

A clinical-EUS model for GIST probability in resection-selected gastric subepithelial lesions: development and external validation

Jiahuang Huang

The First Affiliated Hospital of Shenzhen University

Chong Chen

The First Affiliated Hospital of Shenzhen University

Fei Li

The First Affiliated Hospital of Shenzhen University

Dayong Sun

The First Affiliated Hospital of Shenzhen University

Shuo Yang

Peking University Shenzhen Hospital

Hongqian Tao

Peking University Shenzhen Hospital

Xi Li

Peking University Shenzhen Hospital

Chaohui Liu

1103249411@qq.com

The First Affiliated Hospital of Shenzhen University

Article

Keywords: gastrointestinal stromal tumor, gastric subepithelial lesion, endoscopic ultrasound, external validation

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11094894/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Preoperative distinction of gastrointestinal stromal tumors (GISTs) from other gastric subepithelial lesions remains difficult. We evaluated a three-variable clinical and endoscopic ultrasound (EUS) model for GIST probability in patients undergoing endoscopic resection. This retrospective study included 546 development patients. The three-variable logistic model used age, fundic location, and muscularis propria origin in 528 patients with evaluable predictors. Internal validation used 2000 bootstrap resamples. The original coefficients were tested without refitting in 80 patients from a second hospital in the same city (34 GISTs). Apparent and optimism-corrected areas under the receiver operating characteristic curve were 0.805 and 0.803. External discrimination was 0.783 (95% CI 0.680–0.885), with Brier score 0.192, calibration intercept – 0.339 (95% CI – 0.849 to 0.172), and slope 0.823 (95% CI 0.408–1.237). Adding tumor size > 15 mm increased AUC by 0.007 ($P = 0.133$). In this small external cohort, the model retained moderate discrimination but calibration was uncertain. It may help estimate GIST probability when EUS wall-layer origin is clearly assigned, but it should not be used alone to determine surveillance, biopsy, or resection.

Introduction

Gastric subepithelial lesions encompass mesenchymal tumors, ectopic tissue, and other pathological mimics. A multicenter screening study found upper gastrointestinal subepithelial lesions in 1.60% of examinees, with the stomach the most frequent site; its pathological spectrum included leiomyoma, gastrointestinal stromal tumor (GIST), and ectopic pancreas [1]. GIST requires particular consideration because of its malignant potential, but the diagnostic question of GIST versus non-GIST differs from prognostic risk assessment after GIST is established [2, 3].

Endoscopic ultrasound (EUS) characterizes lesion size, location, wall-layer origin, and morphology, yet cannot reliably distinguish all GISTs from their mimics [4–6]. A quantitative probability estimate could complement EUS findings when counseling patients already being evaluated for resection.

Prediction-model evaluation requires calibration and overall error alongside discrimination and internal and external validation [7–12]. Earlier EUS-based models and image-based artificial-intelligence systems illustrate the diagnostic information available from sonography [13–16]. Our aim was to evaluate a parsimonious three-variable logistic equation using continuous age, binary fundic location, and binary muscularis-propria origin. We applied the original coefficients unchanged at a second hospital in the same city and assessed discrimination, calibration, and prediction error.

Methods

Study design and populations

This retrospective diagnostic prediction-model study included consecutive patients undergoing endoscopic resection for gastric subepithelial lesions at two hospitals in Shenzhen, China. The

development cohort at Shenzhen Second People's Hospital covered January 2016–September 2025; the independent same-city external cohort at Peking University Shenzhen Hospital covered January 2020–January 2026. Each patient contributed one resected lesion with histopathological confirmation. Cohort sizes reflected all available eligible cases rather than a prospective sample-size calculation: 546 development patients and 80 external patients. The model-development set contained 234 GIST events for three predictor terms; external evaluation contained 34 events. These resection-selected cohorts do not represent all incidentally detected gastric lesions.

The study was approved by the Ethics Committee of Shenzhen Second People's Hospital (No. 2026-182-01PJ), which waived written informed consent for retrospective use of deidentified clinical data. The external-center dataset was used under approval from the institutional ethics committee of Peking University Shenzhen Hospital, with written informed consent likewise waived for retrospective use of deidentified clinical data. Both datasets were deidentified.

Outcome and predictor assessment

The outcome was histopathological GIST in the resection specimen, coded 1 versus non-GIST coded 0. This was diagnosis of an existing lesion, without a future prediction horizon. The archived subgroup definition prioritized any GIST component; among remaining lesions, any leiomyoma component defined the leiomyoma group. The exploratory GIST-versus-leiomyoma analysis applied the same M0 equation without refitting.

At both centers, endosonographers had more than five years of EUS experience. Equipment manufacturers included InnerMedical, Lepu Medical Technology, and Olympus. EUS findings were recorded before resection and pathological diagnosis. Both centers used the same core definitions and coding: age10 = age in years/10; fundus = 1 for fundic location and 0 otherwise; mp_origin = 1 for muscularis propria origin and 0 for other reliably assigned layers. Location, wall-layer origin, and tumor size were abstracted from routine preoperative reports (Supplementary Table S2). External validation introduced no center-specific reinterpretation or post hoc recoding. Formal assessor blinding and outcome-assessor qualifications were not recorded.

