

Spatial artificial intelligence in gastric endoscopy for premalignant lesions and early gastric cancer: a systematic review

Mohammad Kazem Amirbeigy

Shahid Sadoughi University of Medical Sciences and Health Services

Alireza Haddadi

Shahid Sadoughi University of Medical Sciences and Health Services

AmirHossein Abotorabi Zarchi

Shahid Sadoughi University of Medical Sciences and Health Services

Shahab Sheikhalishahi

s.shahab.sheikhalishahi@gmail.com

Shahid Sadoughi University of Medical Sciences and Health Services

Article

Keywords:

Posted Date: September 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10740130/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

Artificial intelligence in gastric endoscopy increasingly produces spatial outputs to localize abnormal mucosa and define lesion extent, yet their clinical validity is uncertain. We conducted a task-specific systematic review nested within a prospectively registered parent review and identified 39 study families evaluating localization, segmentation, boundary identification, or margin delineation in gastric premalignant lesions and early gastric cancer. Most studies assessed segmentation on stored images. Histopathology often confirmed diagnosis, but spatial references were usually expert-defined endoscopic masks rather than pathology-registered boundaries. Only three studies reported patient-disjoint internal testing, six external validation, six a prospective component, and two live spatial evaluations; one study showed procedure-altering biopsy guidance. Dice, IoU, boundary, and margin measures were heterogeneous and were not pooled. Spatial AI has moved toward clinically relevant mapping and margin assessment, but stronger reference standards and validation during endoscopy are needed before these outputs can support procedural decisions.

Introduction

Gastric cancer remains a major global cancer burden, with approximately 980,000 new cases and 642,000 deaths estimated worldwide in 2024 (1). For appropriately selected early gastric cancers, endoscopic resection offers curative treatment, while gastric atrophy and intestinal metaplasia define premalignant states that guide surveillance and biopsy strategies (2–4). Clinical management therefore depends on more than recognizing that a lesion is present. Before endoscopic resection, the location and lateral extent of early gastric cancer must be defined sufficiently to guide lesion marking and complete resection; in premalignant mucosa, targeted sampling similarly depends on identifying abnormal areas (2–4). High-definition white-light endoscopy, chromoendoscopy, and magnifying image-enhanced techniques improve mucosal assessment, but the horizontal margin remains difficult to delineate in a subset of lesions, including some undifferentiated cancers (5, 6). Spatial assessment is therefore a distinct clinical problem from binary lesion detection.

Artificial intelligence (AI) research in gastric endoscopy has been dominated by lesion detection and diagnostic characterization, and recent meta-analyses have accordingly focused on sensitivity, specificity, and overall diagnostic accuracy for early gastric cancer and gastric intestinal metaplasia (7–9). Spatial applications have developed alongside this work, including semantic segmentation of intestinal metaplasia, pixel-level segmentation of early gastric cancer, and direct delineation of lesion boundaries or resection margins (10–14). These tasks are not interchangeable. Dice and intersection-over-union quantify overlap between a predicted region and a reference annotation, but neither alone establishes that a boundary is sufficiently reliable for biopsy targeting or endoscopic resection; their interpretation depends on the target and reference annotation, unit and independence of evaluation, imaging modality, and validation setting (15).

Spatial AI studies are distributed across several gastric targets and endoscopic modalities, and use different reference annotations, units of analysis, and performance measures (10, 15). These differences complicate direct comparison of reported spatial performance. Recent systematic reviews have remained focused mainly on diagnostic detection or characterization, with segmentation and margin-related outcomes incorporated into broader AI evidence syntheses rather than evaluated as a separate body of evidence (7–10, 16). The extent to which spatial models have progressed from curated still-image development to independent, external, video-based, or clinical evaluation is also not well characterized. Accordingly, we systematically examined AI systems developed for endoscopic localization, segmentation, and boundary or margin delineation of gastric premalignant lesions and early gastric cancer, with particular attention to spatial reference standards, performance measures, validation design, and clinical applicability.

Results

Study selection and study characteristics

Among the 217 study families included in the parent review, 39 met the task-specific eligibility criteria for the present spatial analysis after full-text reassessment (Fig. 1). Twenty-five study families addressed early gastric cancer or early gastric neoplasia, 11 gastric intestinal metaplasia, two gastric atrophy or chronic atrophic gastritis, and one multiple stages of gastric carcinogenesis.

The included evidence was published between 2017 and 2026, with 33 study families represented by reports published from 2021 onward. Studies were geographically concentrated in East and Southeast Asia: 18 study families were from China, nine from Japan, and six from Thailand; the remainder were from South Korea, Taiwan, India, Russia, or a collaborative Japan–Taiwan cohort. Most were retrospective model-development or validation studies based on still endoscopic images. Six incorporated a prospective component, and six included external validation. Formal evaluation remained predominantly image based, with only a small subset extending assessment to video or live endoscopy. White-light endoscopy and narrow-band imaging, including magnifying NBI, were the principal imaging modalities; linked-color or blue-light imaging, chromoendoscopy, and photodynamic imaging were used less often. Segmentation was the predominant spatial task, while demarcation-line or boundary identification, margin delineation, and region-level localization accounted for smaller parts of the evidence base (Table 1).

Table 1
Characteristics of study families included in the spatial AI review

Study	Country / region	Target	Endoscopic modality	Spatial task	Study size as reported	Validation setting
Ishida 2017 (34)	Japan	EGC/early gastric neoplasia	M-NBI	Segmentation + demarcation	50 images	internal/technical
Kanesaka, 2018 (35)	Japan/Taiwan	EGC/early gastric neoplasia	M-NBI	CADx + delineation	127 patients; 207 images	internal/technical
Takemoto, S., 2019 (36)	Japan	EGC/early gastric neoplasia	WLI	Patch classification + localization	1,000 images	internal/technical
Yoon HJ 2019 (37)	South Korea	EGC/early gastric neoplasia	WLI	Lesion localization	800 patients; 11,539 images	patient-disjoint
An P 2020 (13)	China	EGC/early gastric neoplasia	WLI/chromoendoscopy	Resection-margin delineation	536 patients; 1,244 images	internal/technical
Du, W., 2021 (38)	China	EGC/early gastric neoplasia	WLI	Segmentation	4,520 images	patient-disjoint
Ling T 2021 (14)	China	EGC/early gastric neoplasia	M-NBI	Differentiation + margin delineation	561 patients; 6,639 images	external, multicenter
Teramoto/Fujita family, 2021 (39)	Japan	EGC/early gastric neoplasia	WLI	Segmentation + CADe/classification	1,741 images	internal/technical
Zhang L 2021 (17)	China	EGC/early gastric neoplasia	WLI	Classification + pixel segmentation	22,308 images	internal/technical
Cui, 2022 (40)	China	EGC/early gastric neoplasia	Gastroscopy	CADe + segmentation	NR	internal/technical
Jin J 2022 (41)	China	EGC/early gastric neoplasia	White-light and narrow-band imaging	Early gastric cancer detection and segmentation	7,133 images; 10 videos	prospective
Liu L 2023 (18)	China	EGC/early gastric neoplasia	M-NBI	Neoplasia segmentation	1,196 patients; 7,777 images	external, multicenter
Okumura S 2022 (42)	Japan	EGC/early gastric neoplasia	M-NBI	Demarcation-line identification	11 patients; 23 images	internal/technical
Qiu, 2022 (43)	China	EGC/early gastric neoplasia	ME-NBI	Recognition + segmentation	NR	internal/technical
Teramoto A 2022 (44)	Japan	EGC/early gastric neoplasia	Endoscopy	Detection + region segmentation	187 patients; 2,378 images	internal/technical
Yamashita T 2022 (45)	Japan	EGC/early gastric	5-aminolevulinic-acid photodynamic	Detection + lesion-area segmentation	20 patients	prospective

Abbreviations: AI, artificial intelligence; BLI, blue-light imaging; CADe, computer-aided detection; CADx, computer-aided diagnosis; CAG, chronic atrophic gastritis; EGC, early gastric cancer; GIM, gastric intestinal metaplasia; HGIN, high-grade intraepithelial neoplasia; LCI, linked-color imaging; ME-NBI, magnifying endoscopy with narrow-band imaging; M-NBI, magnifying narrow-band imaging; NBI, narrow-band imaging; NR, not reported; OLGA, Operative Link on Gastritis Assessment; WLI, white-light imaging.

