SUPPLEMENTARY TABLES

Supplementary Table 1. Protocol-freeze and prediction audit trail

Artifact/event	Recorded time	SHA-256 or verification record
Statistical analysis plan	2026-08-05 15:16:36 +0800	ef12db81a8ccaf1327ed5e46b541d2cbebd3c5953cd8849a60ad50288d9cba45
Configuration registry	2026-08-05 15:16:55 +0800	4ea921c401b5e0e387582e39b00953740183a99f9c983cfa9b633d82608739bc
Final CatBoost artifact lock	2026-08-05 16:00:30 +0800	1bf94cabf45ef62c476f32061617aa81cd86ab27cc8d2590759785833a76c324
Second-institution prediction lock	2026-08-05 16:06:50 +0800	Prediction-file and upstream hashes recorded in lock file

The current fitting script had no second-institution input and recorded no access during that run. This narrow computational fact does not negate the investigators' historical exposure to the same cohort in earlier project work, which is disclosed in the manuscript.

Supplementary Table 2. Participant-flow and data-availability audit

Flow item	Development institution	Second institution
Records assessed at source	1651 (earlier cohort-assembly report)	Not retained in locked package
Not included in locked analytical cohort	616 total; reason-specific counts not retained	Not calculable
Final analytical cohort	1035	212
Documented recurrence within 3 years	331	67
Missing predictor/outcome values in final locked table	0	0
Excluded-patient covariates available for comparison	No	No
Individual recurrence/censoring/death/transfer times	Not retained	Not retained

The table reports only information supported by the archived analytical package and an earlier cohort-assembly report. Missing reason-specific counts were not reconstructed or imputed.

Supplementary Table 3. Prespecified predictor dictionary

Predictor	Type/unit	Coding	Source	Measurement time
Age	Continuous, years	Recorded value	Clinical record	Index record at definitive resection
Sex	Categorical	Female/Male	Clinical record	Index record
Anatomical location	Categorical	Upper extremity/lower extremity/trunk	Clinical record	Index record
Tumor depth	Binary	Superficial/deep	Operative and pathology record	Definitive resection
Tumor size	Continuous, cm	Maximum recorded diameter	Pathology/clinical record	Available at definitive resection
FNCLCC grade	Ordinal	1/2/3	Pathology record	Available at definitive resection
Surgical margin	Binary	Negative/positive	Pathology record	Definitive resection
Radiotherapy	Binary	No/yes	Treatment record	Status available in the index record
Chemotherapy	Binary	No/yes	Treatment record	Status available in the index record
Bone or neurovascular invasion	Binary	Absent/present	Imaging, operative, and pathology record	Available at definitive resection
Unplanned surgery	Binary	No/yes	Referral and operative history	Available at definitive resection
Symptom duration	Continuous, months	Recorded duration	Clinical history	Before definitive resection

The same analysis mappings were applied to both institutions. All model inputs were required to be available in the index record; later follow-up information was not used as input. FNCLCC, French Federation of Cancer Centers Sarcoma Group.

Supplementary Table 4. Reconciliation of prespecified candidate configurations

Model family	Configurations
CatBoost	24
Elastic-net logistic regression	35
Extremely randomized trees	18
Full 12-variable logistic regression	1
Histogram gradient boosting	18
k-nearest neighbors	18
Shrinkage linear discriminant analysis	6
LightGBM	24
Multilayer perceptron	18
Random forest	18
Six-variable logistic regression	1
Gradient boosting	18
Radial-basis-function support vector machine	20
XGBoost	24
Total base-model configurations	243
Prespecified ensemble strategies	3
Total strategies compared after family tuning	17

The 243 configurations belong to 14 base-model families. The three ensembles are strategy-level combinations selected within nested resampling and are not additional configuration-grid cells.

Supplementary Table 5. Development-fitted logistic-regression coefficients

Term	Six-variable coefficient	Full 12-variable coefficient
(Intercept)	-2.729623	-2.100807
Age		-0.012915
GenderMale		0.146321
Location2		-0.638751
Location3		-0.083517
Depth2	-0.619904	-0.472581
Size	0.063839	0.084831
FNCLCC2	1.072890	1.165459
FNCLCC3	2.091615	2.310283
Margin1	1.292928	1.283542
Radiotherapy1		-0.058091
Chemotherapy1		-0.402918
Bone_Neurovascular_Invasion1	1.717287	1.694706
Unplanned_Surgery1	1.819967	1.790852
Symptom_Duration		0.001253

For each model, predicted probability = 1/[1 + exp(-linear predictor)]. Reference categories are Female, upper extremity, superficial depth, FNCLCC grade 1, negative margin, no radiotherapy, no chemotherapy, no bone or neurovascular invasion, and no unplanned surgery. Continuous variables retain the units listed in Supplementary Table 3.