Model specification and missing data

The primary model M0 used three clinically selected predictors based on routine availability, anatomical plausibility, and interpretability. A working protocol (V1.0, 5 June 2026) and the signed two-center protocol (V1.1, 30 June 2026) document the retained specification. Because earlier analysis files already contained the fitted model, we do not describe the analysis as prospectively prespecified. M0 was retained without variable reselection or coefficient changes (Supplementary File S1).

Age and fundic location were complete. In 18 development patients, an indistinct lesion–surrounding tissue interface or suspected adherence prevented reliable EUS layer assignment. These findings were clinically indeterminate, not negative for muscularis propria origin. The patients remained in the descriptive cohort but were excluded from M0 by complete-case analysis, leaving 528 patients (234 GISTs; 294 non-GIST lesions). No imputation was performed because a wall-layer assignment had not been observed. The missingness mechanism was not assumed to be random, and complete-case exclusion may introduce selection bias. The model is not applicable when this input is unreliable. Other comparisons used available cases.

Statistical analysis

The archived supplementary comparison lists 61 preoperative variables. Four univariable results absent from the archived comparison table were reproduced from archived source data using the original test framework (BMI, $P=0.347$; white blood cell count, $P=0.919$; alpha-fetoprotein, $P=0.155$; carbohydrate antigen 19-9, $P=0.065$); workflow details are provided in Supplementary File S1. Benjamini–Hochberg correction was applied across all 61 P values. These comparisons did not serve as a screening procedure to select the three M0 predictors. Derived indicators were not separately added when their parent categories were already represented. Continuous comparisons used Welch’s t test or Mann–Whitney U testing; categorical comparisons used Pearson chi-square, Fisher exact, or fixed-seed Monte Carlo exact procedures, as appropriate (Supplementary Table S1).

M0 was developed by multivariable logistic regression. Its locked linear predictor was:

$$\text{logit}[P(\text{GIST})] = -5.747204 + 0.632798 \times \text{age}_{10} + 1.951026 \times \text{fundus} + 1.634058 \times \text{mp_origin}.$$

Where lp denotes the linear predictor above, $P(\text{GIST}) = 1/\{1 + \exp(-lp)\}$. Internal validation used 2000 bootstrap resamples (seed 20260629), refitting within each resample to estimate optimism in the original analysis. Performance measures were the area under the receiver operating characteristic curve (AUC), DeLong 95% confidence interval (CI), optimism-corrected AUC, calibration-in-the-large, calibration intercept and slope, grouped calibration, and Brier scores. The external cohort was evaluated with the locked formula without refitting, variable reselection, or primary recalibration.

Exploratory decision-curve analysis compared model net benefit with treat-all and treat-none across probabilities 0.05–0.80, using a 0.01 grid and strict greater-than comparisons. Fixed reporting strata were low ($P<0.30$), intermediate ($0.30 \leq P<0.70$), and high ($P \geq 0.70$). Observed proportions used Wilson 95% CIs; ordered trends used the Cochran–Armitage test. These thresholds were not optimized or intended as treatment rules.

The archived tumor-size extension assessed >15 mm in 516 complete cases with evaluable M0 predictors and tumor-size data, 12 fewer than the M0 analysis set. Comparisons with M0 used the same 516 patients, paired DeLong testing [17], likelihood-ratio testing, and Brier-score changes where archived.

The size cutoff and $\Delta\text{AUC} \geq 0.01$ benchmark were pragmatic exploratory choices, not validated clinical thresholds. A secondary exploratory assessment against leiomyoma, the most common mimic, applied the same M0 coefficients. Original analyses used Python 3.11 and standard scientific libraries; statistical details are in the Supporting Information. The descriptive pathology supplement reports counting of the 307 original non-GIST pathology records, with conservative synonym merging and one count per mixed diagnosis.

AI-assisted manuscript preparation. ChatGPT (OpenAI) was used for English-language polishing of the manuscript. All authors reviewed the final text and take responsibility for its content. ChatGPT was not listed as an author.

Code availability. The statistical analysis code is available from the corresponding author upon reasonable request, subject to institutional approval and applicable privacy regulations.

Results

Cohorts and pathological spectrum

The development cohort comprised 546 patients: 239 GISTs (43.8%) and 307 non-GIST lesions (56.2%). M0 included 528 patients after exclusion of 18 clinically indeterminate layers (Figure 1). Baseline characteristics appear in Table 1. Six of the 61 comparisons remained significant after false-discovery-rate correction: age, hypertension, tumor location, layer of origin, echo intensity, and hemoglobin (Supplementary Table S1).

