Supplementary Methods

The primary objective was to develop a pre-treatment multivariable prediction model using routinely available preoperative variables to estimate the probability of pathological ≥pT3 invasion, thereby supporting multidisciplinary discussion, intensified preoperative staging, consideration of neoadjuvant therapy, and operative strategy. Variables were selected a priori based on clinical availability before definitive treatment. Conventional EUS uT/uN classifications were recorded to describe baseline staging; however, they were not included as predictors in the primary model to improve portability and to avoid reliance on operator-dependent categorical staging, while preserving a model centred on quantitative EUS-derived thickness and other objective preoperative factors.

Lesion thickness was entered as a continuous predictor and modelled using restricted cubic splines (RCS) with four knots at the 5th/35th/65th/95th percentiles of the observed distribution. This specification was chosen to flexibly capture potential non-linear associations between thickness and the log-odds of ≥pT3. All other predictors entered the model as linear terms or categorical factors with clinically intuitive reference categories. Overall and non-linear effects of thickness were evaluated using Wald χ^2 and likelihood-ratio statistics from an analysis-of-variance decomposition of the fitted model.

Internal validation and optimism correction

Internal validation and correction for potential overfitting were performed using bootstrap resampling (B = 1000). For each bootstrap sample, the full model was refitted and its performance was evaluated both in the bootstrap sample and in the original dataset. The average difference ("optimism") in the C-statistic, calibration slope, Nagelkerke's R^2, and Brier score between bootstrap and test performance was subtracted from the apparent estimates to obtain optimism-corrected values. Apparent and bootstrap-corrected calibration curves were generated using the calibrate function within rms, with predictions grouped into deciles of predicted risk.

Uniform shrinkage and intercept recalibration (final model for implementation)

To improve calibration and reduce overfitting, we applied uniform shrinkage to the final multivariable model. The bootstrap-corrected calibration slope from rms::validate (B=1000) was used as the shrinkage factor (0.9155). All regression coefficients except the intercept were multiplied by this factor. The intercept was then re-estimated by solving for the value that makes the mean predicted probability equal to the observed event rate in the derivation cohort (prevalence = 0.6834). The final shrunk intercept was −4.087426. The exact knot locations for the restricted cubic spline of thickness were 3.691, 8.360, 12.300, and 21.415 mm (5th/35th/65th/95th percentiles).

Alternative model specifications

To assess the added value of covariates beyond EUS thickness and the impact of the functional form of thickness, we evaluated two alternative model specifications:

1. Thickness-only model: logistic regression including RCS-modelled EUS thickness as the sole predictor.
2. Full linear-thickness model: multivariable model including the same set of preoperative predictors as the primary model, but with thickness entered as a single linear term.

For each model, we calculated the C-statistic (AUC) with 95% confidence intervals, Brier score, Nagelkerke's R^2, and calibration intercept and slope. These performance metrics are summarised in Supplementary Table 6.

DeLong tests for pairwise AUC comparisons

Pairwise differences in AUC between models fitted to the same patients were assessed using DeLong's test for correlated ROC curves, implemented in the pROC package in R. We performed two pre-specified comparisons:

1.Full multivariable RCS model vs thickness-only RCS model;

2.Full multivariable RCS model vs full multivariable linear-thickness model.

For each comparison, we report the AUC for both models, the absolute

difference in AUC (ΔAUC), and the corresponding two-sided P value from DeLong's test. These results are reported in the main text and summarised in Supplementary Table 6.

Decision-curve analysis

The potential clinical utility of the full multivariable model was assessed using decision-curve analysis. Net benefit was calculated across a range of risk thresholds and compared with default "treat-all" and "treat-none" strategies, where "treat" was defined as classifying a patient as high risk for ≥pT3 invasion to trigger intensified staging and multidisciplinary treatment planning (including consideration of neoadjuvant therapy and operative strategy). Decision curves were generated using the same set of predictors as in the primary model and were used as a representative example in the main text.

In addition, we derived three representative decision thresholds from the receiver-operating characteristic curve to illustrate how different risk cut-offs might be used in practice. These thresholds represent example risk cut-offs at which clinicians may consider a patient "high risk" for ≥pT3 invasion and therefore escalate pre-treatment planning (e.g., intensified staging, MDT discussion, and consideration of neoadjuvant therapy and operative strategy). For each threshold, we calculated classification measures in the study cohort (Supplementary Table 4). These thresholds are exploratory and should not be interpreted as prescriptive clinical cut-offs without external validation.

Sensitivity analyses

Three sensitivity analyses were performed to assess the robustness of the primary findings:

1. Firth penalised logistic regression: The same set of predictors and RCS specification for thickness were fitted using Firth's bias-reduced logistic regression (logistf package) to mitigate small-sample bias and potential quasi-separation.
2. Mixed-effects logistic regression: A random intercept for individual endoscopists was added to the multivariable model to explore clustering

by operator. Between-operator variance and its 95% confidence interval were estimated.