Supplementary Table 6. Model card for the public research interface

Field	Specification
Model	CatBoost CFG-159; 300 trees, depth 7, learning rate 0.03, L2 leaf regularization 5, random strength 1
Version identifier	SHA-256 1bf94cabf45ef62c476f32061617aa81cd86ab27cc8d2590759785833a76c324
Target population	Surgically treated soft tissue sarcoma patients represented by the two tertiary-center analytical cohorts
Output	Binary 3-year recurrence-status probability conditional on ascertainable status
Intended use	Research reproducibility and independent evaluation; hypothesis generation for future surveillance studies
Non-intended use	Diagnosis, treatment selection, causal inference, autonomous surveillance scheduling, or direct patient care
Required inputs	All 12 prespecified fields; no missing-input strategy has been validated
Treatment-timing limitation	Radiotherapy and chemotherapy dates were not preserved; strict landmark availability cannot be verified
Known failure modes	Missing or miscoded inputs, new categories, distribution shift, different pathology or surveillance practices, and use outside tertiary sarcoma settings
Fairness	Subgroup fairness was not established; only 67 second-institution events limited reliable subgroup assessment
Privacy and security	No direct identifiers are requested, but no independent privacy, cybersecurity, retention, or regulatory review has been completed; identifiable information must not be entered
Clinical threshold	None established; outputs must not trigger a clinical action without prospective decision-impact evaluation

Supplementary Table 7. Distributional comparison between institutions

Variable	Level used for categorical maximum	Development	Second institution	Standardized mean difference
Age	continuous	47.78 ± 13.68	48.88 ± 13.57	-0.080
Bone_Neurovascular_Invasion	1	233 (22.5%)	49 (23.1%)	-0.014
CatBoost predicted probability	descriptive	0.312 ± 0.266	0.293 ± 0.230	0.078
Chemotherapy	0	793 (76.6%)	154 (72.6%)	0.091
Depth	1	307 (29.7%)	60 (28.3%)	0.030
FNCLCC	3	325 (31.4%)	72 (34.0%)	-0.055
Gender	Male	578 (55.8%)	122 (57.5%)	-0.034
Location	2	504 (48.7%)	95 (44.8%)	0.078
Margin	0	952 (92.0%)	190 (89.6%)	0.082
Radiotherapy	0	658 (63.6%)	110 (51.9%)	0.238
Size	continuous	5.95 ± 4.23	6.39 ± 4.12	-0.106
Symptom_Duration	continuous	24.66 ± 13.21	24.86 ± 13.84	-0.015
Unplanned_Surgery	0	822 (79.4%)	178 (84.0%)	-0.118

For categorical predictors, the displayed row is the level with the largest absolute standardized difference; the sign is development minus second institution. Development predicted probability is apparent after complete-cohort fitting and is presented descriptively, not as a validation estimate.

