonly reported in Western series [70]. Similarly, Chinese nasopharyngeal cancer cohorts, which model *delivered* rather than planning dose, demonstrate dose–response behavior that differs from older Western datasets [71], with several studies from Chinese populations reporting steeper dose–response slopes than QUANTEC-derived estimates [72]. Systematic reviews further confirm substantial inter-study variability in NTCP model parameters across cohorts, including Asian NPC populations [73].

Collectively, these findings underscore the dependence of NTCP model parameters on endpoint definition, cohort characteristics and treatment paradigm and they reinforce QUANTEC's recommendation for institutional validation and local calibration of dose–response parameters rather than universal adoption of published values [15, 25, 61, 70, 71, 74].

Performance of traditional NTCP models

Traditional radiobiological models demonstrated moderate to strong rank-order associations with observed toxicity grades (ρ^2 = 0.37–0.63), indicating effective relative risk stratification despite uncertainty in absolute probability calibration. This observation is consistent with systematic reviews showing that NTCP models generally perform better at distinguishing higher-risk from lower-risk patients than at providing precise absolute risk estimates without local calibration [7, 75, 76]. In the context of treatment planning optimisation and comparative plan evaluation, such relative risk ranking may be sufficient to support clinical decision-making [77]. In contrast, applications requiring individualised risk communication or patient selection for clinical trials necessitate calibrated absolute probability estimates [78]. Among the evaluated formulations, the LKB model demonstrated consistently higher discrimination than the EUD- and RS-based approaches across endpoints. This finding may reflect the greater flexibility afforded by the LKB model's parameterisation, which separately accounts for dose–response steepness (m), volume dependence (n) and overall sensitivity (TD50). In comparison, the lower performance of the RS model for parotid toxicity (AUC 0.59) highlights the importance of matching model assumptions to underlying organ architecture, as application of serial-organ formulations to predominantly parallel structures may reduce predictive performance [12, 13].

ML: promises and pitfalls in small datasets

Our exploratory ML analysis was designed to generate hypotheses rather than to meet standards for clinical model validation. The limited number of Grade ≥2 toxicity events (3–5 per endpoint) constrains the stability of multivariable prediction and necessitates cautious interpretation of model performance [41, 42, 58, 79]. Established guidance based on EPV considerations suggests that substantially larger event counts are required for reliable multivariable ML, and model performance estimates derived from small event numbers are known to be statistically unstable [41, 42]. Simulation studies have further demonstrated that ML models trained on very small numbers of events can exhibit wide variability in discrimination metrics, including large fluctuations in AUC estimates across resampling procedures [55, 80]. Further, in this study, the extreme gradient boosting (XGBoost) results, with near-perfect training performance but substantially lower test set discrimination, exemplify the high risk of overfitting when complex, flexible algorithms are applied to small datasets with limited events. This negative finding underscores that ML approaches are not inherently superior and require rigorous validation with adequate sample sizes before any clinical consideration.

Within these limitations, the observed numerical differences between ML and traditional models are best interpreted as hypothesis-generating. When comparing ANN against the best-performing traditional model (LKB) for each organ, differences were modest across all endpoints: parotid glands (ΔAUC +0.04), oral cavity (ΔAUC +0.04) and mixed-architecture larynx (ΔAUC +0.03), no statistically significant advantage for ML was observed for any organ architecture. Parallel organs are characterised by functional subunit redundancy, spatial dose heterogeneity and compensatory mechanisms that may introduce non-linear dose–response relationships not fully captured by parametric radiobiological models [81]. In principle, ML approaches, which are not constrained by predefined functional forms, may be capable of modeling such complex interactions [31, 82, 83]. However, confirmation of this hypothesis requires substantially larger cohorts, with recommended event counts of at least 50 per endpoint, preferably through multi-institutional data pooling [63, 66].

Recent large-scale studies provide context. Dean *et al* [84] (*n* = 130, 45 events) found ML advantages for oral mucositis prediction (AUC 0.82 versus 0.71 traditional). El Naqa *et al* [31] used multivariable ML approaches for radiotherapy outcomes modeling, noting organ-architecture-dependent differences, supporting the organ-architecture hypothesis. However, a systematic review of ML-based toxicity prediction in radiotherapy concluded that consistent superiority of ML over conventional models remains undemonstrated, primarily due to inadequate external validation, heterogeneous study designs and small sample sizes across the published literature [38].