Study	Country / region	Target	Endoscopic modality	Spatial task	Study size as reported	Validation setting
		neoplasia	endoscopy			
Du W 2023 (12)	China	EGC/early gastric neoplasia	WLI	Segmentation	664 patients; 4,979 images	external, multicenter
Ma, L., 2023 (46)	China	EGC/early gastric neoplasia	C-WLI, NBI, ME-NBI and chromoendoscopy	Classification + localization/segmentation	387 patients; 4,697 images	internal/technical
Shibata 2023 (47)	Japan	EGC/early gastric neoplasia	Conventional endoscopy	Lesion segmentation + classification	187 patients; 2,377 images	internal/technical
Takemoto S 2023 (48)	Japan	EGC/early gastric neoplasia	WLI	Detection + demarcation	205 patients; 1,158 images	internal/technical
Zhang K 2024 (49)	China	EGC/early gastric neoplasia	Gastrosocopy	Detection + lesion segmentation	1,120 images	internal/technical
Hong K 2025 (19)	China	EGC/early gastric neoplasia	NBI and near-focus NBI	Boundary segmentation	2,861 images	internal/technical
Wang, B., 2025 (50)	China	EGC/early gastric neoplasia	WLI	Multi-task classification and segmentation	565 patients; 4,156 images	external, multicenter, patient-disjoint
Huang, C.-T., 2026 (51)	Japan	EGC/early gastric neoplasia	NBI	Segmentation	126 images	internal/technical
Menaga, K., 2026 (52)	India	EGC/early gastric neoplasia	WLI not otherwise specified	Segmentation	312 patients; 2,450 images	internal/technical
Wang, 2019 (53)	China	GIM	Endoscopy	Localization + segmentation	150 patients; 200 images	internal/technical
Siripoppohn, 2021 (54)	Thailand	GIM	Endoscopy	Segmentation	NR	internal/technical
Siripoppohn V 2022 (11)	Thailand	GIM	White-light and narrow-band imaging	Detection + segmentation	802 images	internal/technical
Tiankanon 2022 (55)	Thailand	GIM	WLI and NBI	Detection + semantic segmentation	136 patients; 962 images	internal/technical
Pornvoraphat P 2023 (56)	Thailand	GIM	White-light and narrow-band imaging	Detection + segmentation	2,179 images	internal/technical
Chen, Y.-T., 2024 (57)	Taiwan	GIM	NBI	Segmentation	454 images	internal/technical
Pornvoraphat, P. 2024 (58)	Thailand	GIM	White-light and narrow-band imaging	Gastric intestinal metaplasia segmentation	2,772 images	internal/technical
Tiankanon 2024 (59)	Thailand	GIM	WLI and NBI	Detection + segmentation assistance	75 patients	prospective, live
Abbreviations: AI, artificial intelligence; BLI, blue-light imaging; CADe, computer-aided detection; CADx, computer-aided diagnosis; CAG, chronic atrophic gastritis; EGC, early gastric cancer; GIM, gastric intestinal metaplasia; HGIN, high-grade intraepithelial neoplasia; LCI, linked-color imaging; ME-NBI, magnifying endoscopy with narrow-band imaging; M-NBI, magnifying narrow-band imaging; NBI, narrow-band imaging; NR, not reported; OLGA, Operative Link on Gastritis Assessment; WLI, white-light imaging.						

Study	Country / region	Target	Endoscopic modality	Spatial task	Study size as reported	Validation setting
Obukhova 2025 (60)	Russia	GIM	Endoscopic still images, modality NR	Segmentation	2,000 images	internal/technical
Yu 2026 (61)	China	GIM	NBI/LBI/LCI	Few-shot segmentation	889 images	internal/technical
Zhang, M., 2026 (62)	China	GIM	LCI	Segmentation	552 patients; 1,884 images	prospective
Zhao Q 2023 (21)	China	Gastric atrophy/CAG	Conventional gastroscopy video	Atrophy mapping + CAG/OLGA assessment	1,306 patients; 5,290 images	external, multicenter, prospective, live
Chen H 2025 (20)	China	Gastric atrophy/CAG	Conventional gastroscopy images	Atrophic-region segmentation	3,050 images	internal/technical
Gong EJ 2023 (22)	South Korea	Multiple stages	White-light gastroscopy	Multistage segmentation/classification	522 patients; 13,672 images; 522 videos	external, multicenter, prospective
Abbreviations: AI, artificial intelligence; BLI, blue-light imaging; CADE, computer-aided detection; CADx, computer-aided diagnosis; CAG, chronic atrophic gastritis; EGC, early gastric cancer; GIM, gastric intestinal metaplasia; HGIN, high-grade intraepithelial neoplasia; LCI, linked-color imaging; ME-NBI, magnifying endoscopy with narrow-band imaging; M-NBI, magnifying narrow-band imaging; NBI, narrow-band imaging; NR, not reported; OLGA, Operative Link on Gastritis Assessment; WLI, white-light imaging.						

Spatial tasks and performance

Across the 39 study families, semantic or region-level segmentation was the most frequent spatial task, often incorporated into systems that also performed lesion detection or classification. The distribution of study families across gastric targets and spatial tasks is shown in Fig. 2. Other studies evaluated lesion localization, demarcation-line or boundary identification, or delineation of the extent or resection margin of a lesion. Histopathology was frequently used to establish or confirm the lesion diagnosis, whereas the spatial reference itself was usually an endoscopist-defined mask, contour, or region. Annotation procedures ranged from a single experienced endoscopist to two-expert review, senior adjudication, or multi-expert consensus; several reports provided limited information on how disagreement between annotators was handled. Among the dedicated EGC margin studies, An et al. used direct point-to-point endoscopy–pathology mapping, whereas Ling et al. used postresection histopathology with expert-defined endoscopic margins (13, 14). A smaller group used coarser spatial references, including rectangular lesion regions, superpixel-based regions, or block-level localization, rather than pixel-precise lesion contours. These differences in reference construction were retained when spatial-performance results were summarized (Table 2). Study-level details of spatial reference construction, evaluation format, and reported spatial performance are provided in **Supplementary Table 2**.

Table 2
Spatial tasks, reference annotation, and representative performance

Target and spatial task	Study families, n	Modalities	Reference annotation	Representative reported spatial performance	Validation context
EGC/early neoplasia: lesion segmentation or extent mapping	18	WLI, NBI/ME-NBI, conventional gastroscopy	Predominantly expert-drawn pixel/region masks; histopathology commonly established lesion diagnosis. Some studies used coarse rectangular or region-level labels rather than pathology-registered contours.	Reported examples included mIoU 0.698 (Zhang 2021 (17)); DSC 0.911 and JSI 0.845 on an unseen dataset (Du 2023 (12)); Dice 0.867 and mIoU 0.584 (Liu 2023 (18)); external Dice 0.898–0.907 and IoU 0.818–0.834 (Wang 2025 (50)); and DSC 0.923 in a patient-disjoint internal test (Du 2021 (38)).	Mostly stored still images; limited patient-disjoint or external testing.
EGC/early neoplasia: lesion localization or region prediction	3	Mainly WLI/conventional endoscopy	Expert-defined lesion regions, blocks, or masks; spatial granularity varied across studies.	Takemoto 2019 (36) correctly identified 73.3% of cancer blocks and 98.8% of normal blocks in an unseen-image localization test. Other studies quantified lesion-region localization using overlap- or region-based criteria rather than a common metric.	Offline image-based evaluation.
EGC/early neoplasia: demarcation-line or boundary identification	5	M-NBI, NBI/nF-NBI, WLI	Endoscopist-drawn lesion boundaries or demarcation lines; pathology linkage varied.	Boundary-specific models reported Dice 0.831 with NBI and 0.853 with nF-NBI (Hong 2025 (19)). Other studies used lesion-area detection, boundary distance, or delineation agreement rather than Dice/IoU.	Predominantly internal/offline.
EGC: resection-margin delineation	2	WLI/chromoendoscopy; ME-NBI	Expert-defined endoscopic margin annotations with postresection histopathology; An 2020 (13) used direct point-to-point endoscopy–pathology mapping, whereas Ling 2021 (14) used pathology-confirmed EGC with expert-defined endoscopic margins.	An 2020 (13) reported 85.7% accuracy with chromoendoscopy and 88.9% with WLI at overlap ≥ 0.60 . Ling 2021 (14) reported margin-delineation accuracy of 82.7% for differentiated and 88.1% for undifferentiated EGC at overlap ≥ 0.80 .	Included video-based assessment; not equivalent to routine live deployment.
GIM: segmentation/localization	11	WLI, NBI, LCI, BLI	Expert masks or region annotations; some were histology-linked, whereas others relied on expert endoscopic labels without direct biopsy registration.	Internally evaluated models reported mIoU values of approximately 0.55–0.733. One ROI-based study reported IoU 0.88. A BiSeNet system reported mIoU 0.570 and 31.53 frames/s.	Predominantly internal image datasets; limited prospective/live assessment.
Gastric atrophy/CAG: spatial mapping	2	WLI/conventional gastroscopy	Expert atrophic-region annotations with	Chen 2025 (20) reported mean Dice	One retrospective

Spatial-task categories were not mutually exclusive; a study family could contribute to more than one spatial task. Reported performance measures were retained as defined by the original studies and were not pooled or converted to a common scale. Abbreviations: BLI, blue-light imaging; CAG, chronic atrophic gastritis; CDSS, clinical decision support system; DSC, Dice similarity coefficient; EGC, early gastric cancer; GIM, gastric intestinal metaplasia; IoU, intersection over union; JSI, Jaccard similarity index; LCI, linked-color imaging; ME-NBI, magnifying endoscopy with narrow-band imaging; M-NBI, magnifying narrow-band imaging; NBI, narrow-band imaging; nF-NBI, near-focus narrow-band imaging; ROI, region of interest; WLI, white-light imaging.