Supplementary Table 8. Complete repeated internal-evaluation performance

Strategy	AUC	Brier	Log loss	Cal. intercept	Cal. slope	O/E
CatBoost	0.8126 ± 0.0052	0.1584 ± 0.0017	0.4813 ± 0.0051	0.011 ± 0.030	1.070 ± 0.080	1.005 ± 0.015
Equal-weight top-3 ensemble	0.8134 ± 0.0043	0.1586 ± 0.0016	0.4801 ± 0.0038	0.018 ± 0.014	1.086 ± 0.044	1.009 ± 0.007
Brier-weighted top-3 ensemble	0.8133 ± 0.0043	0.1586 ± 0.0016	0.4801 ± 0.0038	0.018 ± 0.014	1.086 ± 0.044	1.009 ± 0.007
Gradient boosting	0.8144 ± 0.0031	0.1589 ± 0.0011	0.4814 ± 0.0026	0.007 ± 0.012	1.117 ± 0.019	1.004 ± 0.006
XGBoost	0.8124 ± 0.0035	0.1595 ± 0.0011	0.4820 ± 0.0031	0.016 ± 0.017	1.054 ± 0.027	1.008 ± 0.009
Logistic top-3 stack	0.8117 ± 0.0044	0.1608 ± 0.0023	0.4880 ± 0.0057	0.001 ± 0.021	0.962 ± 0.038	1.000 ± 0.010
Histogram gradient boosting	0.8092 ± 0.0026	0.1613 ± 0.0012	0.4855 ± 0.0037	0.025 ± 0.017	0.912 ± 0.018	1.012 ± 0.008
Random forest	0.8058 ± 0.0053	0.1620 ± 0.0024	0.4865 ± 0.0049	0.018 ± 0.006	1.134 ± 0.073	1.009 ± 0.003
RBF support vector machine	0.8005 ± 0.0045	0.1634 ± 0.0022	0.4932 ± 0.0055	0.001 ± 0.009	0.976 ± 0.051	1.001 ± 0.004
Extra Trees	0.8015 ± 0.0056	0.1637 ± 0.0030	0.4885 ± 0.0059	0.008 ± 0.007	1.067 ± 0.090	1.004 ± 0.003
Elastic-net logistic	0.7994 ± 0.0017	0.1640 ± 0.0009	0.4930 ± 0.0022	-0.004 ± 0.006	1.042 ± 0.011	0.998 ± 0.003
LightGBM	0.8023 ± 0.0048	0.1642 ± 0.0018	0.4953 ± 0.0073	0.033 ± 0.014	0.837 ± 0.035	1.015 ± 0.006
Multilayer perceptron	0.7955 ± 0.0087	0.1647 ± 0.0026	0.5028 ± 0.0142	0.059 ± 0.055	0.867 ± 0.119	1.030 ± 0.028
Full 12-variable logistic	0.7998 ± 0.0021	0.1647 ± 0.0011	0.4937 ± 0.0029	-0.006 ± 0.006	0.907 ± 0.013	0.997 ± 0.003
Shrinkage LDA	0.7974 ± 0.0021	0.1657 ± 0.0011	0.4968 ± 0.0028	0.000 ± 0.021	0.878 ± 0.015	1.000 ± 0.010
Six-variable logistic benchmark	0.7948 ± 0.0013	0.1663 ± 0.0005	0.4959 ± 0.0010	-0.002 ± 0.004	0.948 ± 0.009	0.999 ± 0.002
k-nearest neighbors	0.7629 ± 0.0077	0.1802 ± 0.0018	0.5437 ± 0.0167	0.299 ± 0.004	1.012 ± 0.178	1.195 ± 0.004

Values are mean ± SD across five correlated outer cycles. The SDs summarize resampling variability and are not inferential confidence intervals. Cal., calibration; LDA, linear discriminant analysis; O/E, observed-to-expected event ratio; RBF, radial basis function.

Supplementary Table 9. Global CatBoost feature attribution in the development cohort

Rank	Feature	Mean absolute SHAP value
1	FNCLCC grade	0.772
2	Bone or neurovascular invasion	0.478
3	Unplanned surgery	0.476
4	Tumor size	0.294
5	Anatomical location	0.229
6	Tumor depth	0.143
7	Surgical margin	0.135
8	Age	0.115
9	Symptom duration	0.090
10	Sex	0.055
11	Chemotherapy	0.044
12	Radiotherapy	0.020

SHAP values are model attributions on the CatBoost raw-output scale and do not represent causal effects. FNCLCC, French Federation of Cancer Centers Sarcoma Group; SHAP, SHapley Additive exPlanations.

Supplementary Table 10. Post hoc sensitivity analysis excluding treatment-status predictors

Measure	Estimate (95% confidence interval)
Area under the receiver operating characteristic curve	0.7984 (0.7291 to 0.8595)
Brier score	0.1640 (0.1358 to 0.1940)
Log loss	0.4935 (0.4231 to 0.5715)
Calibration intercept	0.0989 (-0.2773 to 0.5437)
Calibration slope	0.9665 (0.6968 to 1.3173)
Observed-to-expected ratio	1.0651 (0.8749 to 1.2588)
Paired contrast	**Estimate (95% confidence interval)**
Area under the curve difference, sensitivity minus locked	-0.0067 (-0.0193 to 0.0053)
Brier-score difference, sensitivity minus locked	0.0026 (-0.0019 to 0.0072)

The sensitivity model used the locked CFG-159 hyperparameters but excluded radiotherapy and chemotherapy. It was fitted on the complete development cohort without re-tuning on second-institution outcomes and did not replace the primary model.

Supplementary Table 11. Selected numerical decision-curve results in the second institution

Threshold probability	CatBoost net benefit	Six-variable logistic net benefit	Treat-all net benefit	Treat-none net benefit
0.05	0.284	0.285	0.280	0.000
0.10	0.249	0.256	0.240	0.000
0.15	0.222	0.216	0.195	0.000
0.20	0.199	0.198	0.145	0.000
0.25	0.168	0.181	0.088	0.000
0.30	0.150	0.152	0.023	0.000
0.40	0.110	0.108	-0.140	0.000
0.50	0.071	0.094	-0.368	0.000
0.60	0.045	0.068	-0.710	0.000

These values are exploratory. No threshold was clinically prespecified, and no row should be interpreted as a recommended surveillance cutoff.