The 307-record pathology source contained 235 pure leiomyomas (76.5%), 21 ectopic-pancreas diagnoses (6.8%), seven pure lipomas (2.3%), and five neuroendocrine tumors (1.6%). The remaining labels included other mesenchymal lesions, epithelial neoplasia, inflammatory findings, and combined diagnoses. Thus, non-GIST did not mean uniformly benign. Complete original-label mapping and mutually exclusive counts are provided in Supplementary Table S7. These counts describe only the full development non-GIST cohort, not the 294 M0 non-GIST patients or 46 external non-GIST patients.

Model performance and risk strata

Age per 10 years (odds ratio [OR] 1.883, 95% CI 1.540–2.302), fundic location (OR 7.036, 4.529–10.930), and muscularis propria origin (OR 5.125, 2.305–11.394) were associated with GIST (all $P < 0.001$; Table 2A). The apparent AUC was 0.805 (95% CI 0.768–0.842); optimism-corrected AUC was 0.803. Mean bootstrap training and test AUCs were 0.806 and 0.804, respectively. Corrected calibration intercept and slope were -0.004 and 0.983 . The Brier score was 0.176 (null 0.247; scaled 0.286). Apparent calibration was close to ideal, as expected on development data (Table 2B; Figure 3).

The low-, intermediate-, and high-risk groups contained 190, 205, and 133 patients, respectively. Observed GIST rates were 20.0% (95% CI 14.9%–26.3%), 41.0% (34.5%–47.8%), and 84.2% (77.1%–89.4%) (Table 2D; Supplementary Figure S1). The ordered trend was significant ($Z=11.23$, $P<0.001$). Development decision curves favored the model over both default strategies across the evaluated range (Figure 3C; Supplementary Table S5).

Secondary analyses

Among 516 patients with tumor-size data (233 GISTs; 283 non-GIST lesions), same-sample AUCs were 0.805 for M0 and 0.812 for its size extension ($\Delta\text{AUC}=+0.007$; DeLong $P=0.133$). The likelihood-ratio test favored the extension ($P=0.013$), and >15 mm was associated with GIST (OR 2.655, 95% CI 1.218–5.791; $P=0.014$), but discrimination gain remained below the exploratory 0.01 benchmark (Figure 2; Supplementary Table S6).

Comparisons of the 18 indeterminate-layer patients with evaluable patients found no significant differences in age, sex, GIST proportion, fundic location, or hypertension. The excluded group was too small to rule out selection bias (Supplementary Table S3).

The secondary exploratory GIST-versus-leiomyoma subgroup contained 459 patients (234 GISTs; 225 leiomyomas). Without coefficient refitting, AUC was 0.774 (95% CI 0.731–0.817), Brier score 0.192, scaled Brier score 0.231, calibration-in-the-large 0.142, and calibration slope 0.946 (Supplementary Table S4; Figure 5A).

Preliminary external validation

The independent same-city external center contributed 80 patients (34 GISTs; 46 non-GIST lesions). External AUC was 0.783 (95% CI 0.680–0.885). Calibration-in-the-large was -0.366 (95% CI -0.897 to 0.166), calibration intercept -0.339 (-0.849 to 0.172), and slope 0.823 (0.408 – 1.237). Brier score was 0.192 (null 0.244; scaled 0.216). The negative calibration intercept indicates average overprediction, and the wide intervals indicate limited precision (Table 2C; Figure 4).

External strata contained 25, 32, and 23 patients, with GIST rates of 16.0%, 43.8%, and 69.6% ($Z=3.75$, $P<0.001$; Table 2D; Figure 5B). Exploratory model net benefit exceeded both default strategies at 0.05–0.11, 0.13–0.69, and 0.71 on the evaluated 0.01 grid (Supplementary Table S5). These pointwise results are exploratory and do not define clinical thresholds. Available development–external descriptors are compared in Supplementary Table S8; complete external predictor-distribution summaries were unavailable, so case-mix contributions to the observed calibration shift could not be quantified.

Discussion

This three-variable clinical-EUS equation showed minimal bootstrap optimism and retained moderate discrimination at a second hospital. Because the full equation, calibration statistics, and Brier score are reported, its probability estimates can be assessed directly when EUS wall-layer origin is reliable.

External point estimates suggested average overprediction and a tendency toward overly extreme predictions. With 80 patients and 34 GIST events, the wide calibration intervals leave substantial uncertainty about calibration in other populations.

Tumor size retained an association with diagnosis but added little discrimination in this resection-selected population with predominantly small lesions. This finding does not conflict with size-based GIST management recommendations [2–6]; the diagnostic role of size here differs from its role in prognostic assessment after GIST is established. The leiomyoma subgroup result cannot be extrapolated to every non-GIST subtype. The pathology source includes rare malignant and non-mesenchymal diagnoses, making “non-GIST” an outcome category rather than a uniform benign comparator.

No reporting stratum provides a treatment mandate. GIST occurred in 20.0% of development and 16.0% of external low-risk patients; a low score cannot rule it out. High-risk enrichment likewise does not automatically justify resection. Exploratory decision curves describe model net benefit within these selected cohorts but do not define thresholds for surveillance, sampling, or intervention [12].

