Electronic Journal of General Medicine
2026, 23(6), em764
e-ISSN: 2516-3507
https://www.ejgm.co.uk/

MODESTUM

Original Article

OPEN ACCESS

A pilot machine learning study for preoperative prediction of pathological lymph node status in PIPAC-treated gastric cancer in Kazakhstan: A single-center study

Kathiresan Sivakumar [1,2], Aidana Rakhmankulova [3], Altay Kerimkulov [4], Tomiris Sarina [4], Aliya Kuanyshpekova [4], Abduzhappar Gaipov [5,6*]

[1] School of Computing and Artificial Intelligence, Nazarbayev University, Astana, KAZAKHSTAN
[2] Private Institution "Nazarbayev University Research Administration", Nazarbayev University, Astana, KAZAKHSTAN
[3] Department of Biomedical Sciences, School of Medicine, Nazarbayev University, Astana, KAZAKHSTAN
[4] Multidisciplinary Department, National Research Oncology Center, Astana, KAZAKHSTAN
[5] Department of Medicine, School of Medicine, Nazarbayev University, Astana, KAZAKHSTAN
[6] Department of Internal Medicine, Clinical Academic, University Medical Center, Astana, KAZAKHSTAN
***Corresponding Author:** abduzhappar.gaipov@nu.edu.kz

Citation: Sivakumar K, Rakhmankulova A, Kerimkulov A, Sarina T, Kuanyshpekova A, Gaipov A. A pilot machine learning study for preoperative prediction of pathological lymph node status in PIPAC-treated gastric cancer in Kazakhstan: A single-center study. Electron J Gen Med. 2026;23(6):em764. https://doi.org/10.29333/ejgm/19775

ARTICLE INFO

Received: 15 Sep. 2026

Accepted: 13 Oct. 2026

ABSTRACT

Pathological lymph-node status after neoadjuvant treatment and surgery (ypN) is prognostic in gastric cancer but unavailable before resection. We developed a pilot prediction model to estimate ypN risk from seven variables recorded before radical surgery in patients treated within a single-center PIPAC program; the timing of blood counts relative to neoadjuvant treatment could not be established. Of 62 patients, 47 with recorded ypN were analyzed (18 ypN-positive). Predictors were clinical T and N stage, tumor grade, age, body-mass index, white-cell count, and platelet count. Ridge logistic regression (primary), random forest, and Firth logistic regression were evaluated by cross-validation. Out-of-fold AUCs were 0.701 (95% confidence interval 0.539-0.851), 0.667 (0.488-0.831), and 0.703 (0.532-0.851), respectively; Brier scores were 0.227, 0.227, and 0.222 (prevalence-only reference 0.236). Performance was modest and imprecise; because the predictor panel was selected using the same cross-validation, estimates may be optimistic. These findings are an exploratory signal requiring independent validation.

Keywords: stomach neoplasms, lymphatic metastasis, machine learning, random forest, logistic models, pilot projects

INTRODUCTION

Gastric cancer is among the most common cancers in Kazakhstan, where it was the second most prevalent cancer type in 2014-2022 [1]; the 5-year survival rate is only 17.09%, largely because 36.61% of cases are diagnosed at stage III [2]. One of the most challenging patterns of progression is peritoneal dissemination, which carries substantial morbidity and poor survival, and systemic chemotherapy penetrates peritoneal deposits poorly. Pressurized intraperitoneal aerosol chemotherapy (PIPAC) is an emerging minimally invasive method in which chemotherapy is aerosolized under pressure during laparoscopy and delivered directly into the peritoneal cavity; a recent systematic review and meta-analysis suggested that it may improve histopathological response and survival in carefully selected patients, although further randomized evidence is required [3]. Because PIPAC can be repeated during treatment, it also allows sequential assessment of the peritoneal cavity [4]. Pathological lymph-node status after neoadjuvant treatment and surgery (ypN) is one of the strongest determinants of prognosis in gastric cancer and is not knowable with certainty until a pathologist examines the resected specimen [5]; residual nodal involvement (ypN-positive status) is an independent prognostic factor for overall and relapse-free survival after neoadjuvant chemotherapy and gastrectomy [6]. A model estimating ypN positivity from information available before resection could support risk stratification, counseling, and postoperative surveillance planning, provided that it uses no variable that becomes available only after surgery. In the present study the model is an exploratory research tool, not a substitute for pathological staging or a basis for treatment decisions.

Machine learning (ML) has been widely applied across medicine [7, 8], including to lymph-node metastasis (LNM) prediction in gastric cancer: from early artificial neural-network models [9, 10] to classifiers developed, and in several studies externally validated, in cohorts of hundreds to thousands of patients [11-15]; radiomics-, imaging-, and nomogram-based approaches [16-21]; ML models for LNM in gastric neuroendocrine neoplasms [22]; and a meta-analysis of 41 studies reporting a pooled test c-index of 0.811 [23]. Risk

Copyright © 2026 by Author/s and Licensed by Modestum. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

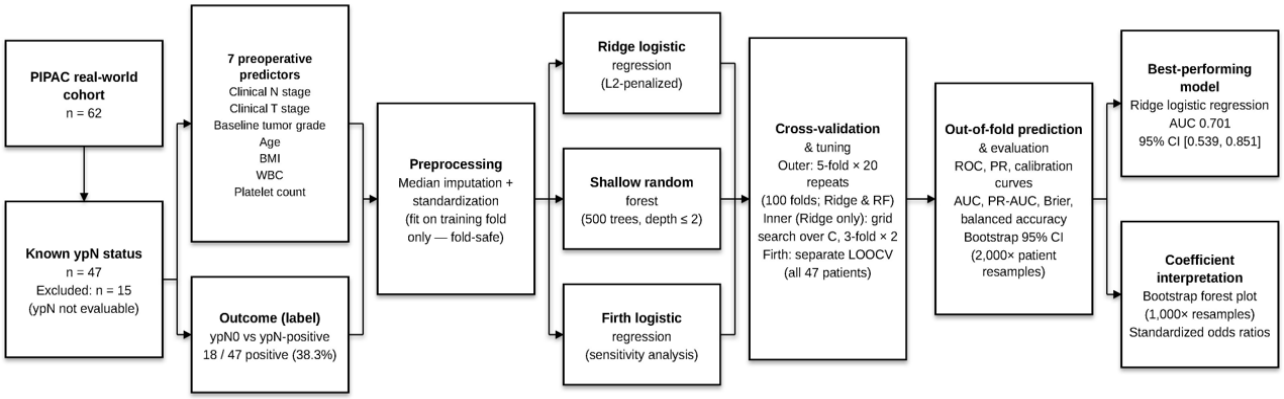


Figure 1. Overview of the study workflow (cross-validation and hyperparameter tuning applied to the ridge logistic regression and random forest models only; the Firth logistic regression sensitivity model was evaluated separately by leave-one-out cross-validation over all 47 patients) (created by the authors based on the study methodology)

factors for LNM in early gastric cancer were characterized earlier with logistic-regression models and risk scores [24-29], and ML was later applied to the same question [30]. ML has also been used for gastric cancer prognosis [31-33] and for predicting response to neoadjuvant chemotherapy [34]. Almost all of these studies used cohorts and predictor sets far larger than ours, mostly in early gastric cancer or at diagnosis, and none, to our knowledge, included patients treated within a PIPAC program.

This paper reports a small, single-center pilot study that asks a narrower question: using seven variables recorded before radical surgery and a cross-validation design in which no model sees a patient's own outcome during training, how well can ypN status be ranked in 47 PIPAC-treated patients? We follow the TRIPOD+AI guidance for prediction-model studies [35] and general principles for developing and validating clinical prediction models [36], adapted to the scale of a pilot feasibility analysis rather than a confirmatory one.

METHODS

Study Design and Participants

This is a retrospective, single-cohort, hypothesis-generating pilot study. The source dataset comprised 62 patients with gastric cancer treated between January 2025 and August 2026 within the single-center PIPAC program of National Research Oncology Center in Kazakhstan, all within a common clinical pathway. The parent program enrolled adults with histologically confirmed locally advanced gastric adenocarcinoma (cT3-4, cN0-3, M0), ECOG performance status 0-2, negative peritoneal cytology, and staging laparoscopy; key exclusion criteria were distant metastases, previous gastric-cancer-directed treatment, severe comorbidity, contraindications to laparoscopy/PIPAC, hypersensitivity to cisplatin or doxorubicin, and active infection. All patients received neoadjuvant treatment before planned resection. Adherence to these criteria cannot be verified patient by patient from the de-identified dataset (five patients were recorded as cT2 and one as cNx), and treatment was heterogeneous in completion of neoadjuvant chemotherapy, the interval between laparoscopy and surgery, and the number and timing of PIPAC cycles (see limitations). Of the 62 patients, 47 (75.8%) had a recorded pathological lymph-node stage (ypN) and were analyzed; the other 15 were excluded because the outcome was undefined for them: no surgery recorded at the time of data extraction (n = 8; one also with documented peritoneal-carcinomatosis progression); a documented gastrectomy with no ypN recorded in the available dataset (n = 4; the reason could not be established from the available records); biopsy only, with disease progression precluding resection (n = 2); and planned resection aborted because of positive peritoneal cytology at second-look laparoscopy (n = 1). The model therefore applies only to patients who undergo resection with available pathological assessment. No patient was excluded on the basis of a predictor value. The protocol of the parent prospective study has been published previously [4]; no separate prespecified protocol was prepared for this retrospective, exploratory analysis.