Clinical implementation barriers

Despite the acceptable performance of traditional radiobiological models in this and prior studies, several practical factors continue to limit the routine clinical adoption of NTCP-based evaluation. These include the computational complexity of model implementation, which often requires specialised software or in-house expertise; the absence of universally accepted parameter sets, necessitating institutional validation and recalibration; limited prospective evidence demonstrating that NTCP-optimised treatment planning improves patient outcomes compared with standard DVH-constraint-based approaches [64, 85] and the lack of widespread integration of NTCP models into regulatory-approved commercial treatment planning systems. Collectively, these considerations help explain why, despite more than four decades of development since Lyman's foundational work [8], NTCP models have remained more commonly used in research and planning studies than in routine clinical workflows.

Limitations and future directions

In addition to the low incidence of Grade ≥2 toxicity events, several methodological considerations should be acknowledged when interpreting these findings. The retrospective study design may introduce selection and information biases, and the single-institution setting limits generalisability to other populations and treatment environments. Follow-up duration was heterogeneous (median 8 months, range 3–24 months), which may reduce sensitivity for detecting late-onset toxicities. Baseline patient-reported outcome measures were not uniformly available, limiting the ability to quantify treatment-related changes over time. Furthermore, the absence of validation in independent external cohorts constrains assessment of model transportability, and the use of binary toxicity endpoints, while clinically practical, does not fully exploit time-to-event information that could better accommodate censoring and temporal patterns of toxicity [86]. The selection of Grade ≥2 as the binary toxicity threshold was guided by clinical relevance, as Grade ≥2 events typically represent functionally significant complications warranting clinical intervention; however, this threshold substantially reduced the number of positive events per endpoint (3–5 events), thereby limiting statistical power. The sensitivity analysis at Grade ≥1 yielded consistent directional findings with improved event counts, providing supporting biological plausibility for the observed dose–response relationships while acknowledging that Grade 1 events may include clinically insignificant toxicities. Future investigations should aim to address these considerations through coordinated multi-institutional data aggregation with sufficient event counts per endpoint to support robust model development and validation [66]. Integration of genetic and molecular biomarkers, such as single-nucleotide polymorphisms in DNA repair pathways, alongside dosimetric features, may enhance individualised toxicity prediction [87]. Spatially resolved dose analyses using voxel-based or regional approaches could further refine structure–function relationships [88]. Prospective studies comparing NTCP-guided treatment planning with conventional DVH-based strategies are needed to determine clinical benefit, and federated learning frameworks offer a promising pathway for collaborative model training across institutions while preserving data privacy [89]. Furthermore, variation in organ-at-risk dose and biologically effective dose

across different α/β values for conventional, moderate and ultra-hypofractionated treatment regimens represents an important consideration when evaluating the transferability of NTCP model parameters derived from standard fractionation cohorts, such as the present one, to emerging hypofractionation schedules [90].

An additional methodological consideration concerns the selective application of the RS model – applied only to the parotid gland and not to larynx or oral cavity. This deliberate choice was guided by historical precedent (QUANTEC parameters available for comparison) and the opportunity to empirically confirm that serial models underperform for parallel organs (AUC 0.59), reinforcing the organ-architecture theme of this study. The absence of established RS parameters for larynx/dysphagia and oral cavity/mucositis endpoints precluded their evaluation. Readers should therefore interpret the parotid RS results primarily as a historical reference and hypothesis test rather than a clinically validated model. This asymmetry is inherent to our study design and discussed in Table 2 notes.

Conclusion

This study establishes a reproducible dual-pipeline framework for NTCP modeling and derives population-specific parameters for an Indian head and neck cancer cohort. Following local calibration, traditional radiobiological models demonstrated robust toxicity risk stratification, with the LKB model achieving the highest discrimination (AUC 0.88–0.91) and institution-specific parameters differing from QUANTEC values. Exploratory ML analyses suggest a potential organ-architecture–dependent performance pattern, warranting validation in adequately powered multi-institutional cohorts. The validated framework provides a foundation for future prospective studies assessing the clinical value of NTCP-guided treatment planning.

List of abbreviations

3DCRT, three-dimensional conformal radiotherapy; AJCC, American Joint Committee on Cancer; ANN, artificial neural network; AUC, area under the curve; CI, confidence interval; CT, computed tomography; CTCAE, Common Terminology Criteria for Adverse Events; DVH, dose-volume histogram; EORTC, European Organisation for Research and Treatment of Cancer; EPV, events per variable; EUD, equivalent uniform dose; gEUD, generalised equivalent uniform dose; LKB, Lyman–Kutcher–Burman; ML, machine learning; NTCP, normal tissue complication probability; OAR, organ at risk; QUANTEC, Quantitative Analyses of Normal Tissue Effects in the Clinic; RBMODv1, RBMODELv1(Radiobiological Model Version 1); ROC, receiver operating characteristic; RS, relative seriality; SMOTE, Synthetic Minority Over-sampling Technique; VMAT, volumetric modulated arc therapy; XGBoost, extreme gradient boosting

Conflicts of interest

The authors declare no conflicts of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Ethical approval

This study involved retrospective analysis of de-identified clinical and dosimetric data and had no impact on patient management. In accordance with institutional policy, formal ethical approval was exempted by the ethics committee (Banaras Hindu University).