3. Alternative functional form for thickness: The spline-based model was compared with the full linear-thickness model to assess whether conclusions regarding the importance and shape of the thickness–risk association depended on the chosen functional form.

Performance metrics for these sensitivity models are summarised alongside the primary model in Supplementary Table 6.

Supplementary Individualised Risk Examples

To illustrate how the multivariable model can be applied in pre-treatment evaluation, we present two hypothetical patients with different EUS thickness values and baseline characteristics. In both examples, the predicted probability of ≥pT3 invasion is interpreted in the context of multidisciplinary discussion, intensified staging, consideration of neoadjuvant therapy, and operative strategy, rather than as a stand-alone determinant of any single treatment.

Example 1 – "Borderline" lesion with modest thickness and favourable features

A 62-year-old man presents with a solitary gastric lesion in the middle third of the stomach. Preoperative biopsy shows an intestinal-type, moderately differentiated adenocarcinoma. Serum CEA and CA19-9 levels are within the reference range. On EUS, the tumour appears as a focal wall thickening with preserved layer structure and a maximum lesion thickness of 6 mm.

When these characteristics are entered into the full multivariable model, the predicted probability of ≥pT3 invasion falls within a low-risk range. This estimate may support standard preoperative staging and surgical planning according to institutional practice, while reserving intensified staging and MDT discussion for cases with discordant clinical or imaging findings. The model output is intended to complement routine staging rather than replace pathological assessment.

Example 2 – Thick, diffuse-type lesion with unfavourable markers

A 65-year-old woman is referred with a circumferential lesion in the lower third

of the stomach. Endoscopy shows an irregular, infiltrative tumour consistent with advanced Borrmann morphology. Preoperative biopsy reveals diffuse-type adenocarcinoma with poor differentiation. Serum CA19-9 is mildly elevated, while CEA is normal. EUS demonstrates marked wall thickening with loss of normal layer architecture and a maximum lesion thickness of 12 mm.

For this patient, the full model yields a predicted probability of ≥pT3 invasion in a high-risk range. This estimate may prompt intensified preoperative staging and multidisciplinary discussion regarding consideration of neoadjuvant therapy and operative strategy, in conjunction with nodal and systemic assessment. Final treatment decisions should remain based on comprehensive staging and clinical judgement.

These examples highlight how modest differences in EUS lesion thickness can translate into meaningful differences in predicted risk and may help prioritise further evaluation when findings are discordant. The model is intended to support pre-treatment risk stratification and planning and does not define any single therapeutic indication.

Handling of missing data

Records with missing values in candidate predictors or outcome were excluded during dataset curation (complete-case approach) before model fitting. Therefore, no missing values remained in the final derivation cohort (n = 518) or temporal validation cohort (n = 246). Variable-level missingness in the final analysis datasets is summarised in Supplementary Table 2.

Final model equation for external validation

RCS knots for thickness_mm (mm): 3.691, 8.360, 12.300, 21.415 (5th/35th/65th/95th percentiles).

$$\text{logit}(p) = -4.087426 + 0.460184 \cdot \text{rcs1}(\text{thickness\_mm}) - 0.493957 \cdot \text{rcs2}(\text{thickness\_mm}) + 0.689097 \cdot \text{rcs3}(\text{thickness\_mm}) + 0.485610 \cdot \text{I}(\text{Lauren}=2) + 0.768401 \cdot \text{I}(\text{Lauren}=3) + 0.002170 \cdot \text{age} - 0.239545 \cdot \text{I}(\text{gender}=1) + 0.058594 \cdot \text{I}(\text{site}=2) + 0.109563 \cdot \text{I}(\text{site}=3) + 0.108398 \cdot \text{CEA} + 0.008604 \cdot \text{CA199}$$

$p = 1/(1+\exp(-\text{logit}(p)))$.

（Note: rcs1–rcs3 are the three spline basis terms generated by rms::rcs(thickness_mm, 4) using the knot values above.）

Supplementary Table 1 Final uniformly shrunk coefficients for external validation and implementation of the ≥pT3 prediction model

	Beta (final shrunk)	OR (final)
Intercept	-4.087	N/A
Thickness (spline 1)	0.460	N/A
Thickness (spline 2)	-0.494	N/A
Thickness (spline 3)	0.689	N/A
Lauren: Diffuse-type vs Intestinal-type	0.486	1.628
Lauren: Mixed-type vs Intestinal-type	0.768	2.162
Age	0.002	1.002
Sex: male vs female	-0.240	0.786
Location: middle vs upper	0.059	1.061
Location: lower vs upper	0.110	1.116
CEA	0.108	1.115
CA19-9	0.009	1.009