Target and spatial task	Study families, n	Modalities	Reference annotation	Representative reported spatial performance	Validation context
			histopathology used for disease confirmation/staging where applicable.	0.94 and mean IoU 0.90 for atrophic-region segmentation. Zhao 2023 (21) used real-time AI marking of suspected atrophic areas within histopathology-linked CAG/OLGA assessment.	image-based study and one prospective real-time clinical study.
Multiple stages of gastric carcinogenesis	1	Routine endoscopy	Stage-specific expert/reference labels with segmentation outputs for atrophy and intestinal metaplasia.	Gong 2023 (22) reported an internal lesion-segmentation rate of 93.4% and subsequently evaluated the CDSS in prospective screening endoscopy.	External plus prospective clinical evaluation.
Spatial-task categories were not mutually exclusive; a study family could contribute to more than one spatial task. Reported performance measures were retained as defined by the original studies and were not pooled or converted to a common scale. Abbreviations: BLI, blue-light imaging; CAG, chronic atrophic gastritis; CDSS, clinical decision support system; DSC, Dice similarity coefficient; EGC, early gastric cancer; GIM, gastric intestinal metaplasia; IoU, intersection over union; JSI, Jaccard similarity index; LCI, linked-color imaging; ME-NBI, magnifying endoscopy with narrow-band imaging; M-NBI, magnifying narrow-band imaging; NBI, narrow-band imaging; nF-NBI, near-focus narrow-band imaging; ROI, region of interest; WLI, white-light imaging.					

Among the 25 study families addressing EGC or early gastric neoplasia, spatial evaluation included lesion segmentation, region localization, boundary recognition, and margin delineation across white-light endoscopy, NBI or magnifying NBI, chromoendoscopy, and less frequently other optical modalities. Performance was reported using several non-equivalent spatial measures. DeepLabv3 segmentation in a WLI study yielded a mean IoU of 0.698 on 1,091 test images (17), whereas a two-center ResUNet study reported a DSC of 0.911 and Jaccard similarity index of 0.845 on an unseen dataset (12). In a multicenter ME-NBI study, neoplastic-lesion segmentation reached a Dice coefficient of 0.867 and mean IoU of 0.584 (18), while boundary-specific models achieved Dice coefficients of 0.831 with NBI and 0.853 with near-focus NBI (19). Margin-focused studies used clinically oriented spatial endpoints in addition to conventional overlap metrics. ENDOANGEL achieved accuracies of 85.7% with chromoendoscopy and 88.9% with WLI at an overlap threshold of 0.60 and was further evaluated using point-to-point endoscopy–pathology mapping (13); a separate ME-NBI system achieved margin-delineation accuracies of 82.7% for differentiated and 88.1% for undifferentiated EGC at an overlap threshold of 0.80 (14). External or independent spatial testing was reported in only a subset of EGC studies, while most evaluations remained based on stored endoscopic images (Table 2).

Gastric intestinal metaplasia accounted for most of the spatial evidence on premalignant conditions. Segmentation was evaluated mainly with WLI and NBI, with additional studies using LCI or other forms of image-enhanced endoscopy. In a histologically confirmed GIM dataset, a real-time BiSeNet model achieved a mean IoU of 0.57 and an inference speed of 31.53 frames per second (11). Other internally evaluated GIM segmentation models reported mean IoU values from 0.55 to 0.733, whereas one ROI-based study reported an IoU of 0.88 using expert annotation without reported histological confirmation of the spatial reference. Reference construction varied from histology-linked masks reviewed by two endoscopists to expert consensus without direct biopsy linkage. The evidence for gastric atrophy was smaller but included both image-based segmentation and clinical evaluation: a vision transformer-based study reported overlap of the model and senior-endoscopist atrophic-region segmentations above 0.90 (20), while a prospective U-Net study automatically marked atrophic areas during real-time endoscopy and evaluated the system against histopathology as part of CAG diagnosis and OLGA staging (21). The study encompassing multiple stages of gastric carcinogenesis combined segmentation of atrophy and intestinal metaplasia with external testing and prospective screening endoscopy, reporting an internal lesion-segmentation rate of 93.4% (22). Most remaining premalignant-condition segmentation studies were evaluated on internal image datasets.

Validation and clinical applicability

External validation was reported in six of the 39 study families, six included a prospective component, and six involved multicenter validation. Explicit patient-disjoint partitioning of development and internal test data was documented in three studies. In many of the remaining image- and pixel-level evaluations, multiple observations were derived from individual patients and patient-level separation was either not reported or unclear. Prospective work ranged from prospectively acquired image datasets and video feasibility testing to evaluation during clinical endoscopy. The CAG U-Net system underwent prospective real-time video evaluation (21), while the multistage gastric-carcinogenesis CDSS combined internal atrophy and intestinal-metaplasia segmentation with external testing and prospective assessment in 522 consecutive screening examinations (22). Video-based spatial testing was also incorporated into EGC margin-delineation studies (13, 14). Reference annotations ranged from expert-drawn masks linked to histopathology to endoscopic annotations without direct pathological boundary registration; several reports did not fully

describe annotator agreement or consensus procedures. Outcomes derived from multiple images, pixels, regions, or biopsy sites within the same patient were present in several datasets. Among the 13 study families with a conventional diagnostic-accuracy component, all were judged at high overall risk of bias using QUADAS-2. High-risk judgments were most frequent for patient selection (11/13), index test (10/13), and flow and timing (11/13), whereas one study was judged at high risk in the reference-standard domain (**Supplementary Table 1**; Table 3).

Table 3
Validation design and clinical applicability of the spatial AI evidence base

Validation or implementation feature	Frequency	Operational definition	Representative studies or examples
Explicit patient-disjoint internal development/test separation	3/39 (7.7%)	Development and internal test data were explicitly separated at patient level.	Yoon 2019 (37); Wang 2025 (50); Du 2021 (38)
Independent external validation	6/39 (15.4%)	Testing on an independent external center, cohort, or geographic dataset.	Included EGC, GIM/atrophy, and multistage systems
Multicenter validation	6/39 (15.4%)	Evaluation incorporated data from more than one center; not all multicenter datasets were institution-held-out external tests.	Ling 2021 (14); Wang 2025 (50); Zhao 2023 (21); Gong 2023 (22)
Prospective component	6/39 (15.4%)	Prospectively collected data, prospective video evaluation, or prospective clinical assessment.	Zhao 2023 (21); Gong 2023 (22); selected EGC/GIM systems
Live spatial evaluation	2/39 (5.1%)	Spatial AI output was evaluated during live endoscopy rather than only on stored images/video.	Zhao 2023 (21); Tiankanon 2024 (59)
Formal endoscopist/reader comparison	9/39 (23.1%)	Standalone AI or AI-assisted performance was compared with one or more endoscopists/readers.	Across EGC, GIM, and CAG studies
Spatial output directly altered the procedure	1/39 (2.6%)	AI-marked spatial findings triggered an additional procedure-level action.	Zhao 2023 (21): additional biopsy of AI-marked atrophic areas
Margin/resection-planning evaluation	2/39 (5.1%)	Spatial output was assessed specifically for EGC margin definition relevant to resection planning.	An 2020 (13); Ling 2021 (14)
Open source code	2/39 (5.1%)	A code repository was reported as openly available.	Zhang 2026 (62); Chen 2024 (57)
Openly available trained model	0/39 (0%)	Publicly accessible trained weights/model were identified.	None identified
Failure-case analysis	2 formal + 1 limited	Explicit review of failure/error cases beyond aggregate performance metrics.	Two families formal; one additional family limited
QUADAS-2 applicable DTA subset	13 families	QUADAS-2 was applied only to study families contributing a conventional diagnostic-accuracy component.	13/13 high overall RoB; D1 high 11/13, D2 10/13, D3 1/13, D4 11/13
Percentages use the 39 included study families as the denominator unless otherwise stated. QUADAS-2 was applied only to the 13 study families with a conventional diagnostic-accuracy component; it was not used to grade segmentation, localization, or boundary-delineation outcomes. The live-evaluation and procedure-modification categories refer specifically to spatial AI output rather than diagnostic classification alone. Abbreviations: CAG, chronic atrophic gastritis; DTA, diagnostic test accuracy; EGC, early gastric cancer; GIM, gastric intestinal metaplasia; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies 2; RoB, risk of bias.			

Only a minority of studies extended spatial evaluation beyond stored still images. Nine study families included a formal endoscopist comparison or reader evaluation. Direct use of a spatial AI output to modify the procedure was documented in the prospective CAG study: suspected atrophic areas marked by the U-Net system but not initially identified by the endoscopist prompted additional biopsies during gastroscopy (21). EGC margin delineation represented the clearest application to treatment planning. ENDOANGEL was evaluated on ESD videos and through point-to-point correspondence between endoscopic and pathological margins (13), whereas the ME-NBI system of Ling et al. was tested for real-time margin delineation on unprocessed EGC videos (14). The multistage CDSS was prospectively applied during screening endoscopy, although its reported atrophy and intestinal-metaplasia segmentation result was obtained from internal testing (22). Inference speed or hardware specifications were reported in several technical studies, generally in research-workstation or offline settings. Open-source code was identified for two study families, but no openly available trained model was identified. Formal failure-case analysis was reported in two study families, with limited error analysis in one additional family. Across the 39 families, most studies remained at development or internal-validation stages; external, prospective, and procedure-integrated evaluations formed smaller subsets (Table 3). The study-level profile of validation design, clinical translation, reproducibility, and error-analysis features is shown in **Supplementary Fig. 1**, with corresponding implementation details provided in **Supplementary Table 3**.