Use of AI-assisted tools

AI-assisted tools (Claude, Anthropic) were used for language editing and manuscript formatting during the revision process. All scientific content, data analyses, interpretations and conclusions were performed solely by the listed authors, who take full responsibility for the accuracy and integrity of the work.

Author contributions

KM: Conceptualisation, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualisation. AM: Methodology, Validation, Writing – Review & Editing, Supervision. AV: Resources, Writing – Review & Editing, Supervision. GP: Conceptualisation, Methodology, Software, Validation, Writing – Review & Editing, Supervision, Project Administration. All authors reviewed and approved the final manuscript.

Reference

1. Nutting CM, Morden JP, and Harrington KJ, *et al* (2011) **Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial** *Lancet Oncol* **12**(2) 127–136 https://doi.org/10.1016/S1470-2045(10)70290-4 PMID: 21236730 PMCID: 3033533
2. Gupta T, Agarwal J, and Jain S, *et al* (2012) **Three-dimensional conformal radiotherapy (3D-CRT) versus intensity modulated radiation therapy (IMRT) in squamous cell carcinoma of the head and neck: a randomized controlled trial** *Radiotherapy Oncol* **104**(3) 343–348 https://doi.org/10.1016/j.radonc.2012.07.001
3. Ghosh–Laskar S, Yathiraj PH, and Dutta D, *et al* (2016) **Prospective randomized controlled trial to compare 3-dimensional conformal radiotherapy to intensity-modulated radiotherapy in head and neck squamous cell carcinoma: long-term results** *Head Neck* **38**(S1) E1481–E1487 https://doi.org/10.1002/hed.24263
4. Mayo C, Martel MK, and Marks LB, *et al* (2010) **Radiation dose–volume effects of optic nerves and chiasm** *Int J Radiat Oncol Biol Phys* **76**(3) S28–S35 https://doi.org/10.1016/j.ijrobp.2009.07.1753 PMID: 20171514
5. Emami B, Lyman J, and Brown A, *et al* (1991) **Tolerance of normal tissue to therapeutic irradiation** *Int J Radiat Oncol Biol Phys* **21**(1) 109–122 https://doi.org/10.1016/0360-3016(91)90171-Y PMID: 2032882
6. Marks LB, Yorke ED, and Jackson A, *et al* (2010) **Use of normal tissue complication probability models in the clinic** *Int J Radiat Oncol Biol Phys* **76**(3) S10–S19 https://doi.org/10.1016/j.ijrobp.2009.07.1754 PMID: 20171502 PMCID: 4041542
7. Kierkels RGJ, Korevaar EW, and Steenbakkers RJHM, *et al* (2014) **Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer radiotherapy produces clinically acceptable treatment plans** *Radiotherapy Oncol* **112**(3) 430–436 https://doi.org/10.1016/j.radonc.2014.08.020
8. Lyman JT (1985) **Complication probability as assessed from dose-volume histograms** *Radiat Res Suppl* **8** S13–S19 https://doi.org/10.2307/3583506 PMID: 3867079

9. Kutcher GJ, Burman C, and Brewster L, *et al* (1991) **Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations** *Int J Radiat Oncol Biol Phys* **21**(1) 137–146 https://doi.org/10.1016/0360-3016(91)90173-2 PMID: 2032884

10. Niemierko A (1997) **Reporting and analyzing dose distributions: a concept of equivalent uniform dose** *Med Phys* **24**(1) 103–110 https://doi.org/10.1118/1.598063 PMID: 9029544

11. Choi B and Deasy JO (2002) **The generalized equivalent uniform dose function as a basis for intensity-modulated treatment planning** *Phys Med Biol* **47**(20) 3579–3589 https://doi.org/10.1088/0031-9155/47/20/302 PMID: 12433121

12. Källman P, Ågren A, and Brahme A (1992) **Tumour and normal tissue responses to fractionated non-uniform dose delivery** *Int J Radiat Biol* **62**(2) 249–262 https://doi.org/10.1080/09553009214552071 PMID: 1355519