Several limitations warrant consideration. Both cohorts were retrospective and selected for endoscopic resection, so treatment-selection and spectrum bias are likely; patients managed by surveillance, sampling, surgery, or no intervention may differ. The external cohort was small and from the same city, and complete predictor-distribution summaries were unavailable, precluding a formal case-mix comparison. The non-GIST group was heterogeneous, and no additional subtype-specific AUCs were estimated. The separate 307-record pathology description was not used to infer the complete-case pathology composition. Temporal stability across the long accrual period and performance across demographic groups were not formally assessed.

EUS layer assignment can vary between examiners [18], but interobserver and intraobserver reliability were not measured. Layer of origin was indeterminate in 18 patients, and their exclusion may have introduced bias despite the absence of significant baseline differences. Prospective work should standardize EUS acquisition and explicitly record uncertainty.

The predictors were clinically selected rather than prospectively prespecified; the protocol chronology and component-based mixed-pathology rule are detailed in Methods and Supplementary File S1.

In conclusion, the original three-variable model retained moderate discrimination in a small external cohort, although calibration remained uncertain. It may help clinicians discuss GIST probability in patients selected for resection when EUS layer of origin is reliable, but prospective multicenter evaluation is needed before routine use or treatment decisions.

Declarations

Author contributions

Jiahuang Huang: Conceptualization, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. Chong Chen: Conceptualization, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. Fei Li: Investigation, Data curation, Validation. Dayong Sun: Investigation, Data curation, Validation. Shuo Yang: Data curation, Validation. Hongqian Tao: Data curation, Validation. Chaohui Liu: Supervision, Methodology, Writing – review & editing. Xi Li: Supervision, Methodology, Writing – review & editing. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

Deidentified participant data and statistical analysis code are available from the corresponding author upon reasonable request, subject to institutional approval and applicable privacy regulations.

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Shenzhen Second People's Hospital and by the institutional ethics committee of Peking University Shenzhen Hospital for the external-center retrospective validation dataset(No. 2026-182-01PJ). Written informed consent was waived by the respective institutional ethics committees for retrospective use of deidentified clinical data. All methods were performed in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Use of generative AI in manuscript preparation

The use of ChatGPT for English-language polishing is disclosed in the Methods section; the authors reviewed the final manuscript and retain full responsibility for its content.

Study registration

This retrospective observational study was not prospectively registered.

Patient and public involvement

Patients and members of the public were not involved in the design, conduct, reporting, interpretation, or dissemination planning of this study.

References

1. Choe Y, Cho YK, Kim GH, et al. Prevalence, natural progression, and clinical practices of upper gastrointestinal subepithelial lesions in Korea: a multicenter study. *Clin Endosc* 56, 744–753 (2023). doi:10.5946/ce.2023.005.
2. Casali PG, Blay JY, Abecassis N, et al. Gastrointestinal stromal tumours: ESMO-EURACAN-GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 33, 20–33 (2022). doi:10.1016/j.annonc.2021.09.005.
3. Serrano C, Martín-Broto J, Asencio-Pascual JM, et al. 2023 GEIS Guidelines for gastrointestinal stromal tumors. *Ther Adv Med Oncol* 15, 17588359231192388 (2023). doi:10.1177/17588359231192388.
4. Deprez PH, Moons LMG, O'Toole D, et al. Endoscopic management of subepithelial lesions including neuroendocrine neoplasms: European Society of Gastrointestinal Endoscopy Guideline. *Endoscopy* 54, 412–429 (2022). doi:10.1055/a-1751-5742.
5. Jacobson BC, Bhatt A, Greer KB, et al. ACG Clinical Guideline: Diagnosis and Management of Gastrointestinal Subepithelial Lesions. *Am J Gastroenterol* 118, 46–58 (2023). doi:10.14309/ajg.0000000000002100.
6. Sharzehi K, Sethi A, Savides T. AGA Clinical Practice Update on Management of Subepithelial Lesions Encountered During Routine Endoscopy: Expert Review. *Clin Gastroenterol Hepatol* 20, 2435–2443.e4 (2022). doi:10.1016/j.cgh.2022.05.054.
7. 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* 385, e078378 (2024). doi:10.1136/bmj-2023-078378.
8. Efthimiou O, Seo M, Chalkou K, et al. Developing clinical prediction models: a step-by-step guide. *BMJ* 386, e078276 (2024). doi:10.1136/bmj-2023-078276.
9. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 384, e074819 (2024). doi:10.1136/bmj-2023-074819.
10. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* 384, e074820 (2024). doi:10.1136/bmj-2023-074820.
11. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 17, 230 (2019). doi:10.1186/s12916-019-1466-7.
12. Vickers AJ, van Calster B, Steyerberg EW. Decision curve analysis to evaluate the clinical benefit of prediction models. *Spine J* 21, 1643–1648 (2021). doi:10.1016/j.spinee.2021.02.024.