The dataset records one PIPAC procedure-detail entry per patient (median recorded duration 75 minutes, inter-quartile range [IQR] 66-81), which does not establish that every patient received exactly one PIPAC session; it does not record therapeutic-versus-prophylactic indication, neoadjuvant intent, the specific neoadjuvant regimen, or the number and timing of PIPAC cycles. Of the 62 patients, 54 (87.1%) completed all four planned cycles of neoadjuvant systemic chemotherapy and 8 (12.9%) had an incomplete course. Using the first recorded laparoscopy date as a proxy for the PIPAC date, the median interval to radical surgery among the 47 analyzed patients was 90 days (IQR 83-98.5, range 34-152); this is a proxy, not a confirmed PIPAC-to-surgery interval (further cohort details: S1 in **Appendix A**).

Figure 1 summarizes the overall study workflow, from cohort definition through predictor and outcome specification, fold-safe preprocessing, model training, and out-of-fold evaluation.

Predictors and Outcome Definition

The outcome was binary ypN status: ypN-positive if the recorded pathological nodal stage was greater than 0 (ypN1-ypN3) and ypN-negative if ypN0. Pathological ypTNM staging is routine standard histopathology after surgery, and ypN was abstracted from the pathology report. Seven predictors were used: clinical T stage (cT), clinical N stage (cN), tumor grade, age, body-mass index (BMI), white-cell count (WBC), and platelet count. In this paper, "preoperative" means recorded before radical surgery and before postoperative pathological assessment; no operative, postoperative, or outcome-derived variable was made available to any model, and the outcome

itself was never among the predictors. The dataset does not establish whether each predictor, particularly WBC and platelet count, was measured before the start of neoadjuvant treatment or between neoadjuvant treatment and surgery. Clinical-stage and grade fields carry no flag of a post-treatment measurement, and WBC and platelet count carry a documented timing flag for 8 of 62 patients (12.9%) noting that the sample was drawn after dexamethasone premedication rather than after chemotherapy, but no timing relative to neoadjuvant treatment is recorded for the remaining measurements. We therefore use "preoperative" only in the sense defined above, and we do not describe any predictor as "pretreatment" or "baseline". Adding the interval between the first recorded laparoscopy and radical surgery as an eighth candidate predictor did not improve discrimination (area under curve [AUC] 0.701 with vs. 0.701 without; paired bootstrap difference +0.000, 95% confidence interval [CI] -0.090 to +0.096), so it was not retained in the primary panel. Outcome assessment was performed as part of routine clinical care before development of the prediction model; therefore, pathologists were unaware of model-generated predictions. Formal blinding to individual clinical predictor values was not documented.

Data Preprocessing

Four of 47 patients (8.5%) had a missing or invalid tumor-grade code; these were imputed with the training-fold median within each cross-validation split. Predictors were standardized (zero mean, unit variance) using training-fold statistics only and then applied to the corresponding test fold, so that no information from a held-out patient influenced the imputation or scaling used to predict that patient. Variables were checked for missingness, invalid or out-of-range values, and consistency with the source coding. Clinical T stage, clinical N stage, and tumor grade were entered as ordinal numeric scores (cT coded 1-3 for cT2-cT4, cN coded 0-3, grade coded 1-4), which assumes equal spacing between adjacent categories; coding details and a categorical (dummy-coded) sensitivity analysis are given as S2 in **Appendix A**.

Model Training

Three modeling approaches were used:

1. **Ridge-penalized logistic regression (primary model):** An ℓ2-regularized logistic regression, with the regularization strength selected by an inner cross-validation grid search within each outer training fold (S3 in **Appendix A**). The cT, cN, and grade scores entered as linear ordinal codes; a categorical (dummy-coded) specification was evaluated as a sensitivity analysis and performed similarly (AUC 0.684 vs. 0.701; paired bootstrap difference -0.016, 95% CI -0.102 to +0.079; S2 in **Appendix A**).
2. **Depth-restricted random forest (comparator):** An ensemble of 500 trees, each restricted to a maximum depth of 2 and a minimum of 5 samples per leaf, with class-balanced sample weighting, chosen to limit overfitting given the small sample size.
3. **Firth (bias-reduced) logistic regression (sensitivity analysis):** A penalized-likelihood logistic regression evaluated by leave-one-out cross-validation, included because ordinary maximum-likelihood logistic regression can be unstable with a small number of outcome events per predictor. Because scikit-learn does not include Firth regression, it was fitted with the open-source Python package firthmodels (version 0.8.2; scikit-learn-compatible FirthLogisticRegression class; https://github.com/jzluo/firthmodels), run under Python 3.11; the rationale for this package and its verification against independent implementations are given as S4 in **Appendix A**.

All hyperparameters and preprocessing steps for ridge regression and the random forest were selected using only training-fold data within a repeated stratified outer cross-validation scheme (5 folds × 20 repeats, 100 folds), so that every patient's prediction came from a model that never saw that patient during training or hyperparameter selection; the Firth model followed the same principle under leave-one-out cross-validation. A single fixed random seed (20260924) controlled the fold assignments and the random-forest estimator, and table and figures were generated from the same out-of-fold predictions, so they are internally consistent by construction. Because the cohort contains only 47 patients and 18 events (about 2.6 events per candidate predictor), all models were kept deliberately simple, and ridge regression was the primary model.

Statistical Analysis

Discrimination was quantified using the out-of-fold area under the receiver operating characteristic (ROC) curve (AUC) and the area under the precision-recall curve (PR-AUC, equivalently average precision). Calibration was assessed visually in five quantile-defined bins (finer binning is unstable at this sample size) and by calibration-in-the-large and a logistic recalibration intercept and slope; overall probabilistic accuracy was summarized with the Brier score, including a scaled Brier score ($1\text{-Brier}_{model}/\text{Brier}_{reference}$) against the prevalence-only reference (S5 in **Appendix A**). Uncertainty was quantified by patient-level bootstrap resampling (2,000 resamples for AUC and Brier score; 1,000 for standardized coefficients), reporting the 2.5th and 97.5th percentiles as a 95% interval. A predictor-panel sensitivity analysis compared the five-variable set (cT, cN, grade, age, and BMI) with the seven-variable set that added WBC and platelet count, using identical patients and resampling splits. Patients with known and unknown ypN were compared with Fisher's exact test (the Fisher-Freeman-Halton extension for variables with more than two categories; missing or invalid grade codes excluded) for categorical variables and the Mann-Whitney U test for continuous variables; these two-sided p-values are descriptive and were not adjusted for multiple comparisons. All analyses were performed in Python 3.11 using scikit-learn. A completed TRIPOD+AI reporting checklist is given as a **Supplementary File**.