13. Jackson A, Kutcher GJ, and Yorke ED (1993) **Probability of radiation-induced complications for normal tissues with parallel architecture subject to non-uniform irradiation** *Med Phys* **20**(3) 613–625 https://doi.org/10.1118/1.597056 PMID: 8350812

14. Bentzen SM, Constine LS, and Deasy JO, *et al* (2010) **Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues** *Int J Radiat Oncol Biol Phys* **76**(3 Suppl) S3–S9 https://doi.org/10.1016/j.ijrobp.2009.09.040 PMID: 20171515 PMCID: 3431964

15. Deasy JO, Moiseenko V, and Marks L, *et al* (2010) **Radiotherapy dose-volume effects on salivary gland function** *Int J Radiat Oncol Biol Phys* **76**(3 Suppl) S58–S63 https://doi.org/10.1016/j.ijrobp.2009.06.090 PMID: 20171519 PMCID: 4041494

16. Rancati T, Schwarz M, and Allen AM, *et al* (2010) **Radiation dose-volume effects in the larynx and pharynx** *Int J Radiat Oncol Biol Phys* **76**(3 Suppl) S64–S69 https://doi.org/10.1016/j.ijrobp.2009.03.079 PMID: 20171520 PMCID: 2833104

17. Werner-Wasik M, Yorke E, and Deasy J, *et al* (2010) **Radiation dose-volume effects in the esophagus** *Int J Radiat Oncol Biol Phys* **76**(3 Suppl) S86–S93 https://doi.org/10.1016/j.ijrobp.2009.05.070 PMID: 20171523 PMCID: 3587781

18. Barnett GC, West CML, and Dunning AM, *et al* (2009) **Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype** *Nat Rev Cancer* **9**(2) 134–142 https://doi.org/10.1038/nrc2587 PMID: 19148183 PMCID: 2670578

19. Dylawerska A, Barczak W, and Wegner A, *et al* (2017) **Association of DNA repair genes polymorphisms and mutations with increased risk of head and neck cancer: a review** *Med Oncol* **34**(12) 197 https://doi.org/10.1007/s12032-017-1057-4 PMID: 29143133 PMCID: 5688183

20. Patil N, Ma N, and Mair M, *et al* (2024) **Oral cavity cancers: ethnic differences in radiotherapy outcomes in a majority South Asian Leicester Community** *Clin Oncol* **36**(5) 300–306 https://doi.org/10.1016/j.clon.2024.02.010

21. Muwonge R, Ramadas K, and Sankila R, *et al* (2008) **Role of tobacco smoking, chewing and alcohol drinking in the risk of oral cancer in Trivandrum, India: a nested case-control design using incident cancer cases** *Oral Oncol* **44**(5) 446–454 https://doi.org/10.1016/j.oraloncology.2007.06.002

22. Awan KH and Patil S (2016) **Association of smokeless tobacco with oral cancer - evidence from the South Asian studies: a systematic review** *J Coll Physicians Surg Pak* **26**(9) 775–780 PMID: 27671184

23. Murthy V, Calcuttawala A, and Chadha K, *et al* (2017) **Human papillomavirus in head and neck cancer in India: current status and consensus recommendations** *South Asian J Cancer* **6**(3) 93–98 https://doi.org/10.4103/sajc.sajc_96_17 PMID: 28975111 PMCID: 5615888

24. Wei WI and Sham JST (2005) **Nasopharyngeal carcinoma** *Lancet* **365**(9476) 2041–2054 https://doi.org/10.1016/S0140-6736(05)66698-6 PMID: 15950718

25. Lee AWM, Ma BBY, and Ng WT, *et al* (2015) **Management of nasopharyngeal carcinoma: current practice and future perspective** *J Clin Oncol* **33**(29) 3356–3364 https://doi.org/10.1200/JCO.2015.60.9347 PMID: 26351355

26. Dandekar M, Tuljapurkar V, and Dhar H, *et al* (2017) **Head and neck cancers in India** *J Surgical Oncol* **115**(5) 555–563 https://doi.org/10.1002/jso.24545

27. Zami Z, Pachuau L, and Bawihtlung Z, *et al* (2024) **Treatment regimens and survival among patients with head and neck squamous cell carcinoma from Mizo tribal population in northeast India – a single centre, retrospective cohort study** *Lancet Regional Health - Southeast Asia* **24** 100377 https://doi.org/10.1016/j.lansea.2024.100377

28. Ebert MA, Haworth A, and Kearvell R, *et al* (2010) **Comparison of DVH data from multiple radiotherapy treatment planning systems** *Phys Med Biol* **55**(11) 337 https://doi.org/10.1088/0031-9155/55/11/N04