13. Kim SM, Kim EY, Cho JW, et al. Predictive factors for differentiating gastrointestinal stromal tumors from leiomyomas based on endoscopic ultrasonography findings in patients with gastric subepithelial tumors: a multicenter retrospective study. Clin Endosc 54, 872–880 (2021). doi:10.5946/ce.2021.251.
14. Hirai K, Kuwahara T, Furukawa K, et al. Artificial intelligence-based diagnosis of upper gastrointestinal subepithelial lesions on endoscopic ultrasonography images. Gastric Cancer 25, 382–391 (2022). doi:10.1007/s10120-021-01261-x.
15. Lu Y, Wu J, Hu M, et al. Artificial intelligence in the prediction of gastrointestinal stromal tumors on endoscopic ultrasonography images: development, validation and comparison with endosonographers. Gut Liver 17, 874–883 (2023). doi:10.5009/gnl220347.
16. Cho IR, Park JC, Roh YH, et al. Noninvasive prediction model for diagnosing gastrointestinal stromal tumors using contrast-enhanced harmonic endoscopic ultrasound. Dig Liver Dis 51, 985–992 (2019). doi:10.1016/j.dld.2019.02.017.
17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845 (1988).
18. Gress F, Schmitt C, Savides T, et al. Interobserver agreement for EUS in the evaluation and diagnosis of submucosal masses. Gastrointest Endosc 53, 71–76 (2001). doi:10.1067/mge.2001.111384.

Tables

Table 1 Selected baseline characteristics of the development cohort by GIST status

Variable	GIST	Non-GIST	Available/ Missing N	Statistical test	Effect estimate (95% CI)	Absolute SMD	P value
Age	56.8 ± 9.8	50.1 ± 11.3	546/0	Welch t- test	6.7 (4.9– 8.5)	0.634	<0.001
Female sex	145/239 (60.7)	212/307 (69.1)	546/0	Pearson χ ²	OR 0.691 (0.485– 0.986)	0.176	0.0410
Hypertension	61/239 (25.5)	43/307 (14.0)	546/0	Pearson χ ²	OR 2.104 (1.363– 3.248)	0.293	<0.001
Albumin	41.5 ± 2.5	41.7 ± 2.9	528/18	Welch t- test	−0.15 (−0.62– 0.32)	0.054	0.5322
Tumor size	7.6 [5.6– 10.9]	7.7 [6.0– 10.0]	516/30	Mann– Whitney U	−0.01	—	0.9204
Fundic location	137/239 (57.3)	46/307 (15.0)	546/0	Pearson χ ²	OR 7.621 (5.083– 11.425)	0.897	<0.001
Muscularis propria origin	224/234 (95.7)	235/294 (79.9)	528/18	Pearson χ ²	OR 5.624 (2.807– 11.266)	0.469	<0.001

Development N=546 (GIST 239; non-GIST 307). Values are mean ± SD, median [IQR], or n/N (%). Age is in years, tumor size in mm, and albumin in g/L. Available/Missing N refers to variable-specific availability. Effect estimates in Table 1 are univariable; Table 2A reports multivariable-adjusted model coefficients and ORs. P values are retained from the archived analysis; Table S1 contains multiplicity-adjusted results. SMD denotes standardized mean difference; Table 1 reports absolute SMD.

Table 2 Development and validation performance of the locked M0 model

Panel A Model coefficients

Predictor	β	SE	Wald χ ²	OR (95% CI)	P value
Intercept	−5.747204	0.7361	60.96	0.003 (0.001–0.014)	<0.001
Age, per 10 years	0.632798	0.1025	38.11	1.883 (1.540–2.302)	<0.001
Fundic location	1.951026	0.2247	75.39	7.036 (4.529–10.930)	<0.001
Muscularis propria origin	1.634058	0.4077	16.06	5.125 (2.305–11.394)	<0.001

age10 = age/10. Binary predictors are coded 1 when present and 0 when absent; indeterminate layer is not coded 0. The full six-decimal equation is given in Methods.

Panel B Internal validation

Metric	Value	95% CI
Apparent AUC	0.805	0.768–0.842
Optimism-corrected AUC	0.803	
AUC standard error	0.019	
Brier score	0.176	
Scaled Brier score	0.286	
Apparent calibration-in-the-large (intercept)	0.000	–0.204–0.204
Apparent calibration slope (free slope)	1.000	0.817–1.183
Optimism-corrected intercept	–0.004	
Optimism-corrected slope	0.983	

Table 2 continued

Panel C Preliminary same-city external validation

Metric	Value	95% CI
External cohort N	80	
AUC	0.783	0.680–0.885
AUC standard error	0.052	
Brier score	0.192	
Scaled Brier score	0.216	
Calibration-in-the-large (intercept)	–0.366	–0.897–0.166
Calibration intercept (free slope)	–0.339	–0.849–0.172
Calibration slope (free slope)	0.823	0.408–1.237

Table 2 continued

Panel D Fixed reporting strata

Risk group	Threshold	Local N (GIST/non- GIST)	Local observed rate (95% CI)	External N (GIST/non- GIST)	External observed rate (95% CI)
Low	$P < 0.30$	190 (38/152)	20.0% (14.9%– 26.3%)	25 (4/21)	16.0% (6.4%– 34.6%)
Intermediate	$0.30 \leq P < 0.70$	205 (84/121)	41.0% (34.5%– 47.8%)	32 (14/18)	43.8% (28.2%– 60.7%)
High	$P \geq 0.70$	133 (112/21)	84.2% (77.1%– 89.4%)	23 (16/7)	69.6% (49.1%– 84.4%)

Cochran–Armitage trend: development $Z=11.23$, $P<0.001$; external $Z=3.75$, $P<0.001$. Intervals for observed rates are Wilson 95% CIs. These strata are for reporting and communication, not treatment selection.