RESULTS

Patient Characteristics

Table 1 summarizes characteristics for the 47 patients with known ypN status (analysis cohort) and the 15 patients excluded for lack of a recorded outcome (reasons in methods: no surgery recorded, n = 8; gastrectomy without a recorded ypN, n = 4; biopsy only with progression, n = 2; aborted resection, n = 1). Two characteristics differed between the groups: clinical N stage (Fisher-Freeman-Halton exact p = 0.017; cN3 in 1 of 47 vs. 3 of 15 patients) and platelet count

Table 1. Characteristics of patients with and without a recorded pathological lymph-node stage (ypN)

Characteristic	Known ypN (n = 47)	Unknown ypN (n = 15)	p-value (known vs. unknown ypN)
Age, years, median (IQR)	61 (57.0-65.5)	62 (51.5-66.0)	0.987
BMI, kg/m^2, median (IQR)	25.4 (23.0-29.0)	25.8 (23.1-28.7)	0.928
WBC, ×10^9/L, median (IQR)	7.8 (6.7-8.5)	7.5 (7.0-7.9)	0.455
Platelet, ×10^9/L, median (IQR)	277.0 (203.5-331.5)	231.0 (174.8-250.8)	0.024
Sex, n (%)			
Male	31 (66.0%)	11 (73.3%)	0.755
Female	16 (34.0%)	4 (26.7%)	
Clinical T stage, n (%)			
cT2	4 (8.5%)	1 (6.7%)	1.000
cT3	35 (74.5%)	11 (73.3%)	
cT4	8 (17.0%)	3 (20.0%)	
Clinical N stage, n (%)			
cN0	11 (23.4%)	1 (6.7%)	0.017
cN1	16 (34.0%)	7 (46.7%)	
cN2	19 (40.4%)	3 (20.0%)	
cN3	1 (2.1%)	3 (20.0%)	
cNx	0 (0.0%)	1 (6.7%)	
Tumor grade, n (%)			
G1	3 (6.4%)	0 (0.0%)	0.559
G2	15 (31.9%)	7 (46.7%)	
G3	24 (51.1%)	6 (40.0%)	
G4	1 (2.1%)	0 (0.0%)	
Missing/invalid grade code	4 (8.5%)	2 (13.3%)	
Pathological ypN (n = 47 only), n (%)			
ypN0 (negative)	29 (61.7%)	-	-
ypN1	3 (6.4%)	-	
ypN2	11 (23.4%)	-	
ypN3	4 (8.5%)	-	

Note. WBC and platelet count for the two groups are reported (median (IQR): known ypN 7.8 (6.7-8.5) × 10^9/L and 277.0 (203.5-331.5) × 10^9/L; unknown ypN 7.5 (7.0-7.9) × 10^9/L and 231.0 (174.8-250.8) × 10^9/L, n = 14/15 with recorded values). Regarding the reasons the 15 patients lacked a recorded ypN value: 4 underwent a documented gastrectomy with a recorded operative date but no ypN recorded – whether this reflects incomplete pathology abstraction, unavailable pathology, or another reason could not be established from the available records; 2 underwent biopsy only, both with documented disease progression including peritoneal carcinomatosis that precluded resection; 1 had resection aborted for positive peritoneal cytology at second-look laparoscopy; and 8 have no surgery recorded at all as of data extraction, most likely still awaiting resection (1 of these 8 also shows peritoneal-carcinomatosis progression). Of the 3 cN3 patients in the excluded group (20.0% vs. 2.1% in the known-ypN group), 2 were among the progression/unresectable-disease cases described above; given how small this subgroup is (n = 3), this is reported as a descriptive observation rather than as support for a causal hypothesis about clinical nodal stage and non-resection. The third had surgery performed, so the missing ypN in that case appears to reflect a reporting gap rather than unrespectability. p-values compare patients with known and unknown ypN: Fisher's exact test (Fisher-Freeman-Halton exact test for variables with more than two categories; missing or invalid grade codes excluded) for categorical variables and the Mann-Whitney U test for continuous variables; two-sided, unadjusted for multiple comparisons, and descriptive only.

(median 277.0 vs. 231.0 × 10^9/L; Mann-Whitney p = 0.024). Sex (p = 0.755), clinical T stage (p = 1.00), tumor grade (p = 0.56), age (p = 0.99), BMI (p = 0.93), and WBC (p = 0.46) did not differ. With 15 patients in the unknown-ypN group and eight unadjusted comparisons (neither nominal difference would remain below 0.05 after Bonferroni correction), these comparisons are descriptive and have low power; the imbalances in clinical N stage and platelet count are nevertheless consistent with a possible selection effect, as discussed in limitations.

Feature Selection

No univariate outcome screening or LASSO-based feature selection was used. Two compact, clinically defined panels were compared: cT, cN, grade, age, and BMI (five predictors), and the same panel with WBC and platelet count (seven predictors). The seven-predictor panel was carried forward for the exploratory analysis because it had the higher out-of-fold point AUC. Because this choice was made on the same cross-validated data that yield the reported estimates, the AUC of 0.701 is conditional on data-driven predictor selection and should be read as an exploratory, potentially optimistic estimate rather than an independently validated one. The paired uncertainty in the difference is reported explicitly.

Model Training and Evaluation

Table 2 reports out-of-fold discrimination and calibration metrics for all three models. The ridge logistic regression achieved an AUC of 0.701 (95% bootstrap CI 0.539-0.851); the random forest achieved 0.667 (95% CI 0.488-0.831); and the Firth sensitivity model achieved 0.703 (95% CI 0.532-0.851). Corresponding PR-AUC values were 0.558, 0.590, and 0.576 (outcome-prevalence baseline 0.383), and Brier scores were 0.227, 0.227, and 0.222, for the ridge, random forest, and Firth models, respectively (**Table 2**). The estimates remain imprecise because only 47 outcome-evaluable patients were available. Using a reference Brier score of 0.236 (the Brier score obtained by predicting the outcome prevalence, 38.3%, for every patient), the scaled Brier score (1-Brier$_{model}$/Brier$_{reference}$) was 0.038 for the seven-predictor ridge model, 0.038 for the random forest, and 0.059 for the Firth model – indicating that all three models achieved only modest probabilistic improvement over this trivial reference at the present sample size, with the largest gain for the Firth model.

Table 2. Out-of-fold discrimination and calibration for all three models (n = 47)

Model	AUC (95% CI)	PR-AUC	Brier	Sensitivity	Specificity
Ridge logistic regression	0.701 (0.539-0.851)	0.558	0.227	0.278	0.931
Shallow random forest	0.667 (0.488-0.831)	0.590	0.227	0.611	0.759
Firth logistic (LOOCV)	0.703 (0.532-0.851)	0.576	0.222	0.556	0.793

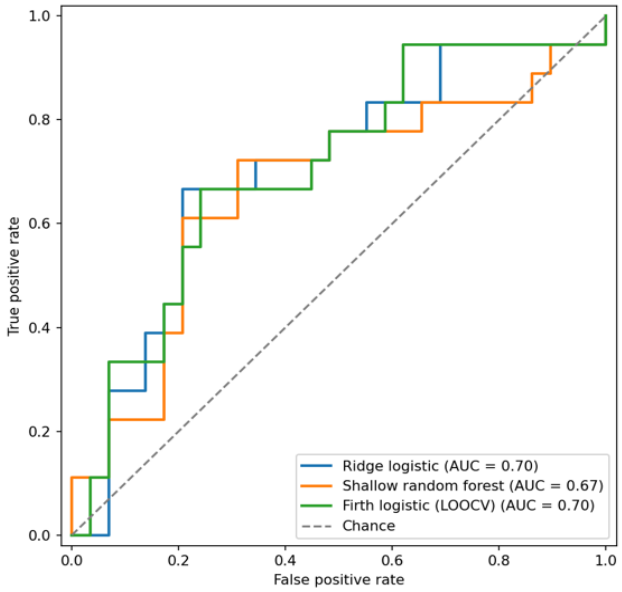


Figure 2. Out-of-fold ROC curves for the three models (the dashed diagonal indicates chance-level discrimination [AUC = 0.5]) (the authors' own elaboration)

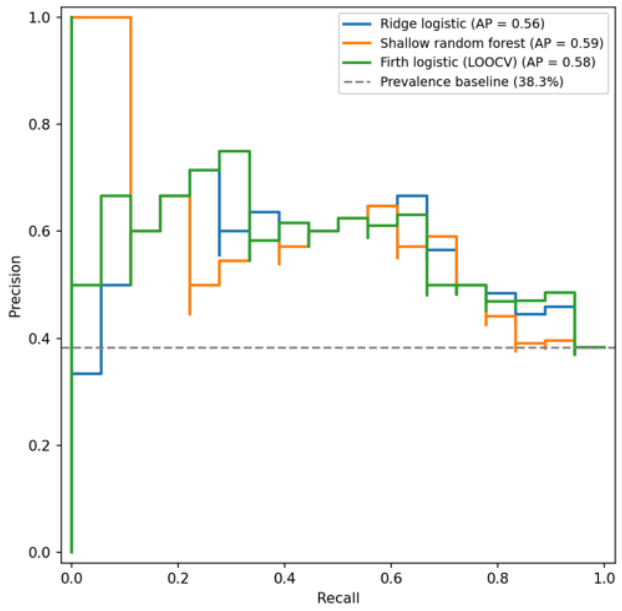


Figure 3. Out-of-fold precision-recall curves (the dashed horizontal line indicates the baseline achievable by predicting the outcome prevalence for every patient [38.3%]) (the authors' own elaboration)

Two calibration procedures are reported: calibration-in-the-large (an intercept-only model with the predicted logit as a fixed offset, slope fixed at 1) was ridge -0.02 (standard error [SE] 0.31), random forest -0.34 (SE 0.31), and Firth -0.08 (SE 0.33) – all close to zero, indicating no strong systematic over- or under-prediction on average. The logistic recalibration intercept and slope (intercept and slope estimated simultaneously) were