29. Walker LS and Byrne JP (2025) **Clinical impact of DVH uncertainties** *Med Dosimetry* **50**(1) 1–7 https://doi.org/10.1016/j.meddos.2024.06.002

30. Pandu B, Khanna D, and Palanisamy M, *et al* (2025) **A narrative review: dose calculation algorithms used in external beam radiotherapy planning systems** *Therapeutic Radiol Oncol* **9**(2025) [https://doi.org/10.21037/tro-24-10]

31. El Naqa I, Bradley J, and Blanco AI, *et al* (2006) **Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors** *Int J Radiat Oncol Biol Phys* **64**(4) 1275–1286 https://doi.org/10.1016/j.ijrobp.2005.11.022 PMID: 16504765

32. Kang J, Schwartz R, and Flickinger J, *et al* (2015) **Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective [Internet]** *Int J Radiat Oncol Biol Phys* **93**(5) 1127–1135 [https://www.sciencedirect.com/science/article/pii/S0360301615030783] https://doi.org/10.1016/j.ijrobp.2015.07.2286 PMID: 26581149

33. Bang C, Bernard G, and Le WT, *et al* (2023) **Artificial intelligence to predict outcomes of head and neck radiotherapy** *Clin Transl Radiat Oncol* **39** 100590 [https://doi.org/10.1016/j.ctro.2023.100590] PMID: 36935854 PMCID: 10014342

34. Volpe S, Pepa M, and Zaffaroni M, *et al* (2021) **Machine learning for head and neck cancer: a safe bet?—a clinically oriented systematic review for the radiation oncologist** *Front Oncol* **11** 772663 https://doi.org/10.3389/fonc.2021.772663

35. Lambin P, Van Stiphout RGPM, and Starmans MHW, *et al* (2013) **Predicting outcomes in radiation oncology —multifactorial decision support systems** *Nat Rev Clin Oncol* **10**(1) 27–40 https://doi.org/10.1038/nrclinonc.2012.196

36. Wang Y, Li W, and Xu N, *et al* (2025) **Artificial intelligence in radiobiology: bridging mechanisms and data analysis** *Radiat Med Prot* **6**(6) 301–311 https://doi.org/10.1016/j.radmp.2025.12.005

37. Gardner LL, Thompson SJ, and O'Connor JD, *et al* (2024) **Modelling radiobiology** *Phys Med Biol* **69**(18) 18TR01 https://doi.org/10.1088/1361-6560/ad70f0

38. Isaksson LJ, Pepa M, and Zaffaroni M, *et al* (2020) **Machine learning-based models for prediction of toxicity outcomes in radiotherapy** *Front Oncol* **10** 790 https://doi.org/10.3389/fonc.2020.00790 PMID: 32582539 PMCID: 7289968

39. El Naqa I, Ruan D, and Valdes G, *et al* (2018) **Machine learning and modeling: data, validation, communication challenges** *Med Phys* **45**(10) e834–e840 https://doi.org/10.1002/mp.12811 PMID: 30144098 PMCID: 6181755

40. Hanley JA and Mcneil BJ (1982) **The meaning and use of the area under a receiver operating characteristic (ROC) curve** *Radiology* **143**(1) 29–36 https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747

41. Peduzzi P, Concato J, and Kemper E, *et al* (1996) **A simulation study of the number of events per variable in logistic regression analysis** *J Clin Epidemiol* **49**(12) 1373–1379 https://doi.org/10.1016/S0895-4356(96)00236-3 PMID: 8970487

42. Vittinghoff E and Mcculloch CE (2007) **Relaxing the rule of ten events per variable in logistic and Cox regression** *Am J Epidemiol* **165**(6) 710–718 https://doi.org/10.1093/aje/kwk052

43. Riley RD, Ensor J, and Snell KIE, *et al* (2020) **Calculating the sample size required for developing a clinical prediction model** *BMJ* **368** 441 https://doi.org/10.1136/bmj.m441

44. Brouwer CL, Steenbakkers RJHM, and Bourhis J, *et al* (2015) **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* **117**(1) 83–90 https://doi.org/10.1016/j.radonc.2015.07.041 PMID: 26277855

45. Patel G, Mandal A, and Bharati A, *et al* (2022) **Development and validation of an indigenous, radiobiological model-based tumor control probability and normal tissue complication probability estimation software for routine plan evaluation in clinics** *J Cancer Res Ther* **18**(6) 1697–1705 https://doi.org/10.4103/jcrt.JCRT_330_20 PMID: 36412432