Panel E Exploratory decision curves

At the 0.01 threshold grid, development net benefit exceeded zero and both default strategies throughout 0.05–0.80. External net benefit exceeded zero at 0.05–0.69 and 0.71; exceeded treat-all at 0.05–0.11 and 0.13–0.80; exceeded treat-none at 0.05–0.69 and 0.71; and exceeded both at 0.05–0.11, 0.13–0.69, and 0.71. Detailed intervals appear in Supplementary Table S5. These pointwise results are exploratory and do not define clinical thresholds.

Figures

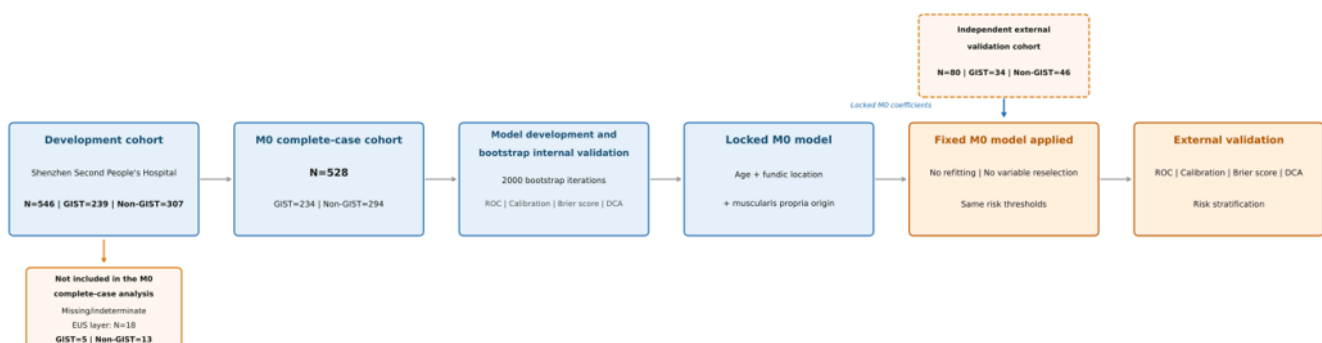


Figure 1

Study flow and model validation. The development cohort comprised 546 consecutive patients who underwent endoscopic resection for gastric subepithelial lesions, with histopathological examination of the resected specimens serving as the reference standard. The M0 complete-case cohort included 528

patients (234 GISTs and 294 non-GIST lesions). In 18 patients, the EUS layer of origin was clinically indeterminate and these patients were not included in the M0 complete-case analysis. The M0 model used age, fundic location, and muscularis propria origin and underwent 2000-resample bootstrap internal validation. The original coefficients were then applied unchanged to an independent same-city external cohort of 80 consecutively endoscopically resected patients (34 GISTs and 46 non-GIST lesions).