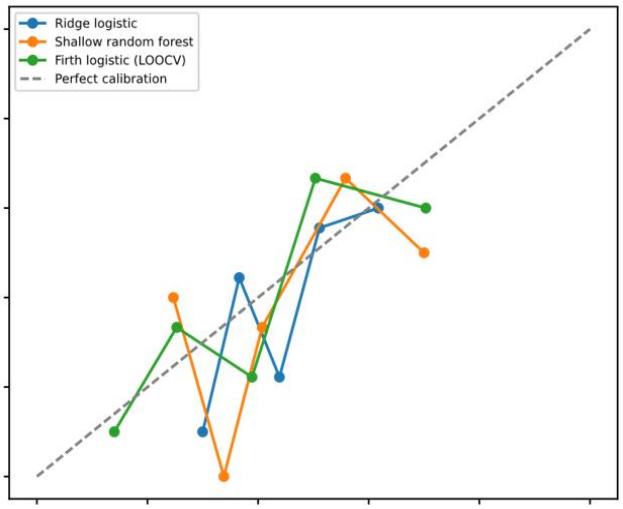


Figure 4. Calibration curves, five quantile-defined bins per model (the dashed diagonal indicates perfect calibration) (the authors' own elaboration)

ridge intercept -0.20 (SE 0.40), slope 0.59 (SE 0.57); random forest intercept -0.31 (SE 0.32), slope 1.25 (SE 0.75); Firth intercept -0.34 (SE 0.33), slope 0.29 (SE 0.27).

Calibration slopes below 1 indicate that predicted probabilities were too extreme relative to the observed outcomes (i.e., over-dispersion/overfitting of the predicted spread), most evident for the Firth predictions (slope 0.29); all calibration estimates are highly imprecise, since only 47 patients were available. These are reported alongside the five-bin calibration plots to aid interpretation at this sample size, since bins of approximately nine patients each are difficult to interpret alone.

Figure 2 shows the out-of-fold ROC curves for all three models against the chance diagonal; **Figure 3** shows the corresponding precision-recall curves against the outcome-prevalence baseline (38.3%). **Figure 4** shows calibration curves in five quantile bins for each model. **Figure 5** shows the bootstrap AUC distributions at n = 47.

Figure 6 shows the seven standardized coefficients for both the ridge and the Firth model, each with bootstrap 95% intervals (n = 1,000 resamples, identical resampled patients used for both models). For the ridge model, six intervals crossed zero; only the age coefficient interval excluded zero. For the Firth model, all seven intervals crossed zero, including age, whose interval only narrowly included zero. These coefficient analyses are descriptive and should not be interpreted as independently validated effects.

The final ridge model (C = 0.1), fitted to the full cohort for descriptive interpretation, had standardized coefficients (odds ratios per SD) of cN 0.237 (1.27), cT 0.049 (1.05), tumor grade 0.194 (1.21), age -0.311 (0.73), BMI 0.060 (1.06), WBC 0.275 (1.32), and platelet count 0.159 (1.17). The intercept, the standardization means and SDs, the predictor units, and the prediction formula needed to apply this model unchanged to new patients are given in **Table A1** in **Appendix A**.

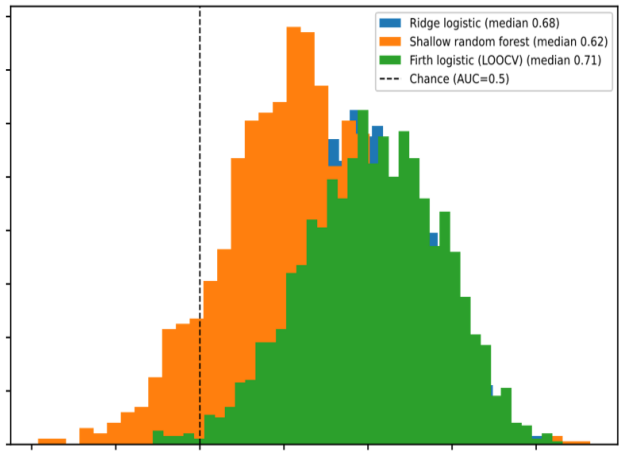


Figure 5. Bootstrap AUC distributions (2,000 patient-level resamples per model), shown as histograms rather than interval endpoints alone to make the width and shape of the sampling uncertainty visible (the authors' own elaboration)

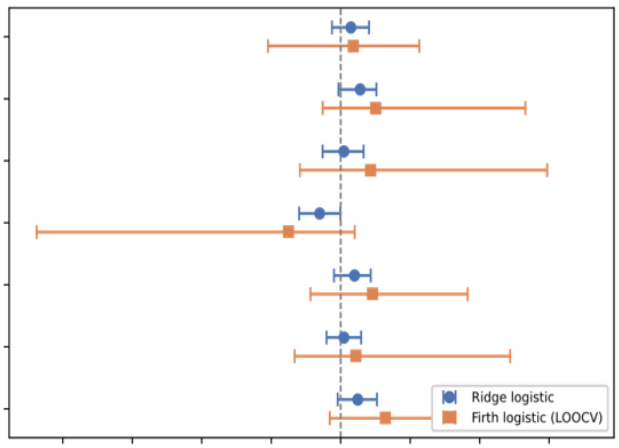


Figure 6. Standardized ridge-logistic and Firth-logistic coefficients for the seven-predictor model (bootstrap median and 95% CI, n = 1,000 paired resamples per model) (the authors' own elaboration)

The corresponding full-cohort Firth (bias-reduced) fit, included as a sensitivity check on the same seven standardized predictors, gave coefficients of cN 0.506 (odds ratio [OR] 1.66), cT 0.198 (OR 1.22), tumor grade 0.364 (OR 1.44), age -0.677 (OR 0.51), BMI 0.365 (OR 1.44), WBC 0.426 (OR 1.53), and platelet count 0.208 (OR 1.23). The direction of every coefficient matched the ridge fit; the larger magnitudes reflect Firth's lower shrinkage relative to the cross-validated ridge penalty, not a different qualitative conclusion.

Predictor-Panel Sensitivity Analysis

Using the same patients, folds, preprocessing, and ridge-tuning procedure, the five-predictor panel (cT, cN, grade, age, and BMI) achieved AUC 0.623, PR-AUC 0.517, and Brier score 0.236. Adding

WBC and platelet count produced AUC 0.701, PR-AUC 0.558, and Brier score 0.227. The paired AUC difference was +0.079 (95% bootstrap CI -0.025 to +0.201). The seven-predictor panel therefore had the higher point AUC and was retained for exploratory analysis, but the interval includes zero and does not establish a reproducible improvement. The five-predictor panel's scaled Brier score was 0.000 relative to the prevalence-only reference (0.236), indicating minimal probabilistic improvement over always predicting the outcome prevalence.

Clinical-Staging Baseline Comparison

To contextualize the seven-predictor model against routine clinical staging, clinical N stage alone and clinical T plus N stage combined were evaluated as baseline models under the identical repeated stratified cross-validation scheme (ridge-penalized logistic regression, 5 folds × 20 repeats) used for the primary models. The cN-alone panel achieved AUC 0.517, PR-AUC 0.391, and Brier score 0.242; the cT+cN panel achieved AUC 0.510, PR-AUC 0.398, and Brier score 0.243. Both baseline panels therefore performed close to chance. The paired bootstrap AUC difference between the seven-predictor ridge model and the cT+cN baseline was +0.191 (95% CI -0.002 to +0.372), so the seven-predictor panel had a higher point AUC than routine clinical staging alone, but because the interval includes zero these data do not establish that it discriminates better than the cT+cN baseline.

DISCUSSION

This pilot study evaluated whether postoperative lymph-node positivity could be predicted from seven preoperative variables, that is, variables recorded before radical surgery: clinical T stage, clinical N stage, tumor grade, age, BMI, WBC, and platelet count. All three modeling approaches gave out-of-fold AUC point estimates between 0.667 and 0.703. Because the seven-predictor panel was chosen on the same cross-validated data, these estimates are conditional on data-driven predictor selection and are best read as an exploratory signal of modest, imprecise discrimination, not as evidence of established predictive ability or of independent validation: the 95% intervals were wide (lower limits 0.488-0.539), the random forest interval included 0.5, and patients who were ultimately ypN-positive overlapped substantially with ypN0 patients in predicted probability.