Discussion

This review focuses on a part of gastric endoscopic AI that is not captured by diagnostic accuracy alone: defining where abnormal mucosa lies and how far it extends. Spatial models have moved beyond image segmentation in selected datasets, with some systems addressing demarcation lines, resection margins, and outputs linked to procedural decisions. Good spatial performance and clinical usefulness, however, are different questions. Interpretation depends not only on agreement with a reference annotation, but on whether that reference is clinically meaningful, testing is independent, and the output remains reliable in the workflow in which it is intended to be used.

Earlier reviews of gastric endoscopic AI addressed a different question from the one examined here. Meta-analyses of EGC have primarily pooled sensitivity and specificity for lesion detection or diagnosis and compared AI with endoscopists, while the corresponding GIM literature has similarly focused on diagnostic accuracy (7–9). These analyses are informative about whether AI recognizes disease, but they do not establish whether the pixels identified by a model define the spatial extent needed for a clinical decision. A system can classify an image correctly, or even achieve substantial overlap with an annotated lesion, while still placing the relevant boundary incorrectly. This distinction is particularly important in EGC, where the horizontal extent of a lesion determines treatment planning and may remain difficult to define even after the lesion itself has been recognized (5, 6). Dedicated EGC margin studies addressed this problem more directly by evaluating spatial output against a treatment-relevant lesion boundary rather than treating segmentation as an end in itself (13, 14). The spatial question in GIM is different again. Here, the potential value of segmentation is not to define a resection margin, but to characterize the distribution of metaplastic mucosa and potentially support the areas selected for assessment or biopsy; current European guidance already links endoscopic characterization of gastric precancerous conditions with biopsy-based staging (3). A segmentation mask is therefore clinically meaningful only in relation to the decision it is expected to support. Framing the literature in this way shifts the question from whether AI can identify abnormal mucosa to whether its spatial output represents the disease extent with sufficient validity for the intended endoscopic task.

For spatial models, the reference annotation determines what a high Dice or IoU actually means. In many studies in this review, histopathology confirmed the lesion diagnosis, but the mask used for training or evaluation was still defined on the endoscopic image by one or more endoscopists. These are not equivalent forms of validation: histology can establish what the lesion is without establishing its exact endoscopic boundary. Expert-drawn masks are often the only practical spatial reference, and consensus or adjudication reduces dependence on a single observer, but it does not create an independent ground truth. Inter-annotator studies in medical image segmentation have shown that expert contours vary and that the construction of the reference set can affect the benchmark against which an algorithm is judged (23, 24). The EGC margin studies illustrate why the construction of the reference matters. An et al. used point-to-point endoscopy–pathology mapping, allowing the predicted resection margin to be related directly to the histologic cancer boundary (13). Ling et al. studied pathology-confirmed EGC but evaluated spatial agreement against endoscopically defined margins and subsequently tested the system on unprocessed ME-NBI videos (14). The two approaches therefore validate different aspects of performance: confirmation of the disease does not, by itself, provide histologic validation of every point along an endoscopic boundary.

The same caution applies to the performance metrics themselves. Dice and IoU are useful measures of spatial overlap, but their values depend on the size and geometry of the target structure and on how observations are aggregated; neither indicates whether disagreement occurs at a clinically consequential part of the boundary (15). Two predictions with similar overlap may therefore have different implications if one preserves the lesion margin while the other displaces the lateral edge relevant to biopsy or resection planning. A recent systematic review of intraoperative computer vision in robotic surgery identified a similar reliance on overlap-based segmentation metrics and limited clinical validation, indicating that the gap between technical performance and procedural relevance extends beyond gastric endoscopy (25). The validation design adds a separate source of uncertainty. When multiple images or frames from the same patient are treated as independent observations, or when development and test sets are separated at the image rather than patient level, correlated information can occur across partitions and inflate apparent test performance; this effect has been demonstrated experimentally in medical imaging datasets (26, 27). In the present review, explicit patient-disjoint internal testing and external validation were uncommon. High performance on an internal image-level test set therefore did not necessarily establish that the same spatial accuracy would be retained in new patients, centers, devices, or endoscopic conditions. Differences in target geometry, aggregation, unit of analysis, and independence of testing also limit direct comparison of Dice or IoU values across studies. The decision not to pool spatial-performance estimates consequently reflects more than statistical heterogeneity: many of the reported values were generated under different reference structures and validation conditions and cannot be assumed to measure the same quantity.

Figure 3 shows that validation has generally not progressed at the same pace as model development. Much of the spatial evidence was generated from stored images, while independent external testing, prospective evaluation, and assessment during live endoscopy were uncommon. The distinction between these stages is clinically important. A model that can process endoscopic frames at video speed has demonstrated computational feasibility, but this does not establish that its spatial output remains accurate when image quality, viewing angle, insufflation, motion, mucus, bleeding, and operator behavior vary during an actual procedure. Prerecorded video provides a more realistic test than selected still images because the model must process sequential and changing views, but it remains an offline evaluation in which the output cannot influence what the endoscopist does. Reader-comparison studies answer a different question again: they show how AI performs relative to endoscopists, or how endoscopists perform with AI assistance when an assisted-reader design is used, but do not necessarily show that the spatial information changes biopsy or treatment decisions. The strongest evidence of clinical translation therefore comes from studies in which the system is evaluated prospectively during endoscopy and its output is allowed to enter the decision process. This distinction between

technical real-time capability, live clinical evaluation, and procedure-altering use is consistent with DECIDE-AI, which defines early clinical evaluation around AI systems used in live settings where supported decisions can affect patient care (28). The pattern also provides a useful context for the 2026 World Endoscopy Organization position statement, which judged the evidence supporting clinical use of AI for gastric epithelial neoplasia to remain insufficient and concentrated in advanced institutions in a limited number of countries (29). The apparently greater representation of some validation domains in CAG and multiple-stage studies should also be interpreted cautiously, because these groups contained only two and one study families, respectively; these proportions therefore describe the distribution of the available evidence rather than comparative maturity between gastric targets.

The studies that came closest to this level of use are more informative about clinical translation than offline overlap scores alone. For EGC, An et al. brought spatial evaluation closer to treatment planning by assessing a pathology-linked margin in ESD video, although the system was not evaluated as a live interventional tool (13). Zhao et al. went further in CAG: during real-time gastroscopy, the model marked suspected atrophic sites, the endoscopist was informed of AI-identified sites that had not already been recognized, and additional biopsies were obtained from those areas (21). The importance of this design lies not simply in real-time deployment, but in showing that the spatial output entered the clinical decision process. These examples also clarify how clinical value should be judged according to intended use. For a margin-delineation system, the relevant question is whether the predicted boundary can support appropriate resection planning; for mucosal mapping or biopsy guidance, evaluation should address whether the spatial output improves where tissue is sampled, diagnostic yield, biopsy burden, or another prespecified procedural outcome. Dice, IoU, and inference speed remain useful technical measures, but they cannot substitute for evidence that the spatial output performs its intended function during endoscopy.

Failure reporting and reproducibility were also limited. Formal failure-case analysis was reported in only two study families, with limited analysis in one additional study, despite the importance of understanding whether spatial errors occur under recognizable endoscopic conditions or at clinically relevant lesion boundaries. Code was openly available for only two study families, and no study provided an openly available trained model. These findings do not imply that closed systems are intrinsically less accurate, but they make independent verification and comparison more difficult. For spatial AI, this matters because an aggregate Dice or IoU gives little information about the circumstances in which a model mislocalizes a lesion edge, fails under changing endoscopic views, or produces a region that would be misleading if acted upon. Failure analysis should therefore be considered part of clinical validation rather than an optional technical appendix, particularly for systems intended to influence biopsy or resection decisions. Transparent reporting of software, model availability, and clinically contextualized errors would also make it easier to determine whether reported performance can be reproduced outside the development environment (30).

Future studies should be designed around the clinical use of the spatial output rather than incremental gains on curated image sets. Development and internal testing should be separated at the patient level, with subsequent validation in independent cohorts that include different centers, endoscopy systems, and operators; repeated images or frames from the same patient should either remain within the same partition or be accounted for analytically (26, 27, 30). Reference annotation also requires greater attention. For margin-oriented tasks, endoscopy–pathology registration should be incorporated where feasible, whereas other spatial tasks should use prespecified multi-expert annotation with a defined adjudication process and reporting of observer variation (30). Evaluation should then move from selected images to continuous, unselected procedure video and, ultimately, prospective use during endoscopy. Spatial overlap measures can remain part of this evaluation, but boundary- or distance-sensitive measures should accompany Dice or IoU when the location of the lesion edge is clinically relevant (15). Clinical studies should assess the incremental value of AI when its output is available to the endoscopist, using endpoints matched to the intended task (for example, biopsy targeting, missed lesion extent, adequacy of resection margins, unnecessary biopsy burden, procedure time, or completeness of resection) rather than standalone model accuracy alone. DECIDE-AI similarly emphasizes early clinical evaluation of actual system performance and the human factors surrounding its use before larger-scale implementation (28).