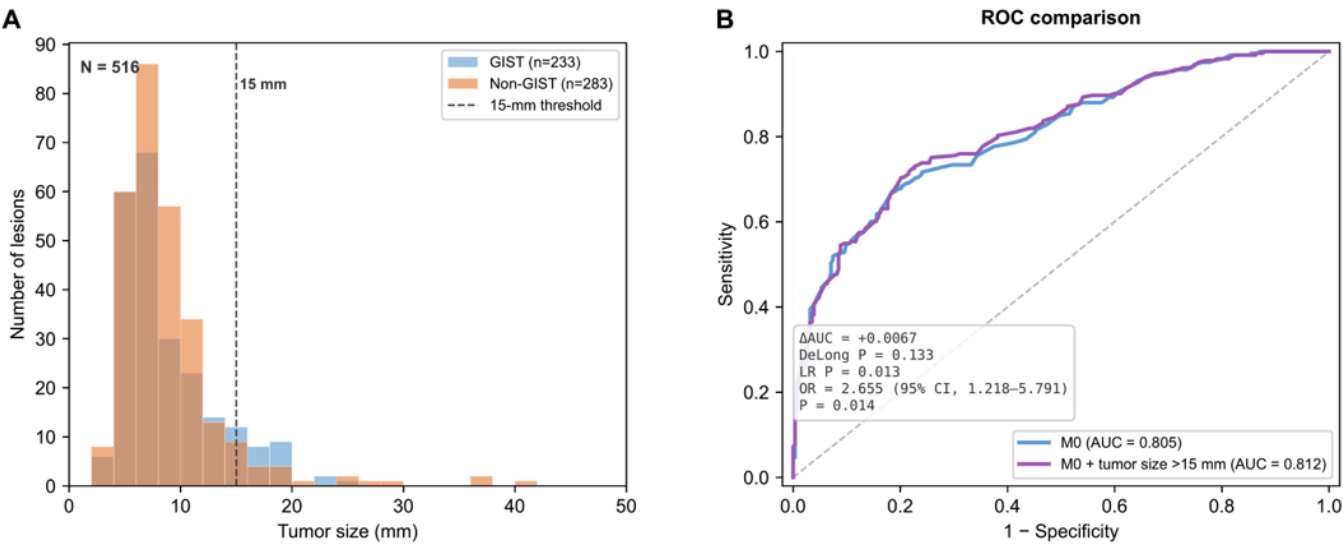


Figure 2

Tumor-size distribution and incremental value of tumor size >15 mm. Tumor size was evaluated in 516 patients with available size data (233 GISTs and 283 non-GIST lesions). Panel A shows the tumor-size distribution according to GIST status and the pragmatic 15-mm exploratory threshold. Panel B shows patient-level ROC curves for M0 and M0 plus tumor size >15 mm. M0 and the extended model had AUCs of 0.805 and 0.812, respectively ($\Delta AUC = +0.007$; DeLong $P = 0.133$; likelihood-ratio $P = 0.013$). Tumor size >15 mm was associated with GIST (OR=2.655, 95% CI 1.218–5.791; $P = 0.014$), but the increase in discrimination did not reach the pragmatic exploratory $\Delta AUC \geq 0.01$ benchmark.

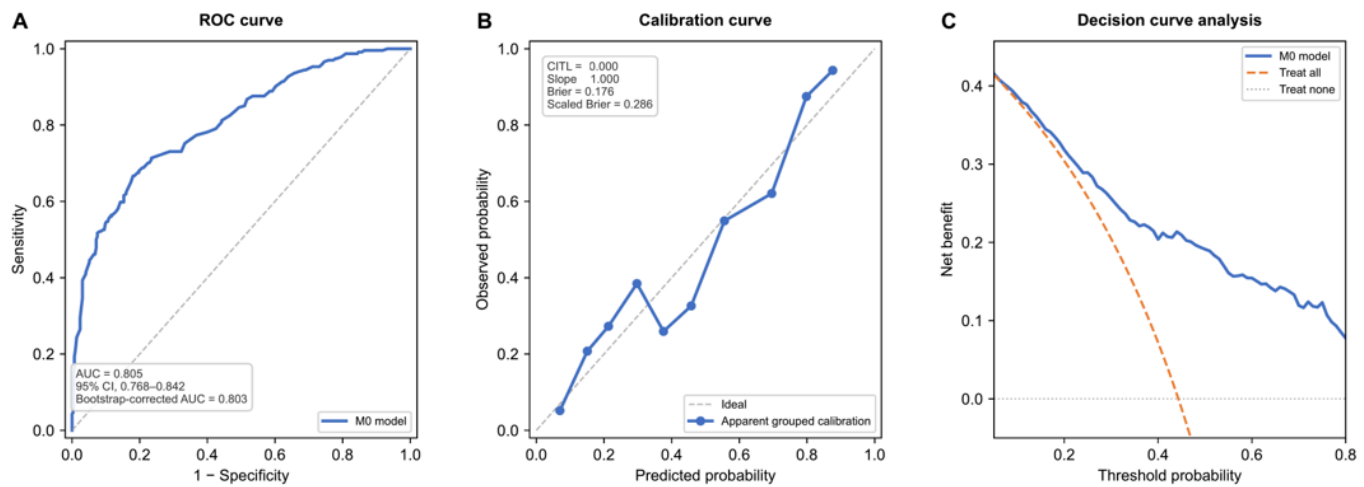


Figure 3

Internal performance of the M0 model. (A) Receiver operating characteristic curve showing an apparent AUC of 0.805 (DeLong 95% CI 0.768–0.842) and an optimism-corrected AUC of 0.803, shown as a point estimate. (B) Apparent grouped calibration is displayed; the optimism-corrected calibration intercept and slope were –0.004 and 0.983, respectively. The apparent Brier score was 0.176 and the scaled Brier score was 0.286. (C) Exploratory decision curve analysis showed positive net benefit and greater net benefit than both default strategies across threshold probabilities from 0.05 to 0.80.