Ridge logistic regression, chosen as the primary approach because coefficient shrinkage can stabilize estimation when the number of predictors is large relative to the number of outcome events, achieved an AUC of 0.701 (95% CI 0.539-0.851), PR-AUC 0.558, and Brier score 0.227. The random forest performed similarly (AUC 0.667, 95% CI 0.488-0.831), giving no evidence at this sample size that nonlinear relationships or predictor interactions offer an advantage over a regularized additive model. Firth logistic regression, evaluated by leave-one-out cross-validation as a bias-reduction sensitivity check appropriate for small, event-limited cohorts, achieved the highest point AUC (0.703, 95% CI 0.532-0.851). All three intervals were wide and overlapping, so these results do not establish that any one approach is superior to the others.

Adding WBC and platelet count to the five-variable panel (cT, cN, grade, age, BMI) raised the point AUC from 0.623 to 0.701, PR-AUC from 0.517 to 0.558, and lowered the Brier score from 0.236 to 0.227, but the paired bootstrap difference (+0.079, 95% CI -0.025 to +0.201) crosses zero, so this apparent gain could reflect sampling variation and needs prospective confirmation. WBC and platelet count are nonetheless routinely available, low-cost measurements with a biologically plausible link to lymph-node status through systemic inflammation and tumor-associated immune response; because this study was designed for prediction rather than

causal inference, their coefficients should not be read as evidence that changing these levels would alter nodal status. Clinical T/N stage and tumor grade contribute complementary anatomical and histopathological information, and age and BMI add basic patient context; the panel was intentionally limited to seven variables to preserve interpretability and limit overfitting with only 18 outcome events.

An AUC of 0.701 means that, given one randomly selected ypN-positive and one ypN0 patient, the ridge model ranks the ypN-positive patient higher about 70.1% of the time – not that 70.1% of individual predictions are correct, and it defines no clinical decision threshold. PR-AUC and the Brier score complement this by reflecting performance on the positive class and calibration, respectively; together with the calibration plots and bootstrap intervals reported above, they give a fuller picture than any single metric.

This performance is lower than most published gastric-cancer lymph-node-metastasis models, which typically use eight or more predictors in cohorts of several hundred to several thousand patients and report external-validation AUCs of roughly 0.72-0.86 [11]. The gap is unsurprising given the difference in event counts, predictor diversity, and validation design, and direct comparison is further complicated because most of those studies predict nodal status at diagnosis or in early gastric cancer, whereas the present outcome is pathological status after neoadjuvant treatment and PIPAC-related management. The aim here was not to match those larger models but to explore whether a transparent, leakage-safe workflow could detect any exploratory signal in a small, clinically distinct regional cohort.

The prediction time point deliberately excludes ypT, nodal counts, lymph-node ratio, regression grade, resection status, operative details, and any other postoperative or outcome-derived variable. Cross-validation alone cannot correct this kind of temporal leakage: predictor eligibility was fixed by the clinical timeline before model fitting, and imputation, standardization, and ridge-penalty selection were then performed on training-fold data only.

To our knowledge, this is the first leakage-safe prediction-model analysis of postoperative lymph-node status in a PIPAC-treated gastric cancer population, and specifically the first in a Kazakhstani cohort; published lymph-node models have generally been developed in general surgical or early gastric cancer cohorts rather than PIPAC-selected patients [11]. Every reported estimate came from out-of-fold predictions for real patients rather than synthetic or simulated data, so the effective evaluation sample was 47 patients throughout.

These results describe feasibility and preoperative risk ranking only. They do not establish that PIPAC is beneficial relative to alternative treatment, and they are not yet sufficient to guide a surgical decision for an individual patient. The logical next step is prospective recruitment under a standardized protocol, testing this seven-variable specification unchanged for a fair temporal or external validation, before considering additional predictors (inflammatory ratios, tumor markers, imaging or radiomic features) or expanding through multicenter collaboration within Kazakhstan and the broader region to increase event counts and assess model transportability.

Limitations

This study should be interpreted as an exploratory, single-center pilot analysis. The cohort included 47 outcome-evaluable patients, of whom 18 were ypN-positive. With about 2.6 outcome events per candidate predictor, even penalized regression is at the limit of what these data can support; the depth-restricted random forest and the Firth model were included as limited comparators, not as evidence that a more complex model is warranted, and none of the models should be regarded as a fully developed or externally validated prediction model. Consequently, the performance estimates remain uncertain and may change when the models are evaluated in larger or more heterogeneous populations. Because this was a small single-center cohort, we could not meaningfully evaluate potential health inequalities or differences in model performance across sociodemographic subgroups. External validation was not possible because an independent PIPAC-treated cohort with comparable preoperative variables and postoperative ypN outcomes was unavailable. Several methodological safeguards were implemented to obtain the most reliable estimates possible from this limited cohort. Ridge logistic regression and random forest were evaluated using repeated patient-level cross-validation. Missing-value imputation, predictor scaling, and ridge-penalty selection were performed using training-fold information only. Firth logistic regression was evaluated separately using leave-one-out cross-validation, so each patient's prediction was generated by a model fitted without that patient. Patient-level bootstrap resampling was used to quantify uncertainty around model discrimination. Together, these procedures reduced information leakage and optimistic evaluation, although they cannot replace independent external validation. The seven-variable panel was retained after comparison with the five-variable panel because it had the higher cross-validated point AUC. Adding WBC and platelet count increased the ridge ROC AUC from 0.623 to 0.701. However, the paired bootstrap 95% confidence interval for this difference included zero (+0.079, 95% CI -0.025 to +0.201). Therefore, the present data suggest, but do not establish, that WBC and platelet count provide additional predictive information. This predictor-panel comparison may also introduce selection-related optimism, and the seven-variable model should consequently be evaluated without further modification in a new cohort. Sensitivity and specificity were calculated at a fixed predicted-probability threshold of 0.50 for descriptive purposes. This threshold was not optimized for clinical utility and should not be interpreted as a recommended treatment or surveillance threshold. The timing of WBC and platelet count relative to the start of neoadjuvant treatment is not recorded for most patients, so "preoperative" here means recorded before radical surgery; counts drawn before and after neoadjuvant therapy may carry different information, and the model should be applied only with a prospectively fixed measurement time. Four patients had missing or invalid tumor-grade values. These values were imputed exclusively within the corresponding training folds, maintaining separation between training and held-out patients, although prospective collection of complete tumor-grade information would be preferable. Missing-data procedures must be established during prospective validation before implementation. The current models predict postoperative lymph-node positivity; they do not estimate the causal effect of PIPAC, determine which treatment is optimal,

or demonstrate that changing treatment according to the predicted probability improves patient outcomes. Furthermore, the cohort represents a selected group of patients with known postoperative ypN status. The findings may therefore not generalize to patients who do not undergo radical surgery or whose postoperative pathology is unavailable. The 15 excluded patients had a more advanced clinical N-stage distribution (cN3 in 3 of 15 vs. 1 of 47; exact p = 0.017) and lower platelet counts (median 231.0 vs. 277.0 × 10^9/L; p = 0.024), so the analysis cohort may differ systematically from patients who did not reach pathological assessment; these comparisons are unadjusted and descriptive. In addition, although all 62 patients were treated within a common program, completion of neoadjuvant chemotherapy (54 of 62 completed all four planned cycles), the interval between the recorded laparoscopy and surgery (34-152 days), and the number and timing of PIPAC cycles varied or could not be verified, so the estimates may partly reflect treatment heterogeneity. Clinical T stage, clinical N stage, and tumor grade were modeled as linear (ordinal-coded) predictors; these variables may not have a linear relationship with log-odds of ypN positivity in reality. A categorical (dummy-coded) sensitivity analysis was evaluated (S2 in **Appendix A**; AUC 0.684 vs. 0.701 for the linear-ordinal specification, both under an identical, unified regularization grid C ∈ {0.001, 0.01,0.1, 1.0, 10.0}), which did not show clear evidence of non-linearity at this sample size but remains an imprecise comparison; a restricted cubic spline or polynomial specification was not additionally evaluated. Relatedly, the present analysis used a dichotomous outcome (ypN-positive vs. ypN0); dichotomizing an underlying ordinal or continuous measure of nodal burden is known to reduce statistical power and can require a substantially larger sample size than an equivalent analysis of a continuous or ordinal surrogate outcome. Additional sensitivity analyses of numerical nodal-burden outcomes are described as S6 in **Appendix A**: a numeric-stage ridge linear regression did not improve discrimination of the binary outcome (AUC 0.62 vs. 0.70; R^2 -0.04), contrary to the general expectation [37, 38], and the exploratory count and ratio models of nodal burden were full-cohort and non-cross-validated and should not be over-interpreted at n = 47. Rerunning the identical pipeline across six further, arbitrarily chosen seeds (1-6) gave ridge AUCs of 0.670-0.701 and random forest AUCs of 0.642-0.688 across all seven seeds (S2 in **Appendix A**), so the result under the prespecified seed is not an outlier.