Several features of this review strengthen the interpretation of a heterogeneous evidence base. The analysis was nested within a prospectively registered parent review, companion reports were linked at the study-family level to avoid double counting, and spatial tasks were considered separately from conventional detection and classification. Reported spatial metrics were retained in their original form rather than converted or pooled across incompatible tasks, reference standards, and units of analysis. The review also examined how spatial references were constructed and distinguished technical performance from external validation, live evaluation, and procedural use. The review also has limitations. This was a task-specific secondary analysis of the parent review rather than a new search designed exclusively around segmentation or boundary-delineation terminology, although the parent search covered major biomedical and engineering databases. Reporting was incomplete in some studies, particularly for patient-level data separation, annotation procedures, and the relation between endoscopic annotations and pathology; conference abstracts and technical proceedings provided less methodological detail in some cases. Variation in target definition, reference annotation, unit of analysis, and spatial metrics precluded a meaningful quantitative synthesis. Finally, the validation and clinical-translation domains were used descriptively to characterize the evidence and were not intended as a validated readiness score.

Spatial AI in gastric endoscopy has progressed from lesion-level recognition toward models that estimate mucosal extent, boundaries, and, in a small number of studies, spatial information relevant to biopsy or treatment planning. Further progress will require evidence that these outputs

correspond to clinically valid disease boundaries and remain reliable in independent cohorts and during actual endoscopic procedures. Clinical validation should therefore be aligned with the action that the predicted region or boundary is intended to guide.

Methods

Study design and study selection

This review was undertaken as a task-specific secondary analysis of the studies identified in a prospectively registered systematic review of artificial intelligence in upper gastrointestinal endoscopy for gastric premalignant lesions and early gastric cancer. Localization, delineation, and segmentation were prespecified eligible AI tasks in the parent protocol. The parent protocol is accessible through PROSPERO under registration CRD420261476860; no amendments were made to the prespecified eligibility of localization, delineation, or segmentation for this task-specific analysis. The present analysis focused on studies that evaluated the spatial location or extent of eligible gastric lesions. Reports arising from the same underlying study or model-development cohort were linked at the study-family level and were not counted as independent studies. The review is reported in accordance with the PRISMA 2020 statement (31).

The parent review searched PubMed/MEDLINE, Embase.com, Scopus, Web of Science Core Collection, Cochrane Library/CENTRAL, and IEEE Xplore from database inception to 6 July 2026. ClinicalTrials.gov was searched on the same date and the WHO International Clinical Trials Registry Platform on 9 July 2026. The search combined controlled vocabulary and free-text terms for artificial intelligence and machine learning, upper gastrointestinal endoscopy, gastric location, and gastric premalignant lesions or early gastric cancer. No restrictions were imposed on publication year, language, publication type, age, or access status. The complete database-specific strategies, search dates, and retrieval counts were retained from the parent review and are provided in the **Supplementary Information**. Search methods and reporting were documented in accordance with PRISMA-S (32). No additional database search was performed for this task-specific analysis.

All study families in the parent review were reassessed against the eligibility criteria for the present review using the final extraction database and corresponding full texts. Eligible spatial outputs included pixel- or region-level segmentation, lesion-extent maps, demarcation-line or boundary identification, and margin delineation. Localization without a segmentation mask was retained when the spatial prediction was evaluated against a reference lesion annotation or directly informed a clinical action, such as targeted biopsy. Studies were excluded when spatial output was limited to coarse bounding-box detection, when activation or saliency maps were used only to interpret model predictions, when AI was restricted to image enhancement followed by human delineation, or when the spatial method relied solely on non-learning image processing. Companion publications could provide additional methodological or outcome information, but each underlying study family was counted once. Task-specific eligibility and spatial-data extraction were performed independently by two reviewers. Disagreements were resolved through discussion and consensus.

Data extraction and methodological assessment

Data collected in the parent review were retained for study design, population, endoscopic modality, AI architecture, validation setting, and report linkage. Full texts were then reviewed for variables specific to spatial evaluation, including target condition, spatial task, input format, unit of analysis, reference annotation, data-splitting strategy, and internal, external, prospective, video-based, or live evaluation. Spatial-performance measures were extracted as reported, including Dice coefficient, intersection over union (IoU) or Jaccard index, pixel-level measures, region overlap, boundary or margin agreement, and distance-based metrics. Reader comparisons, inference speed, use of the spatial output for biopsy or resection planning, code or model availability, and failure analyses were recorded when reported. The selection of AI-specific extraction items was informed by CLAIM 2024 (30), which was used as a reporting framework rather than as a risk-of-bias tool or study-level compliance score.

Methodological assessment was matched to study design and purpose. For the 13 study families with a conventional diagnostic-accuracy component, QUADAS-2 assessments from the parent review were independently rechecked by two reviewers for the present analysis, covering patient selection, index test, reference standard, and flow and timing (33). Disagreements were resolved through discussion and consensus. QUADAS-2 was not used to grade segmentation, localization, or boundary-delineation outcomes. For all 39 study families, spatial methodological features were assessed descriptively, including independence of development and test data, patient-level separation, external or temporal validation, construction of the reference annotation, annotator number and expertise, consensus or adjudication procedures, histopathologic linkage when applicable, unit of analysis, and extension to video, prospective data, or live endoscopy. Each domain was reported separately, and no composite methodological-quality score was calculated (**Supplementary Table 1**).

Spatial-performance outcomes were retained in the form reported by the original studies. Dice and IoU/Jaccard measures were considered overlap metrics, whereas pixel-level classification measures, boundary agreement, and distance-based measures were treated as distinct outcome types. Metrics were not converted into a common scale because their interpretation depends on the spatial task, target structure, evaluation unit, and method of aggregation; this approach is consistent with current recommendations for validation of biomedical image-analysis algorithms (15). Results derived at pixel, region, image, lesion, video, or patient level were therefore kept separate. When several

architectures were compared within a study, the model identified by the investigators as their proposed or principal approach was used for descriptive reporting rather than selecting a model solely because it produced the highest result. Independent or external test results were given precedence over development-set performance when both were available.

Evidence synthesis

Studies were grouped first by target condition and spatial task, distinguishing segmentation, lesion-region localization, demarcation or boundary identification, and margin delineation. Within these groups, findings were examined in relation to endoscopic modality, reference-annotation method, unit of analysis, and validation setting. Quantitative pooling was not undertaken because the included studies differed in the anatomical and pathological target being delineated, definition of the reference boundary, evaluation unit, and spatial-performance metric. Results were instead presented in structured tables and a narrative synthesis, retaining the outcome definitions used by the original investigators. Studies based on selected still images were distinguished from those using patient-disjoint test sets, external validation, prospectively acquired data, continuous or prerecorded video, or live endoscopy. Findings from companion publications were integrated within the corresponding study family.

Clinical applicability was considered separately from spatial accuracy. We recorded whether an evaluation included independent or external validation, prospective or video-based testing, use during live endoscopy, comparison with endoscopists, or an intended clinical function such as biopsy targeting or margin assessment. Implementation-related information, including inference speed, deployment setting, code or model availability, and reported failure analysis, was summarized when available. These features were presented individually and were not combined into a composite readiness score.

Data availability

The study-level dataset generated and analyzed during this systematic review is available in the Zenodo repository at <https://doi.org/10.5281/zenodo.21991723>. The dataset contains the extracted variables used for the task-specific spatial analysis and study-family-level evidence synthesis reported in this article.

Declarations

Data availability

The study-level dataset generated and analyzed during this systematic review is available in the Zenodo repository at <https://doi.org/10.5281/zenodo.21991723>. The dataset contains the extracted variables used for the task-specific spatial analysis and study-family-level evidence synthesis reported in this article.

Acknowledgements

This study received no funding.