46. Chen T and Guestrin C (2016) **XGBoost: a scalable tree boosting system [Internet]** *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco: ACM) 785–794 pp [https://doi.org/10.1145/2939672.2939785]

47. Chawla NV, Bowyer KW, and Hall LO, *et al* (2002) **SMOTE: synthetic minority over-sampling technique** *jair* **16** 321–357 https://doi.org/10.1613/jair.953

48. Fernández A, García S, and Galar M, *et al* (2018) *Learning from Imbalanced Data Sets* [Internet] (Cham: Springer International Publishing) [http://link.springer.com/10.1007/978-3-319-98074-4] Date accessed: 06/02/26

49. Varma S and Simon R (2006) **Bias in error estimation when using cross-validation for model selection** *BMC Bioinf* **7** 91 https://doi.org/10.1186/1471-2105-7-91

50. DeLong ER, DeLong DM, and Clarke-Pearson DL (1988) **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach** *Biometrics* **44**(3) 837–845 https://doi.org/10.2307/2531595 PMID: 3203132

51. Saito T and Rehmsmeier M (2015) **The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets** *PLoS One* **10**(3) 118432 https://doi.org/10.1371/journal.pone.0118432

52. Hauke J and Kossowski T (2011) **Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data** *Quaest Geogr* **30**(2) 87–93 [https://doi.org/10.2478/v10117-011-0021-1]

53. Armstrong RA (2014) **When to use the Bonferroni correction** *Ophthalmic Physiol Opt* **34**(5) 502–508 https://doi.org/10.1111/opo.12131 PMID: 24697967

54. Bland JM and Altman DG (1986) **Statistical methods for assessing agreement between two methods of clinical measurement** *Lancet* **1**(8476) 307–310 https://doi.org/10.1016/S0140-6736(86)90837-8 PMID: 2868172

55. Steyerberg EW, Harrell FE, and Borsboom GJ, *et al* (2001) **Internal validation of predictive models: efficiency of some procedures for logistic regression analysis** *J Clin Epidemiol* **54**(8) 774–781 https://doi.org/10.1016/S0895-4356(01)00341-9 PMID: 11470385

56. Altman DG and Royston P (2006) **The cost of dichotomising continuous variables** *BMJ* **332**(7549) 1080 https://doi.org/10.1136/bmj.332.7549.1080 PMID: 16675816 PMCID: 1458573

57. Wasserstein RL and Lazar NA (2016) **The ASA statement on p-values: context, process, and purpose** *Am Statistician* **70**(2) 129–133 https://doi.org/10.1080/00031305.2016.1154108

58. Japkowicz N and Stephen S (2002) **The class imbalance problem: a systematic study** *Intell Data Anal* **6**(5) 429–449 https://doi.org/10.3233/IDA-2002-6504

59. Athmakoor NK, Potaraju S, and Ardha A, *et al* (2021) **A comparative dosimetric analysis of volumetric modulated Arc therapy (VMAT) versus intensity-modulated radiation therapy (IMRT) for head and neck cancer in an Indian population** *Cureus* **17**(12) e100283 [https://doi.org/10.7759/cureus.100283]

60. Ho KF, Farnell DJJ, and Routledge JA, *et al* (2010) **Comparison of patient-reported late treatment toxicity (LENT-SOMA) with quality of life (EORTC QLQ-C30 and QLQ-H&N35) assessment after head and neck radiotherapy** *Radiother Oncol* **97**(2) 270–275 https://doi.org/10.1016/j.radonc.2010.01.017 PMID: 20554338

61. Miah AB, Gulliford SL, and Clark CH, *et al* (2013) **Dose-response analysis of parotid gland function: what is the best measure of xerostomia?** *Radiother Oncol* **106**(3) 341–345 https://doi.org/10.1016/j.radonc.2013.03.009 PMID: 23566529

62. Venkatesh GH, Manjunath VB, and Mumbrekar KD, *et al* (2014) **Polymorphisms in radio-responsive genes and its association with acute toxicity among head and neck cancer patients** *PLoS One* **9**(3) e89079 https://doi.org/10.1371/journal.pone.0089079 PMID: 24594932 PMCID: 3942321

63. Collins GS, Ogundimu EO, and Altman DG (2016) **Sample size considerations for the external validation of a multivariable prognostic model: a resampling study** *Stat Med* **35**(2) 214–226 https://doi.org/10.1002/sim.6787 PMCID: 4738418