These limitations define the next phase of the research rather than diminishing the feasibility demonstrated by this study. The principal priorities are prospective recruitment, standardized collection of preoperative predictors, expansion to multiple clinical centers, evaluation of prespecified probability thresholds, and temporal and external validation in an independent patient population.

CONCLUSIONS

In this exploratory pilot study of patients with gastric cancer treated within a single-center PIPAC program, seven routinely available variables carried at most a modest and imprecise signal for postoperative lymph-node positivity (ypN), and the estimates are too uncertain to support clinical use. The principal contribution is a transparent, leakage-safe analytical workflow that can be applied unchanged in larger cohorts. The models should be regarded as research prototypes and should not be used to select treatment, determine eligibility for surgery, or replace multidisciplinary clinical assessment. Prospective, multicenter validation of the unchanged seven-variable specification, including calibration, clinical utility, and comparison with conventional clinical staging, is required before any clinical application.

Author contributions: KS: conceptualization, data curation, formal analysis, methodology, software, validation, visualization, writing – original draft; **AR:** writing – original draft; **AK:** conceptualization, investigation, methodology, resources, supervision, writing – review & editing; **TS:** investigation, methodology, validation, writing – review & editing; **AK:** investigation, writing – review & editing; **AG:** conceptualization, methodology, resources, supervision, writing – review & editing. All authors have agreed with the results and conclusions.

Funding: The data analyzed were collected under a grant from the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant number BR24992950 – Creation and implementation of innovative treatment methods for oncological diseases).

Acknowledgments: The authors would like to thank the clinical team responsible for source PIPAC treatment program for data collection.

Ethical statement: This study was approved by the National Research Oncology Center Board on 24 April 2024 with approval number 051. Written informed consent was obtained from all the participants. Participants have the right not to participate in the study and be treated by standard therapy protocol. Participant data will be de-identified before analysis. Each participant will be assigned a unique study ID; the linkage file connecting IDs to personal identifiers will be stored separately on an encrypted, access-restricted server JMIR Res Protoc 2025 | vol. 14 | e78053 | p. 5 (page number not for citation purposes) RenderXJMIR RESEARCH PROTOCOLS and will be accessible only to authorized study personnel. No personal identifiers will be visible during data processing or analysis. For this study, participants receive no monetary compensation.

AI statement: An AI tool (a large language model developed by Anthropic) was used to assist in debugging and optimizing the Python code for the cross-validation, model-fitting, and sensitivity analyses. The final scripts were verified, tested, and fully executed by the research team.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

1. Akhmedullin R, Aimyshev T, Zhakhina G, et al. In-depth analysis and trends of cancer mortality in Kazakhstan: A joinpoint analysis of nationwide healthcare data, 2014-2022. BMC Cancer. 2024;24(1):1340. https://doi.org/10.1186/s12885-024-13128-2 PMid:39482619 PMCid: PMC11526534
2. Rakhmankulova A, Sakko Y, Kerimkulov A, et al. Survival disparities and predictors in gastric cancer: A population-based study from Kazakhstan (2012-2023). Front Oncol. 2025;15:1670082. https://doi.org/10.3389/fonc.2025.1670082 PMid:41357613
3. Ramalho-Vasconcelos F, Gomes R, Bouça-Machado R, et al. Pressurized intraperitoneal aerosol chemotherapy (PIPAC) in the treatment of gastric cancer: Feasibility, efficacy, and safety—A systematic review and meta-analysis. J Clin Med. 2024;13(11):3320. https://doi.org/10.3390/jcm13113320 PMid:38893031 PMCid:PMC11173180

4. Kerimkulov A, Uskenbayev T, Sarina T, et al. Efficacy of preventive pressurized intraperitoneal aerosol chemotherapy in patients with locally advanced gastric cancer: Protocol for a prospective controlled trial. JMIR Res Protoc. 2025;14:e78053. https://doi.org/10.2196/78053 PMid:41218205 PMCid:PMC12648122
5. Li Z, Wang Y, Shan F, et al. ypTNM staging after neoadjuvant chemotherapy in the Chinese gastric cancer population: An evaluation on the prognostic value of the AJCC eighth edition cancer staging system. Gastric Cancer. 2018;21(6):977-87. https://doi.org/10.1007/s10120-018-0830-1 PMid:29748876
6. Su P, Zhang Y, Yu T, et al. Comparison of the predictive value of pathological response at the primary tumour and lymph node status after neoadjuvant chemotherapy in locally advanced gastric cancer. Clin Transl Oncol. 2023;25(8):2462-71. https://doi.org/10.1007/s12094-023-03130-8 PMid:37093455
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58. https://doi.org/10.1056/NEJMra1814259 PMid:30943338 PMCid:PMC12363574
8. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-9. https://doi.org/10.1038/s41591-018-0316-z PMid:30617335
9. Bollschweiler EH, Mönig SP, Hensler K, et al. Artificial neural network for prediction of lymph-node metastases in gastric cancer: A phase II diagnostic study. Ann Surg Oncol. 2004;11(5):506-11. https://doi.org/10.1245/ASO.2004.04.018 PMid:15123460
10. Xue Z, Lu J, Lin J, et al. [Establishment of an artificial neural-network model for predicting lymph node metastasis in patients with stage II-III gastric cancer]. Zhonghua Wei Chang Wai Ke Za Zhi. 2022;25(4):327-35. https://doi.org/10.3760/cma.j.cn441530-20220105-00010 PMid:35461201
11. Li Q, Yan S, Yang W, et al. Machine-learning models for prediction of lymph node metastasis in patients with gastric cancer: A Chinese single-centre study with external validation in an Asian American population. BMJ Open. 2025;15(3):e098476. https://doi.org/10.1136/bmjopen-2024-098476 PMid:40132850 PMCid:PMC11938237
12. Ding F, Huang B, Liu J, et al. Interpretable machine learning analysis of clinicopathological and immunonutritional biomarkers for predicting lymph node metastasis in gastric cancer. Sci Rep. 2025;15:44964. https://doi.org/10.1038/s41598-025-29004-3 PMid:41272102 PMCid:PMC12749999
13. Zhu H, Wang G, Zheng J, et al. Preoperative prediction of lymph-node metastasis in early gastric cancer using interpretable machine learning models: A multicentre study. Surgery. 2022;171(6):1543-51. https://doi.org/10.1016/j.surg.2021.12.015 PMid:35131110
14. Song K, Wu J, Xu M, et al. An ensemble learning model to predict lymph node metastasis in early gastric cancer. Sci Rep. 2025;15:11257. https://doi.org/10.1038/s41598-025-95794-1 PMid:40175493 PMCid:PMC11965319
15. Seo JW, Park KB, Lim ST, Jun KH, Chin HM. Machine learning models for prediction of lymph node metastasis in patients with T1b gastric cancer. Am J Cancer Res. 2024;14(8):3842-51. https://doi.org/10.62347/KREL8138 PMid:39267667 PMCid:PMC11387857
16. Ma T, Zhao M, Li X, Song X, Wang L, Ye Z. A machine learning based radiomics approach for predicting no. 14v station lymph node metastasis in gastric cancer. Front Med (Lausanne). 2024;11:1464632. https://doi.org/10.3389/fmed.2024.1464632 PMid:39493708 PMCid:PMC11527654
17. Yue C, Xue H. Construction and validation of a nomogram for lymph node metastasis in stage II-III gastric cancer based on machine learning algorithms. Front Oncol. 2024;14:1399970. https://doi.org/10.3389/fonc.2024.1399970 PMid:39439953 PMCid:PMC11493538
18. Zheng Y, Qiu B, Liu S, et al. A transformer-based deep learning model for early prediction of lymph node metastasis in locally advanced gastric cancer after neoadjuvant chemotherapy using pretreatment CT images. EClinicalMedicine. 2024;75:102805. https://doi.org/10.1016/j.eclinm.2024.102805 PMid:39281097 PMCid:PMC11402411
19. Lin W, Zhong Q, Guo J, et al. A preoperative prediction model for lymph node metastasis in patients with gastric cancer using a machine learning-based ultrasomics approach. Curr Med Imaging. 2024;20:e15734056291074. https://doi.org/10.2174/0115734056291074240522052725 PMid:38803184
20. He J-Y, Cao M-X, Li E-Z, et al. Development and validation of a nomogram for predicting lymph node metastasis in early gastric cancer. World J Gastrointest Oncol. 2024;16(7):2960-70. https://doi.org/10.4251/wjgo.v16.i7.2960 PMid:39072177 PMCid:PMC11271770
21. Zhang M, Ding C, Xu L, et al. A nomogram to predict the risk of lymph node metastasis in early gastric cancer. Sci Rep. 2021;11(1):22873. https://doi.org/10.1038/s41598-021-02305-z PMid:34819570 PMCid:PMC8613278
22. Liu L, Liu W, Jia Z, et al. Application of machine-learning algorithms to predict lymph node metastasis in gastric neuroendocrine neoplasms. Heliyon. 2023;9(10):e20928. https://doi.org/10.1016/j.heliyon.2023.e20928 PMid:37928390 PMCid:PMC10622622
23. Li Y, Xie F, Xiong Q, Lei H, Feng P. Machine learning for lymph node metastasis prediction in patients with gastric cancer: A systematic review and meta-analysis. Front Oncol. 2022;12:946038. https://doi.org/10.3389/fonc.2022.946038 PMid:36059703 PMCid:PMC9433672
24. Zhao X, Cai A, Xi H, et al. Predictive factors for lymph-node metastasis in undifferentiated early gastric cancer: A systematic review and meta-analysis. J Gastrointest Surg. 2017;21(4):700-11. https://doi.org/10.1007/s11605-017-3364-7 PMid:28120275
25. Gotoda T, Yanagisawa A, Sasako M, et al. Incidence of lymph node metastasis from early gastric cancer: Estimation using a large number of cases at two large centres. Gastric Cancer. 2000;3(4):219-25. https://doi.org/10.1007/PL00011720 PMid:11984739
26. Abe N, Watanabe T, Suzuki K, et al. Risk factors predictive of lymph node metastasis in depressed early gastric cancer. Am J Surg. 2002;183(2):168-72. https://doi.org/10.1016/S0002-9610(01)00860-1 PMid:11918883
27. Lee JH, Choi M-G, Min B-H, et al. Predictive factors for lymph node metastasis in patients with poorly differentiated early gastric cancer. Br J Surg. 2012;99(12):1688-92. https://doi.org/10.1002/bjs.8934 PMid:23023388