Supplementary Table 2 Variable-level missingness summary in the final analysis datasets (complete-case derivation and temporal validation cohorts)

Variable	Development cohort (2017–2020), n = 518 Missing n (%)	Temporal validation cohort (2021–2025), n = 246 Missing n (%)	Overall, n = 764 Missing n(%)
Age	0 (0.0)	0 (0.0)	0 (0.0)
Sex	0 (0.0)	0 (0.0)	0 (0.0)
Tumour location	0 (0.0)	0 (0.0)	0 (0.0)
Lauren classification	0 (0.0)	0 (0.0)	0 (0.0)
EUS lesion thickness, mm	0 (0.0)	0 (0.0)	0 (0.0)
CEA	0 (0.0)	0 (0.0)	0 (0.0)
CA19-9	0 (0.0)	0 (0.0)	0 (0.0)
Pathological T stage (pT)	0 (0.0)	0 (0.0)	0 (0.0)

Footnote: Missingness was assessed for all candidate predictors and the outcome in the final analysis datasets used for model development and temporal validation. Records with missing values were excluded during dataset curation before model fitting (complete-case approach). Therefore, no missing values remained in the final derivation cohort (n = 518) or temporal validation cohort (n = 246).

Supplementary Table 3 Original (unshrunk) regression coefficients of the full multivariable model for pathological ≥pT3 gastric cancer (for reference)

	Beta	SE	Z	P value	OR	Lower CI	Upper CI
Intercept	-4.513	1.039	-4.342	<0.001	0.011	0.001	0.084
Thickness (spline 1)	0.503	0.121	4.150	<0.001	N/A	N/A	N/A
Thickness (spline 2)	-0.540	0.547	-0.986	0.324	N/A	N/A	N/A
Thickness (spline 3)	0.753	1.503	0.501	0.617	N/A	N/A	N/A
Lauren: Diffuse-type vs Intestinal-type	0.530	0.307	1.729	0.084	1.700	0.932	3.100
Lauren: Mixed-type vs Intestinal-type	0.839	0.311	2.696	0.007	2.315	1.258	4.261
Age	0.002	0.011	0.223	0.824	1.002	0.982	1.023
Sex: male vs female	-0.262	0.248	-1.057	0.290	0.770	0.474	1.250
Location: middle vs upper	0.064	0.334	0.192	0.848	1.066	0.554	2.051
Location: lower vs upper	0.120	0.351	0.341	0.733	1.127	0.567	2.240
CEA	0.118	0.048	2.493	0.013	1.126	1.026	1.236
CA19-9	0.009	0.006	1.496	0.135	1.009	0.997	1.022

Note: This appendix table reports the original (unshrunk) regression estimates primarily for interpretability and statistical inference. For external validation and clinical implementation, the final uniformly shrunk coefficients with re-

estimated intercept (implementation-ready model) are provided in Supplementary Table 1.

Supplementary Table 4 Temporal validation of the final prediction model

	Temporal validation
N	246
Events	116
Event Rate	47.15%
AUC	0.845
AUC_L	0.798
AUC_U	0.893
Brier	0.166
Cal_Intercept	-0.466
Cal_Slope	0.623

Supplementary Table 5 Performance of the primary and sensitivity models for preoperative prediction of pathological ≥pT3 gastric cancer

Model	Full multivariable (RCS thickness)	Thickness-only (RCS thickness)	Full multivariable (linear thickness)	Firth penalised multivariable (RCS thickness)
N	518	518	518	518
AUC	0.851	0.825	0.847	0.839
AUC lower	0.817	0.785	0.812	0.803
AUC upper	0.885	0.864	0.881	0.875
Brier	0.142	0.150	0.144	0.146
Nagelkerke's R2	0.444	0.381	0.429	N/A
Cal intercept app	0	0	0	-0.003
Cal slope	1	1	1	1.052
Cal intercept corrected	0.041	N/A	N/A	N/A
Cal slope corrected	0.916	N/A	N/A	N/A
AUC corrected	0.838	N/A	N/A	N/A
Delta AUC vs full	N/A	0.026	0.004	N/A
P value (DeLong vs full model)	N/A	0.001	0.036	N/A

Supplementary Table 6 Classification consequences of three representative decision thresholds derived from the full multivariable model for pathological ≥pT3 gastric cancer

Strategy	Balanced (Youden)	High-sensitivity threshold (Se ≥0.90)	High-specificity threshold (Sp ≥0.90)
Predicted risk threshold	0.784	0.476	0.823
Approximate thickness (mm)	12.6	8.1	13.9
Sensitivity	0.664	0.901	0.602
Specificity	0.866	0.543	0.902
PPV	0.914	0.81	0.93
NPV	0.544	0.718	0.512
TP	235	319	213
FN	119	35	141
FP	22	75	16
TN	142	89	148
Proportion high_risk	0.496	0.761	0.442