64. Klein EE, Drzymala RE, and Purdy JA, *et al* (2005) **Errors in radiation oncology: a study in pathways and dosimetric impact** *J Appl Clin Med Phys* **6**(3) 81–94 https://doi.org/10.1120/jacmp.v6i3.2105 PMID: 16143793 PMCID: 5723492

65. Chang JH, Gehrke C, and Prabhakar R, *et al* (2016) **RADBIOMOD: a simple program for utilising biological modelling in radiotherapy plan evaluation** *Phys Med* **32**(1) 248–254 https://doi.org/10.1016/j.ejmp.2015.10.091

66. Skripcak T, Belka C, and Bosch W, *et al* (2014) **Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets** *Radiother Oncol* **113**(3) 303–309 https://doi.org/10.1016/j.radonc.2014.10.001 PMID: 25458128 PMCID: 4648243

67. Talbot CJ, Tanteles GA, and Barnett GC, *et al* (2012) **A replicated association between polymorphisms near TNFα and risk for adverse reactions to radiotherapy** *Br J Cancer* **107**(4) 748–753 https://doi.org/10.1038/bjc.2012.290 PMID: 22767148 PMCID: 3419947

68. Trotti A, Colevas A, and Setser A, *et al* (2003) **CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment** *Semin Radiat Oncol* **13**(3) 176–181 https://doi.org/10.1016/S1053-4296(03)00031-6 PMID: 12903007

69. Veldeman L, Madani I, and Hulstaert F, *et al* (2008) **Evidence behind use of intensity-modulated radiotherapy: a systematic review of comparative clinical studies** *Lancet Oncol* **9**(4) 367–375 https://doi.org/10.1016/S1470-2045(08)70098-6 PMID: 18374290

70. Obinata K, Nakamura M, and Carrozzo M, *et al* (2014) **Changes in parotid gland morphology and function in patients treated with intensity-modulated radiotherapy for nasopharyngeal and oropharyngeal tumors** *Oral Radiol* **30**(2) 135–141 https://doi.org/10.1007/s11282-013-0151-3 PMID: 24817788 PMCID: 4009139

71. Lou J, Huang P, and Ma C, *et al* (2018) **Parotid gland radiation dose-xerostomia relationships based on actual delivered dose for nasopharyngeal carcinoma** *J Appl Clin Med Phys* **19**(3) 251–260 https://doi.org/10.1002/acm2.12327 PMID: 29664218 PMCID: 5978560

72. Zeng L, Liu S, and Deng H, *et al* (2019) **Comparing different normal tissue complication probability (NTCP) models of radiation-induced temporal lobe injury after intensity-modulated radiation therapy for nasopharyngeal carcinoma** *Int J Radiat Oncol Biol Phys* **105**(1) E398 https://doi.org/10.1016/j.ijrobp.2019.06.1579

73. Stieb S, Lee A, and Van Dijk LV, *et al* (2021) **NTCP modeling of late effects for head and neck cancer: a systematic review** *Int J Part Therapy* **8**(1) 95–107 https://doi.org/10.14338/20-00092

74. Ren G, Xu SP, and Du L, *et al* (2015) **Actual anatomical and dosimetric changes of parotid glands in nasopharyngeal carcinoma patients during intensity modulated radiation therapy** *Biomed Res Int* **2015** 670327 https://doi.org/10.1155/2015/670327 PMID: 25793202 PMCID: 4352457

75. Mohan R, Wu Q, and Manning M, *et al* (2000) **Radiobiological considerations in the design of fractionation strategies for intensity-modulated radiation therapy of head and neck cancers** *Int J Radiat Oncol Biol Phys* **46**(3) 619–630 https://doi.org/10.1016/S0360-3016(99)00438-1 PMID: 10701741

76. Nahum AE and Uzan J (2012) **(Radio)Biological optimization of external-beam radiotherapy** *Comput Math Methods Med* **2012** 329214 https://doi.org/10.1155/2012/329214 PMID: 23251227 PMCID: 3508750

77. Allen Li X, Alber M, and Deasy JO, *et al* (2012) **The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM** *Med Phys* **39**(3) 1386–1409 https://doi.org/10.1118/1.3685447 PMID: 22380372

78. Steyerberg EW, Moons KGM, and Van Der Windt DA, *et al* (2013) **Prognosis research strategy (PROGRESS) 3: prognostic model research** *PLoS Med* **10**(2) e1001381 https://doi.org/10.1371/journal.pmed.1001381 PMID: 23393430 PMCID: 3564751

79. Van Calster B, Mclernon DJ, and Van Smeden M, *et al* (2019) **Calibration: the Achilles heel of predictive analytics** *BMC Med* **17**(1) 230 https://doi.org/10.1186/s12916-019-1466-7 PMID: 31842878 PMCID: 6912996