Author contributions

SS conceived the study. SS and AH developed the protocol and search strategy. SS, AH, and AAZ performed study screening and full-text assessment. SS and AH performed data extraction and methodological and risk-of-bias assessment. SS and AH conducted the data synthesis and descriptive analyses. SS, AH, and MKA contributed to methodological and clinical adjudication. SS prepared the tables and figures. SS drafted the manuscript. All authors contributed to the interpretation of the findings, critically reviewed and revised the manuscript for important intellectual content, and read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Sung H, Filho AM, Laversanne M, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA Cancer J Clin*. 2026;76(4):e70090.
2. Pimentel-Nunes P, Libânio D, Bastiaansen BAJ, Bhandari P, Bisschops R, Bourke MJ, et al. Endoscopic submucosal dissection for superficial gastrointestinal lesions: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2022. *Endoscopy*. 2022;54(6):591–622.
3. Dinis-Ribeiro M, Libânio D, Uchima H, Spaander MCW, Bornschein J, Matysiak-Budnik T, et al. Management of epithelial precancerous conditions and early neoplasia of the stomach (MAPS III): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter

- and Microbiota Study Group (EHMSG) and European Society of Pathology (ESP) Guideline update 2025. *Endoscopy*. 2025;57(5):504–54.
4. Japanese gastric cancer treatment guidelines 2025 (7th edition). *Gastric Cancer*. 2026;29(2):271–99.
 5. Nagahama T, Yao K, Maki S, Yasaka M, Takaki Y, Matsui T, et al. Usefulness of magnifying endoscopy with narrow-band imaging for determining the horizontal extent of early gastric cancer when there is an unclear margin by chromoendoscopy (with video). *Gastrointest Endosc*. 2011;74(6):1259–67.
 6. Kim JW, Jung Y, Jang JY, Kim GH, Bang BW, Park JC, et al. Narrowband imaging with near-focus magnification for discriminating the gastric tumor margin before endoscopic resection: A prospective randomized multicenter trial. *J Gastroenterol Hepatol*. 2020;35(11):1930–7.
 7. Shi Y, Fan H, Li L, Hou Y, Qian F, Zhuang M, et al. The value of machine learning approaches in the diagnosis of early gastric cancer: a systematic review and meta-analysis. *World J Surg Oncol*. 2024;22(1):40.
 8. Li N, Yang J, Li X, Shi Y, Wang K. Accuracy of artificial intelligence-assisted endoscopy in the diagnosis of gastric intestinal metaplasia: A systematic review and meta-analysis. *PLoS One*. 2024;19(5):e0303421.
 9. Liu J, Li D, Zhuo Y, Zhang S. Deep learning for detecting early gastric cancer with white-light endoscopy: a systematic review and meta-analysis. *Front Artif Intell*. 2026;9:1734591.
 10. Klang E, Soroush A, Nadkarni GN, Sharif K, Lahat A. Deep Learning and Gastric Cancer: Systematic Review of AI-Assisted Endoscopy. *Diagnostics (Basel)*. 2023;13(24).
 11. Siripoppohn V, Pittayanon R, Tiankanon K, Faknak N, Sanpavat A, Klaikaew N, et al. Real-time semantic segmentation of gastric intestinal metaplasia using a deep learning approach. *Clin Endosc*. 2022;55(3):390–400.
 12. Du W, Rao N, Yong J, Adjei PE, Hu X, Wang X, et al. Early gastric cancer segmentation in gastroscopic images using a co-spatial attention and channel attention based triple-branch ResUnet. *Comput Methods Programs Biomed*. 2023;231:107397.
 13. An P, Yang D, Wang J, Wu L, Zhou J, Zeng Z, et al. A deep learning method for delineating early gastric cancer resection margin under chromoendoscopy and white light endoscopy. *Gastric Cancer*. 2020;23(5):884–92.
 14. Ling T, Wu L, Fu Y, Xu Q, An P, Zhang J, et al. A deep learning-based system for identifying differentiation status and delineating the margins of early gastric cancer in magnifying narrow-band imaging endoscopy. *Endoscopy*. 2021;53(5):469–77.
 15. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods*. 2024;21(2):195–212.
 16. Lv M, Chen F, Li Q, Xue M, Wang J. Comparative diagnostic accuracy of different artificial intelligence models for early gastric cancer: a systematic review and meta-analysis. *Front Oncol*. 2025;15:1670843.
 17. Zhang L, Zhang Y, Wang L, Wang J, Liu Y. Diagnosis of gastric lesions through a deep convolutional neural network. *Dig Endosc*. 2021;33(5):788–96.
 18. Liu L, Dong Z, Cheng J, Bu X, Qiu K, Yang C, et al. Diagnosis and segmentation effect of the ME-NBI-based deep learning model on gastric neoplasms in patients with suspected superficial lesions - a multicenter study. *Front Oncol*. 2022;12:1075578.
 19. Hong K, Lei C, Kan X, Ouyang Y, Mei Y, Guo Y, et al. Artificial intelligence for early gastric cancer boundary recognition in NBI and nF-NBI endoscopic images. *Scand J Gastroenterol*. 2025;60(7):624–34.
 20. Chen H, Liu S, Miao Y, Yang X, Li T, Wu C, et al. Constructing and validating vision transformer-based assisted detection models for atrophic gastritis: A retrospective study. *Sci Prog*. 2025;108(3):368504251381972.
 21. Zhao Q, Jia Q, Chi T. U-Net deep learning model for endoscopic diagnosis of chronic atrophic gastritis and operative link for gastritis assessment staging: a prospective nested case-control study. *Ther Adv Gastroenterol*. 2023;16:17562848231208669.
 22. Gong EJ, Bang CS, Lee JJ, Jeong HM, Baik GH, Jeong JH, et al. Clinical Decision Support System for All Stages of Gastric Carcinogenesis in Real-Time Endoscopy: Model Establishment and Validation Study. *J Med Internet Res*. 2023;25:e50448.
 23. Yang F, Zamzmi G, Angara S, Rajaraman S, Aquilina A, Xue Z, et al. Assessing Inter-Annotator Agreement for Medical Image Segmentation. *IEEE Access*. 2023;11:21300–12.
 24. Molière S, Hamzaoui D, Granger B, Montagne S, Allera A, Ezziane M, et al. Reference standard for the evaluation of automatic segmentation algorithms: Quantification of inter observer variability of manual delineation of prostate contour on MRI. *Diagn Interv Imaging*. 2024;105(2):65–73.
 25. Korsten T, Kuiper GM, Su R, Ruurda JP, Verhagen A, Mariani M, et al. Artificial intelligence for intraoperative surgical guidance in robotic-assisted ventral cavity surgery: a systematic review on the current state of validation methods. *npj Digital Surgery*. 2026;1(1):13.
 26. Tampu IE, Eklund A, Haj-Hosseini N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Sci Data*. 2022;9(1):580.
 27. Yagis E, Atnafu SW, García Seco de Herrera A, Marzi C, Scheda R, Giannelli M, et al. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci Rep*. 2021;11(1):22544.

28. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924–33.
29. Fujishiro M, Kakushima N, Abe S, Yip HC, Liu X, Kim GH, et al. World Endoscopy Organization Position Statements for Artificial Intelligence in Endoscopic Diagnosis of Gastric Epithelial Neoplasia. *Dig Endosc*. 2026;38(4):e70144.
30. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. 2024;6(4):e240300.
31. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*. 2021;372:n71.
32. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Syst Rev*. 2021;10(1):39.
33. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–36.
34. Ishida N, Yagi N, Kitae H, Tomie A, Ichikawa H, Hiwa S, et al. Machine learning-based automatic detection system for demarcation line of gastric cancer with NBI images. *United European Gastroenterology Journal*. 2017;5(5_suppl).
35. Kanesaka T, Lee TC, Uedo N, Lin KP, Chen HZ, Lee JY, et al. Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging. *Gastrointest Endosc*. 2018;87(5):1339–44.
36. 李 强, 王 强, 李 强, 李 强, 李 强, 李 强, 李 强, et al. CNN 在 胃 癌 早 期 检 测 中 的 应 用. *中华消化内镜杂志*. 2019;85(9):761–4.
37. Yoon HJ, Kim S, Kim JH, Keum JS, Oh SI, Jo J, et al. A Lesion-Based Convolutional Neural Network Improves Endoscopic Detection and Depth Prediction of Early Gastric Cancer. *J Clin Med*. 2019;8(9).
38. Du W, Rao N, Yong J, Wang Y, Hu D, Gan T, et al. Automatic Early Gastric Cancer Segmentation in Gastroscopic Images Based on ResUnet. *Proceedings of the 2021 8th International Conference on Biomedical and Bioinformatics Engineering*; Kyoto, Japan: Association for Computing Machinery; 2022. p. 13–9.
39. Teramoto A, Shibata T, Yamada H, Hirooka Y, Saito K, Fujita H. Automated Detection of Gastric Cancer by Retrospective Endoscopic Image Dataset Using U-Net R-CNN. *Applied Sciences* [Internet]. 2021; 11(23):[11275 p.].
40. Cui ZH, Zhang QY, Zhang JW, Sun X, Wang Q, Lei Y, et al., editors. Application of Improved Mask R-CNN Algorithm Based on Gastroscopic Image in Detection of Early Gastric Cancer. *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*; 2022 27 June–1 July 2022.
41. Jin J, Zhang Q, Dong B, Ma T, Mei X, Wang X, et al. Automatic detection of early gastric cancer in endoscopy based on Mask region-based convolutional neural networks (Mask R-CNN)(with video). *Front Oncol*. 2022;12:927868.
42. Okumura S, Goudo M, Hiwa S, Yasuda T, Kitae H, Yasuda Y, et al. Demarcation Line Determination for Diagnosis of Gastric Cancer Disease Range Using Unsupervised Machine Learning in Magnifying Narrow-Band Imaging. *Diagnostics (Basel)*. 2022;12(10).
43. Qiu K, Bu X, Zhou H, Liu L, editors. Research on ME- NBI Gastric Lesion Recognition System based on Improved UNet Structure. *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*; 2022 14–16 Jan. 2022.
44. Teramoto A, Shibata T, Yamada H, Hirooka Y, Saito K, Fujita H. Detection and Characterization of Gastric Cancer Using Cascade Deep Learning Model in Endoscopic Images. *Diagnostics (Basel)*. 2022;12(8).
45. Yamashita T, Kurumi H, Fujii M, Sakaguchi T, Hashimoto T, Kinoshita H, et al. Objective Methods of 5-Aminolevulinic Acid-Based Endoscopic Photodynamic Diagnosis Using Artificial Intelligence for Identification of Gastric Tumors. *J Clin Med*. 2022;11(11).
46. Ma L, Su X, Ma L, Gao X, Sun M. Deep learning for classification and localization of early gastric cancer in endoscopic images. *Biomedical Signal Processing and Control*. 2023;79:104200.
47. Shibata T, Yamada H, Hirooka Y, Teramoto A. THE APPLICATION OF ARTIFICIAL INTELLIGENCE FOR THE PRE-SECONDARY DIAGNOSIS OF GASTRIC CANCER; FITTING FOR THE CLOUD BASED GASTRIC CANCER SCREENING SYSTEM. *Gastrointestinal Endoscopy*. 2023;97(6):AB780.
48. Takemoto S, Hori K, Yoshimasa S, Nishimura M, Nakajo K, Inaba A, et al. Computer-aided demarcation of early gastric cancer: a pilot comparative study with endoscopists. *J Gastroenterol*. 2023;58(8):741–50.
49. Zhang K, Wang H, Cheng Y, Liu H, Gong Q, Zeng Q, et al. Early gastric cancer detection and lesion segmentation based on deep learning and gastroscopic images. *Sci Rep*. 2024;14(1):7847.
50. Wang B, Lei C, Guo Y, Hong K, Kan X, Ouyang Y, et al. Task-specific prompting SAM for multi-task gastric cancer diagnosis in endoscopic images. *Expert Systems with Applications*. 2025;289:128329.
51. Huang C-T, Lee T-C, Kanesaka T, Uedo N, Wang H-P, Chang H-T, editors. Deep-Learning-Based Early Stage Gastric Cancer Detection in Narrow-Band-Imaging Gastroscopy Images Using Pyramid Spatial Pooling U-Net. *Ubi-Media Computing, Pervasive Systems, Algorithms and Networks*; 2026 2026//; Singapore: Springer Nature Singapore.