28. Sekiguchi M, Oda I, Taniguchi H, et al. Risk stratification and a predictive risk-scoring model for lymph node metastasis in early gastric cancer. J Gastroenterol. 2016;51(10):961-70. https://doi.org/10.1007/s00535-016-1180-6 PMid:26884381
29. Pyo JH, Lee H, Min B-H, et al. A risk prediction model for lymph-node metastasis in poorly differentiated-type intramucosal gastric cancer. PLoS One. 2016;11(5):e0156207. https://doi.org/10.1371/journal.pone.0156207 PMid:27228258 PMCid:PMC4881979
30. Zhou C-M, Wang Y, Ye H-T, et al. Machine learning predicts lymph node metastasis of poorly differentiated-type intramucosal gastric cancer. Sci Rep. 2021;11:1300. https://doi.org/10.1038/s41598-020-80582-w PMid:33446730 PMCid:PMC7809018
31. Ji K, Shi L, Feng Y, et al. Construction and interpretation of machine learning-based prognostic models for survival prediction among intestinal-type and diffuse-type gastric cancer patients. World J Surg Oncol. 2024;22(1):275. https://doi.org/10.1186/s12957-024-03550-y PMid:39407221 PMCid:PMC11481450
32. Xu C, Wang J, Zheng T, Cao Y, Ye F. Prediction of prognosis and survival of patients with gastric cancer using a weighted improved random forest model: An application of machine learning in medicine. Arch Med Sci. 2021;18(5):1208-20. https://doi.org/10.5114/aoms/135594
33. Wang H-N, An J-H, Wang F-Q, Hu W-Q, Zong L. Predicting gastric cancer survival using machine learning: A systematic review. World J Gastrointest Oncol. 2025;17(5):103804. https://doi.org/10.4251/wjgo.v17.i5.103804
34. Chen Y, Wei K, Liu D, et al. A machine learning model for predicting a major response to neoadjuvant chemotherapy in advanced gastric cancer. Front Oncol. 2021;11:675458. https://doi.org/10.3389/fonc.2021.675458 PMid:34141620 PMCid:PMC8204104
35. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine-learning methods. BMJ. 2024;385:e078378.
36. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-31. https://doi.org/10.1093/eurheartj/ehu207 PMid:24898551
37. Chilvers C, Dewey M, Fielding K, et al. Antidepressant drugs and generic counselling for treatment of major depression in primary care: Randomised trial with patient preference arms. BMJ. 2001;322(7289):772-5. https://doi.org/10.1136/bmj.322.7289.772 PMid:11282864 PMCid:PMC30555
38. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. Stat Med. 2006;25(1):127-41. https://doi.org/10.1002/sim.2331 PMid:16217841

APPENDIX A: SUPPLEMENTARY METHODS

Pilot prediction model of postoperative lymph-node status (ypN) in patients treated within a single-center pressurized intraperitoneal aerosol chemotherapy (PIPAC) program. This file accompanies the main manuscript and contains technical details removed from the methods section.

S1. Additional Cohort Details

Neoadjuvant treatment

Of the 62 patients, 54 (87.1%) completed all four planned cycles of neoadjuvant systemic chemotherapy and 8 (12.9%) had an incomplete course. The specific neoadjuvant regimen is not recorded in a dedicated field. The recorded adjuvant-regimen field (FLOT in 33 of the 47 analyzed patients; 2 other; 12 not recorded) describes treatment given after surgery and cannot be used to infer the preceding neoadjuvant regimen.

PIPAC procedure

The dataset records one PIPAC procedure-detail entry (duration, cisplatin dose, doxorubicin dose) per patient for all 62 patients (median recorded duration 75 minutes, IQR 66-81). This confirms one recorded entry per patient but does not establish that every patient received exactly one PIPAC session. Therapeutic-versus-prophylactic indication, neoadjuvant intent, the number of PIPAC cycles, and the interval between cycles are not recorded.

Interval from laparoscopy to surgery

Using the first recorded laparoscopy date as a proxy for the PIPAC date, the median interval to radical surgery among the 47 analyzed patients was 90 days (IQR 83.0-98.5, range 34-152). Because the dataset does not establish that this date corresponds to the PIPAC session, this is a proxy and not a confirmed PIPAC-to-surgery interval. Adding this interval as an eighth candidate predictor did not change discrimination (AUC 0.701 with and without; paired bootstrap difference +0.000, 95% CI -0.090 to +0.096), and it was not retained.

S2. Coding of Ordinal Predictors and Categorical Sensitivity Analysis

Clinical T stage (cT), clinical N stage (cN), and tumor grade were entered as ordinal numeric scores. In the 47-patient analysis cohort, cT was coded 1, 2, and 3 for cT2, cT3, and cT4 (no cT1 patients were present), cN was coded 0-3 for cN0-cN3, and grade was coded 1-4 for G1-G4; an unrecorded or invalid grade was treated as missing and imputed with the training-fold median (see main text, Data preprocessing). In the primary specification each score entered into the model as a single numeric column, which assumes equal spacing between adjacent categories on the log-odds scale; each column was then standardized (training-fold mean and population SD) so that the ridge penalty acted on a common scale. Because of this standardization the choice of origin (for example, starting cT at 1 rather than 2) is immaterial, whereas the spacing of the scores is not. The single patient with an unrecorded clinical N stage (cNx) belonged to the excluded unknown-ypN group and therefore required no coding.

In the categorical sensitivity analysis, cT, cN, and grade were instead entered as unordered indicator (dummy) variables for each observed category, without any assumption of linear spacing, under the same outer and inner cross-validation scheme and the same regularization grid as the primary model (S2), so that both specifications were compared under identical tuning conditions. The categorical specification achieved AUC 0.684, PR-AUC 0.635, and Brier score 0.216, compared with AUC 0.701, PR-AUC 0.558, and Brier score 0.227 for the linear-ordinal specification (paired bootstrap AUC difference -0.016, 95% CI -0.102 to +0.079). The two specifications therefore performed similarly at this sample size, giving no clear evidence that the linear-ordinal coding materially misrepresents the relationship between cT, cN, and grade with ypN positivity, although the comparison is imprecise and a restricted cubic spline or polynomial specification was not evaluated.

S3. Model Tuning, Random Forest Settings, and Seed Robustness

Ridge logistic regression (primary model)

An $\ell 2$-regularized logistic regression (scikit-learn LogisticRegression, lbfgs solver, max_iter = 1,000) was used. The regularization strength C was selected from the grid {0.001, 0.01, 0.1, 1, 10} by an inner cross-validation grid search (3 folds × 2 repeats) within each outer training fold, optimizing the negative Brier score. The inner search saw only the outer training fold.