80. Riley RD, Snell KI, and Ensor J, *et al* (2019) **Minimum sample size for developing a multivariable prediction model: pART II - binary and time-to-event outcomes** *Stat Med* **38**(7) 1276–1296 https://doi.org/10.1002/sim.7992 PMCID: 6519266

81. Jackson A, Marks LB, and Bentzen SM, *et al* (2010) **The lessons of QUANTEC: recommendations for reporting and gathering data on dose-volume dependencies of treatment outcome** *Int J Radiat Oncol Biol Phys* **76**(3 Suppl) S155–S160 https://doi.org/10.1016/j.ijrobp.2009.08.074 PMID: 20171512 PMCID: 2854159

82. El Naqa I, Pater P, and Seuntjens J (2012) **Monte Carlo role in radiobiological modelling of radiotherapy outcomes** *Phys Med Biol* **57**(11) R75–R97 https://doi.org/10.1088/0031-9155/57/11/R75 PMID: 22571871

83. Valdes G, Solberg TD, and Heskel M, *et al* (2016) **Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy** *Phys Med Biol* **61**(16) 6105–6120 https://doi.org/10.1088/0031-9155/61/16/6105 PMID: 27461154 PMCID: 5491385

84. Dean JA, Wong KH, and Welsh LC, *et al* (2016) **Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy** *Radiother Oncol* **120**(1) 21–27 https://doi.org/10.1016/j.radonc.2016.05.015 PMID: 27240717 PMCID: 5021201

85. Dawson LA, Ten Haken RK, and Lawrence TS (2001) **Partial irradiation of the liver** *Semin Radiat Oncol* **11**(3) 240–246 https://doi.org/10.1053/srao.2001.23485 PMID: 11447581

86. Harrell FE, Califf RM, and Pryor DB, *et al* (1982) **Evaluating the yield of medical tests** *JAMA* **247**(18) 2543–2546 https://doi.org/10.1001/jama.1982.03320430047030 PMID: 7069920

87. Andreassen CN, Schack LMH, and Laursen LV, *et al* (2016) **Radiogenomics - current status, challenges and future directions** *Cancer Lett* **382**(1) 127–136 https://doi.org/10.1016/j.canlet.2016.01.035 PMID: 26828014

88. Palma G, Monti S, and Conson M, *et al* (2020) **NTCP models for severe radiation induced dermatitis after IMRT or proton therapy for thoracic cancer patients** *Front Oncol* **10** 344 https://doi.org/10.3389/fonc.2020.00344 PMID: 32257950 PMCID: 7090153

89. Rieke N, Hancox J, and Li W, *et al* (2020) **The future of digital health with federated learning** *NPJ Digit Med* **3** 119 https://doi.org/10.1038/s41746-020-00323-1 PMID: 33015372 PMCID: 7490367

90. Sarkar B and Pradhan A (2025) **Characteristic variation of organs at risk dose and biologically effective dose as a function of different α/β values for conventional, moderate, and ultrahypofractionated breast cancer radiotherapy** *Radiat Environ Biophys* **65** 387–400 https://doi.org/10.1007/s00411-025-01165-9

Supplementary information

Table S1. Detailed performance metrics for extreme gradient boosting (XGBoost) models.

Organ	Training AUC	Test AUC	Test accuracy (%)	Test precision	Test recall	Test F1-score
Parotid	0.99	0.71	88.2	0.33	0.50	0.40
Larynx	0.98	0.67	86.4	0.40	0.67	0.50
Oral cavity	1.00	0.82	85.0	0.25	0.50	0.33

Notes: *Severe overfitting evident:* Near-perfect training performance (AUC 0.98–1.00) with substantially lower test set discrimination (AUC 0.67–0.82) despite L2 regularisation (λ = 1.0) and limited tree depth (max_depth = 3). *Class imbalance impact:* Test sets contained only 1–2 positive events per organ, rendering precision, recall and F1-score estimates statistically unstable. *EPV violation:* With only 3–5 events per endpoint and 20 candidate predictors, XGBoost models grossly violate minimum EPV guidelines (≥10 events per predictor), making overfitting inevitable. *Interpretation:* These results serve as a methodologically important negative finding, demonstrating that complex machine learning algorithms require substantially larger event counts for reliable performance. The overfitting pattern reinforces the cautionary narrative regarding ML in low-event settings. *Comparison:* ANN models (Table 4) showed better generalisation (test AUC 0.91–0.95) than XGBoost, suggesting that simpler architectures may be preferable when events are limited