52. K.Menaga, C. R M, Pandiyarajan P, Preethibha C, Sivakumarraju M, Pinky BB, editors. Graph Learning-Enhanced U-Net for Early Gastric Cancer Detection and Segmentation. 2026 6th International Conference on Pervasive Computing and Social Networking (ICPCSN); 2026 6–8 May 2026.
53. Wang C, Li Y, Yao J, Chen B, Song J, Yang X, editors. Localizing and Identifying Intestinal Metaplasia Based on Deep Learning in Oesophagoscope. 2019 8th International Symposium on Next Generation Electronics (ISNE); 2019 9–10 Oct. 2019.
54. Siripoppohn V, Pittayanon R, Tiankanon K, Faknak N, Sanpavat A, Klaikaew N, et al., editors. Fast segmentation convolutional neural network with edge-guided path for real-time diagnosis of gastric intestinal metaplasia. 2021 25th International Computer Science and Engineering Conference (ICSEC); 2021 18–20 Nov. 2021.
55. Tiankanon K, Pittayanon R, Siripoppohn V, Faknak N, Sanpavat A, Wisedopas N, et al. CONVOLUTIONAL NEURAL NETWORK BASED ARTIFICIAL INTELLIGENT TO PERFORM REAL-TIME SEMATIC SEGMENTATION ON GASTRIC INTESTINAL METAPLASIA. *Gastrointestinal Endoscopy*. 2022;95(6):AB228.
56. Pornvoraphat P, Tiankanon K, Pittayanon R, Sunthornwetchapong P, Vateekul P, Rerknimitr R. Real-time gastric intestinal metaplasia diagnosis tailored for bias and noisy-labeled data with multiple endoscopic imaging. *Comput Biol Med*. 2023;154:106582.
57. Chen YT, Yang EH, Chang WL, Lin J, Cheng HC, Huang CR, editors. Mask Focal Modulation Network for Gastric Intestinal Metaplasia Segmentation. 2024 International Joint Conference on Neural Networks (IJCNN); 2024 30 June–5 July 2024.
58. Pornvoraphat P, Tiankanon K, Pittayanon R, Nupairoj N, Vateekul P, Rerknimitr R. Real-time gastric intestinal metaplasia segmentation using a deep neural network designed for multiple imaging modes on high-resolution images. *Knowledge-Based Systems*. 2024;300:112213.
59. Tiankanon K, Pungpipattrakul N, Pittayanon R, Lerttanatum N, Sanpavat A, Pornvoraphat P, et al. THE PERFORMANCE OF TRAINEE ENDOSCOPISTS IN DIAGNOSING AND SEGMENTING GASTRIC INTESTINAL METAPLASIA WITH AND WITHOUT ARTIFICIAL INTELLIGENCE GUIDANCE: A PRELIMINARY RESULT. *Gastrointestinal Endoscopy*. 2024;99(6):AB13.
60. Obukhova N, Motyko A, Savelev A, Saveleva O, Samarin A, Sigaeva J, editors. Automatic Segmentation of Metaplasia in an Endoscopic Decision Support System. 2025 10th International Conference on Image, Vision and Computing (ICIVC); 2025 16–18 July 2025.
61. Yu X, Li R, Zhang M, Wang Z, Zheng X. Few-shot image segmentation on gastric intestinal metaplasia with attention-based progressive prototype alignment. *Biomedical Signal Processing and Control*. 2026;120:110184.
62. Zhang M, Wang L, Yu Y, Liu L, Tao X, Gu L, et al. A Benchmark Dataset of Endoscopic Images and a Novel Deep Learning Method to Segment Gastric Intestinal Metaplasia Under Linked Color Imaging. *Computational Intelligence*. 2026;42(1):e70177.

Figures

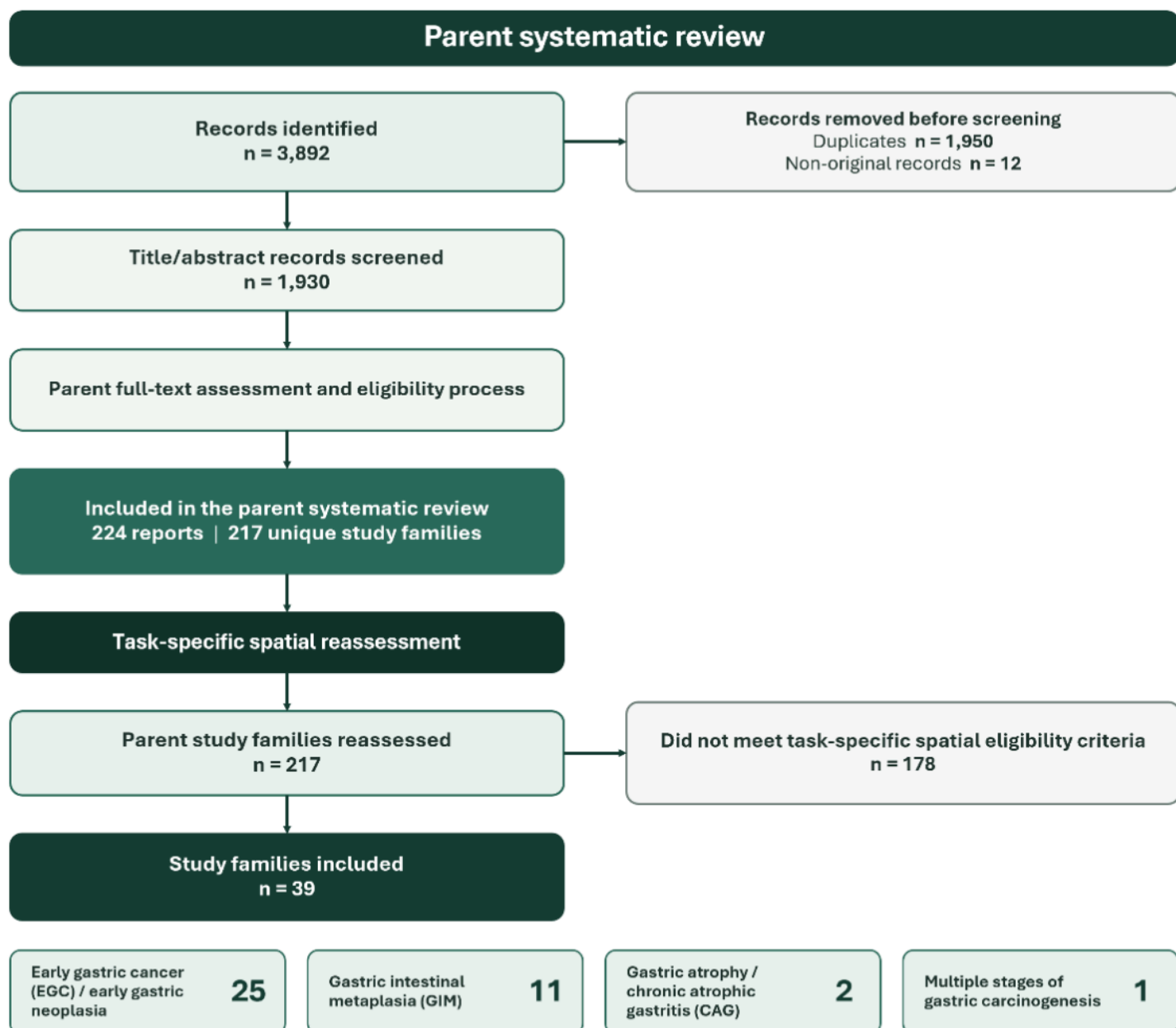


Figure 1

Study derivation and task-specific spatial reassessment flow diagram.