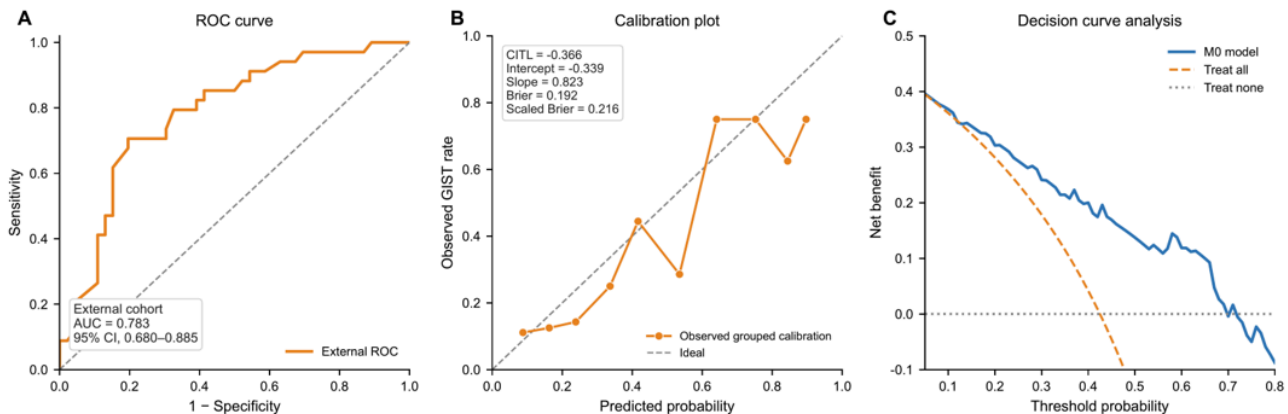


Figure 4

External validation of the M0 model. The independent same-city external cohort included 80 patients (34 GISTs and 46 non-GIST lesions). (A) The AUC was 0.783 (DeLong 95% CI 0.680–0.885). (B) Observed grouped calibration showed calibration-in-the-large of –0.366 (95% CI –0.897 to 0.166), calibration intercept of –0.339 (95% CI –0.849 to 0.172), calibration slope of 0.823 (95% CI 0.408–1.237), Brier score of 0.192, and scaled Brier score of 0.216. The wide calibration intervals reflect limited precision with 34 GIST events. (C) On the pointwise 0.01 threshold grid, model net benefit was positive at 0.05–

0.69 and 0.71; exceeded treat-all at 0.05–0.11 and 0.13–0.80; exceeded treat-none at 0.05–0.69 and 0.71; and exceeded both default strategies at 0.05–0.11, 0.13–0.69, and 0.71.

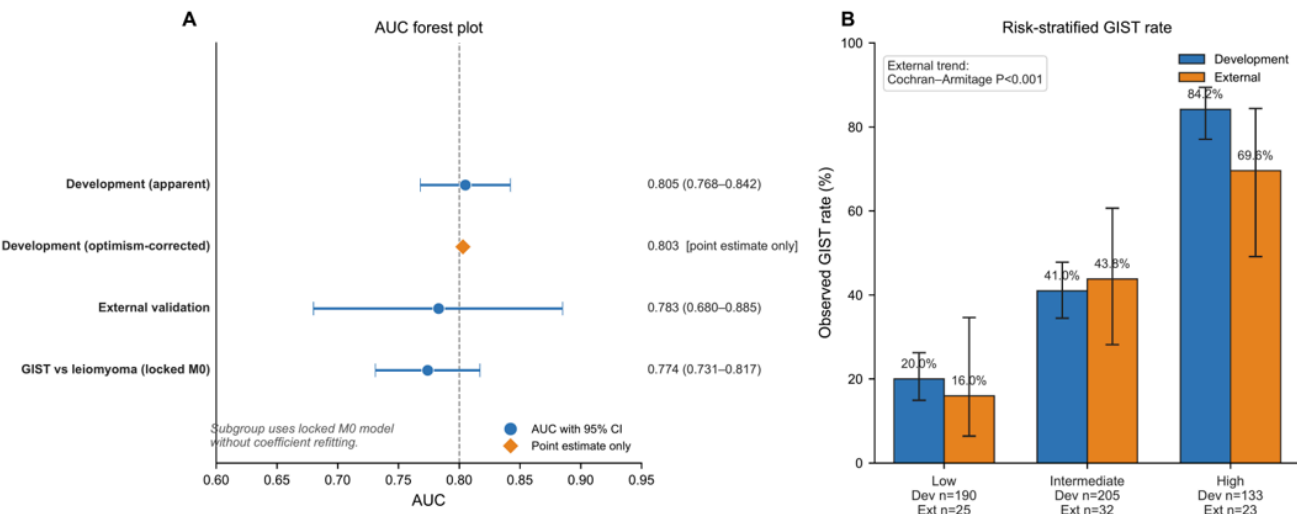


Figure 5

Discrimination and risk-stratification comparison across development, same-city external validation, and subgroup analyses. The left panel compares the development apparent AUC, the optimism-corrected development AUC shown as a point estimate, the external validation AUC, and the GIST-versus-leiomyoma subgroup AUC. The subgroup analysis used the M0 model without coefficient refitting. The right panel compares observed GIST rates across the same low-, intermediate-, and high-risk strata in the development and external cohorts; error bars represent Wilson 95% confidence intervals. The dashed vertical line marks AUC=0.80 as a visual reference.

Figures 1–5 are provided as separate, high-resolution figure files. Supplementary Figure S1 and Supplementary File S1 are included in the combined Supplementary Information.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)