Random forest (comparator)

An ensemble of 500 trees, each restricted to a maximum depth of 2 and a minimum of 5 samples per leaf, with class-balanced sample weighting, was used to limit overfitting given the small sample size. No hyperparameter search was performed for the random forest.

Outer validation

Ridge regression and the random forest were evaluated by repeated stratified cross-validation (5 folds × 20 repeats, 100 folds). Imputation, standardization, and (for ridge regression) penalty selection were refitted within each training fold. A single fixed random seed (20260924), chosen before analysis, controlled the fold assignments and the random-forest estimator. **Table 2** and **Figure 2** and **Figure 3** of the main manuscript were generated from the same out-of-fold predictions.

Seed robustness

Because the AUCs in the main manuscript were obtained under one seed, the identical pipeline was rerun with six further, arbitrarily chosen seeds (1-6, selected before looking at any result). Across all seven seeds the ridge AUC ranged from 0.670 to 0.701 (mean 0.689, SD 0.012) and the random-forest AUC from 0.642 to 0.688 (mean 0.670, SD 0.015). The result under the prespecified seed therefore lies within this range and is not an outlier selected after the fact.

S4. Firth Logistic Regression: Software and Verification

Firth (bias-reduced) logistic regression applies Firth's modified-score penalty (a Jeffreys-invariant-prior penalty) and was included as a sensitivity analysis because ordinary maximum-likelihood logistic regression can be unstable with a small number of events per predictor. It was evaluated by leave-one-out cross-validation, in which each patient's prediction came from a model fitted without that patient, with median imputation and standardization refitted on each training set.

Because scikit-learn does not include Firth regression, all Firth results in the manuscript (**Table 2** and the full-cohort coefficients) were generated with the open-source Python package firthmodels (version 0.8.2; FirthLogisticRegression class, default settings; https://github.com/jzluo/firthmodels) under Python 3.11. This package was used because it provides a scikit-learn-compatible estimator, so that the Firth model sits inside the same fold-wise pipeline as the other models; because it is the actively maintained successor of the earlier firthlogist package, whose repository is no longer maintained; and because it installs in the same Python 3.11 environment as the rest of the analysis. The fit converged in every fold.

Verification

The leave-one-out predictions and full-cohort coefficients were reproduced with an independently written implementation of the same algorithm: predicted probabilities agreed to within 0.00002 and the AUC was identical (0.703). The earlier firthlogist package (version 0.5.0, run in a separate Python 3.10 environment with scikit-learn 1.5.2) gave the same results: its leave-one-out predicted probabilities agreed with those from firthmodels to within 0.00002 and the AUC was identical.

S5. Definitions of Calibration and Accuracy Measures, and Bootstrap Procedure

Calibration plot

Mean predicted probability was compared with the observed outcome frequency within five quantile-defined bins per model; five bins were chosen deliberately because finer binning is unstable at this sample size.

Calibration-in-the-large

An intercept-only logistic model with the out-of-fold predicted logit entered as a fixed offset (slope fixed at 1). Values near zero indicate no strong systematic over- or under-prediction on average: ridge -0.02 (SE 0.31), random forest -0.34 (SE 0.31), and Firth -0.08 (SE 0.33).

Logistic recalibration intercept and slope

The outcome was regressed on the logit of each model's out-of-fold predicted probability, estimating intercept and slope simultaneously: ridge intercept -0.20 (SE 0.40), slope 0.59 (SE 0.57); random forest intercept -0.31 (SE 0.32), slope 1.25 (SE 0.75); Firth intercept -0.34 (SE 0.33), slope 0.29 (SE 0.27). Slopes below 1 indicate predicted probabilities that were too extreme relative to the observed outcome frequencies; the standard errors are wide.

Brier score and scaled Brier score

The Brier score is the mean squared difference between predicted probability and observed outcome. The scaled Brier score is $1\text{-Brier}_{model}/\text{Brier}_{reference}$, where the reference is the Brier score obtained by predicting the outcome prevalence (38.3%) for every patient (0.236). Scaled Brier scores were 0.038 (ridge), 0.038 (random forest), and 0.059 (Firth), that is, only modest improvement over the prevalence-only reference.

Bootstrap

Patient-level bootstrap resampling was used for uncertainty: 2,000 resamples for AUC and Brier score and 1,000 resamples for standardized coefficients, with the 2.5th and 97.5th percentiles reported as the 95% interval. Paired differences (for example, seven- versus five-predictor AUC, or ridge versus clinical-staging baseline) used the same resamples for both models.

S6. Sensitivity Analyses of Numerical Nodal-Burden Outcomes

Numeric ypN stage

The numeric pathological ypN stage (ypN0-ypN3, coded 0-3, already recorded in the dataset) was modeled with ridge linear regression under the identical cross-validation scheme. This is a numeric-stage sensitivity analysis and not an ordinal regression: a proportional-odds or other genuinely ordinal model would be required for that and was not fitted. The numeric-stage model achieved RMSE 1.09 and R^2 -0.04 (no better than predicting the mean stage for every patient), and its continuous predictions discriminated the binary ypN-positive outcome less well than the primary dichotomous model (AUC 0.62 vs. 0.70), with a wider, not narrower, bootstrap confidence interval (width 0.34 vs. 0.31). In this cohort the numeric surrogate therefore did not improve precision relative to the dichotomous outcome, contrary to the general expectation [37, 38 of the main manuscript]; this may reflect the few distinct levels and the concentration of patients at ypN0 and ypN2 (29 and 11 of 47 patients, respectively) rather than a failure of the underlying statistical principle.

Count and ratio models

The dataset also records the number of metastatic lymph nodes, the number of examined lymph nodes, and the lymph-node ratio for all 47 patients. These were explored as secondary numerical outcomes with count and proportion models: a negative-binomial generalized linear model with a log-examined-node-count offset, and a quasi-binomial model with dispersion-corrected standard errors for the ratio, chosen because of substantial zero inflation (29 of 47 patients with zero metastatic nodes) and overdispersion relative to Poisson (Pearson χ^2/df ≈ 5.4). These are exploratory, hypothesis-generating, full-cohort (not cross-validated) analyses and should not be over-interpreted at n = 47. Nodal counts and ratios are postoperative quantities and were never used as predictors.

Supplementary Table. Final Ridge Model Parameters

Final seven-predictor ridge logistic regression (C = 0.1) fitted to the full 47-patient cohort for descriptive interpretation and for application to new patients. The model is exploratory and has not been externally validated; it should not be used for clinical decisions.

Table A1. Final ridge model parameters

Predictor (order in model)	Unit/coding	Mean (μ)	SD (σ)	Standardized coefficient (β)	OR per SD
Clinical T stage (cT)	Ordinal code 1-3 (cT2-cT4)	2.09	0.50	0.049	1.05
Clinical N stage (cN)	Ordinal code 0-3	1.21	0.82	0.237	1.27
Tumor grade	Ordinal code 1-4	2.57	0.64	0.194	1.21
Age	Years	59.4	9.11	-0.311	0.73
BMI	kg/m²	26.0	4.55	0.060	1.06
WCC	$\times 10^9$/L	7.69	2.16	0.275	1.32
Platelet count	$\times 10^9$/L	285.1	95.3	0.159	1.17

Intercept (β_0) = -0.512

Predictor order: cT, cN, tumor grade, age, BMI, WBC, platelet count. Means and SDs are full-cohort values; SDs are population SDs (ddof = 0, scikit-learn StandardScaler default). Missing tumor-grade values were imputed with the whole-cohort median (grade = 3) for this full-cohort fit; training-fold medians were used inside cross-validation. Model: scikit-learn LogisticRegression, lbfgs solver, max_iter = 1,000; Python 3.11.

Prediction formula

For a patient with predictor values $x_1, x_2, \ldots, x_7$, the predicted probability of ypN-positive status is $\hat{p} = 1/(1 + \exp(-(\beta_0 + \Sigma_j \beta_j \cdot (x_j - \mu_j) / \sigma_j)))$, where j indexes the seven predictors in the order shown and μ_j, σ_j, β_j are the values in **Table A1**. Each predictor is first standardized with the mean and SD shown, then multiplied by its standardized coefficient.

For comparison, the full-cohort Firth (bias-reduced) coefficients on the same standardized scale were cT 0.198 (OR 1.22), cN 0.506 (OR 1.66), tumor grade 0.364 (OR 1.44), age -0.677 (OR 0.51), BMI 0.365 (OR 1.44), WBC 0.426 (OR 1.53), and platelet count 0.208 (OR 1.23); all signs matched the ridge fit, and the larger magnitudes reflect lower shrinkage.