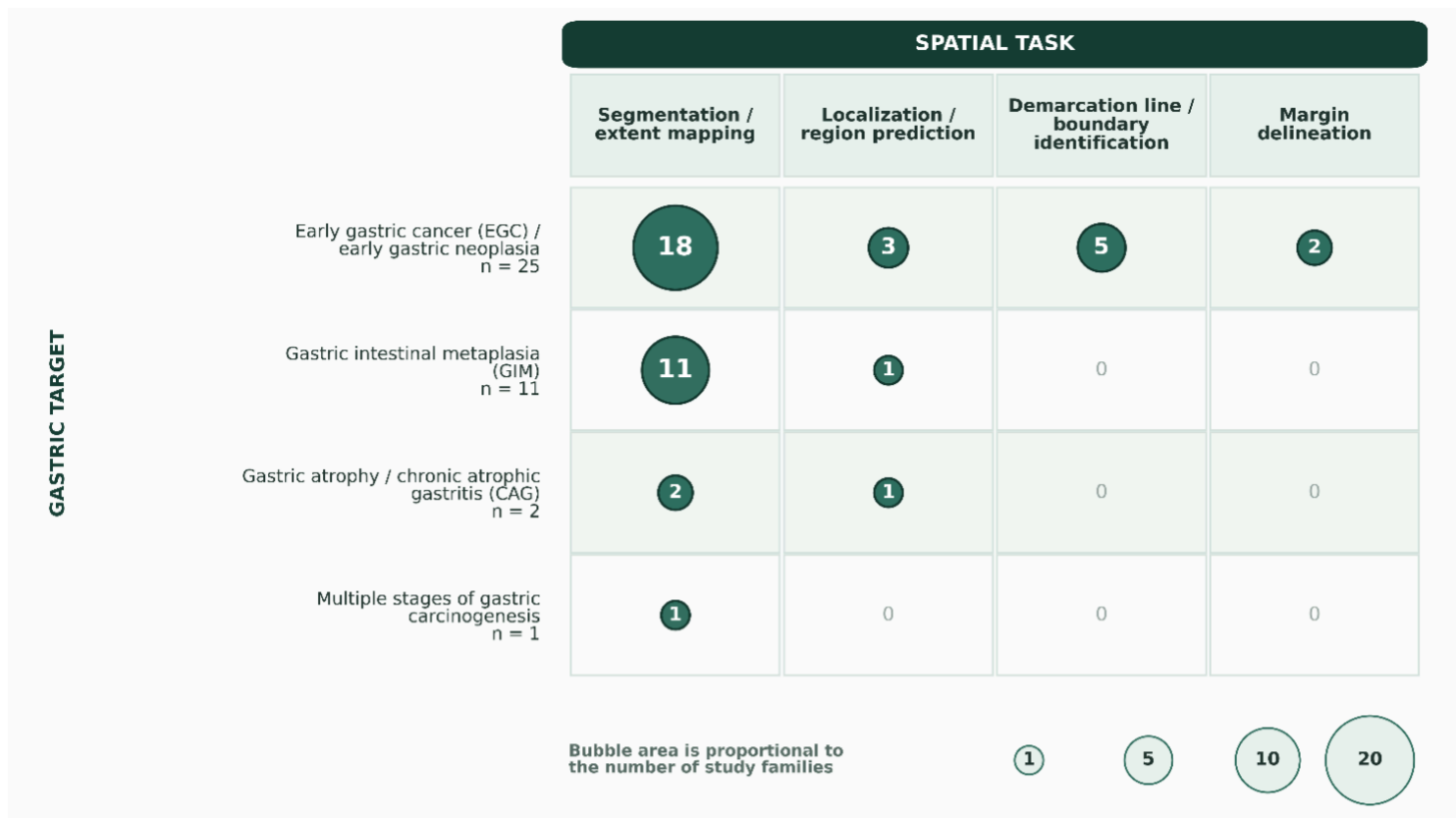


Figure 2

Distribution of spatial AI evidence across gastric targets and spatial tasks. Bubble area is proportional to the number of study families. Row totals indicate the total number of included study families within each gastric target category. Spatial-task categories were not mutually exclusive; a study family could contribute to more than one spatial task. EGC, early gastric cancer; GIM, gastric intestinal metaplasia; CAG, chronic atrophic gastritis.

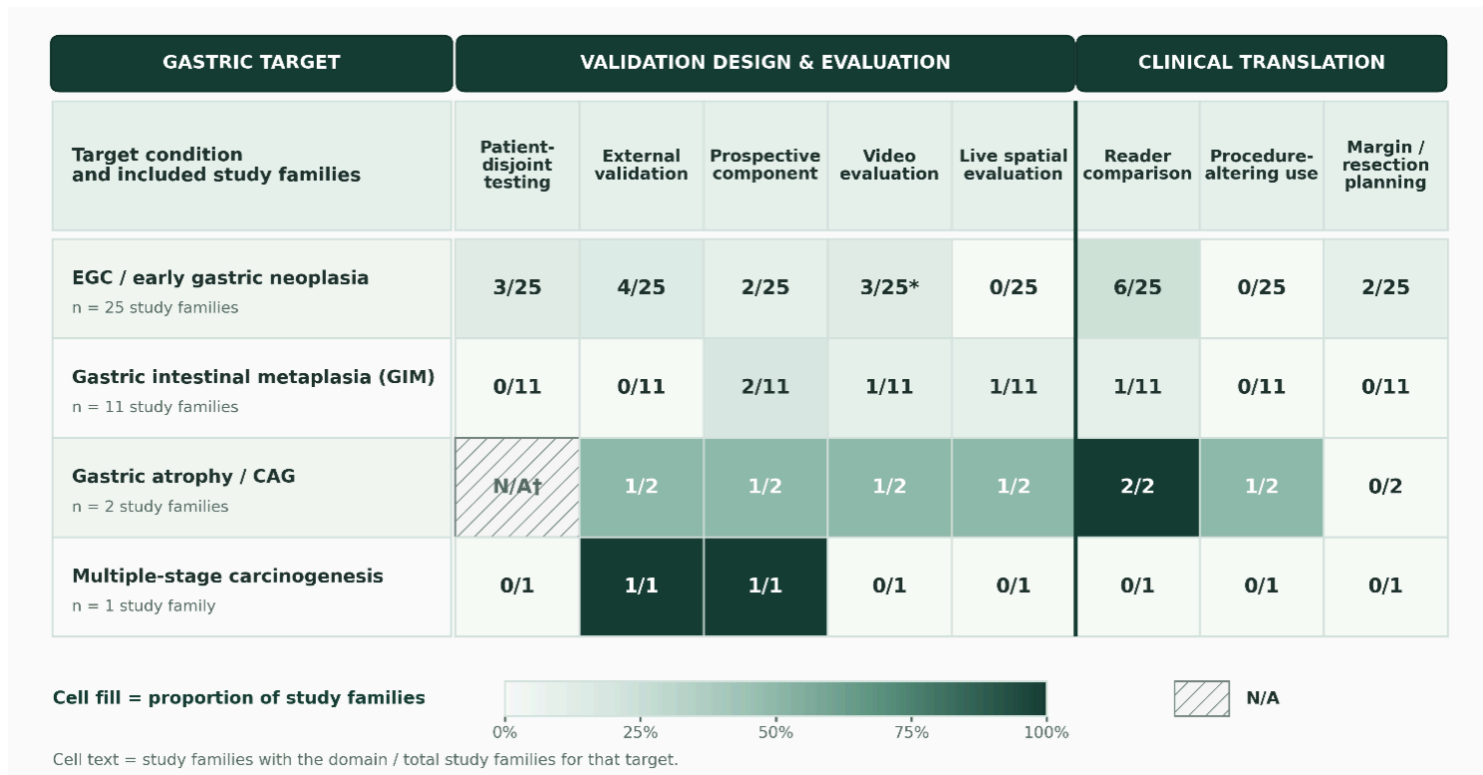


Figure 3

Validation and clinical-translation profile of spatial AI evidence across gastric target groups. Cell shading represents the proportion of study families within each target group meeting the specified validation or clinical-translation feature; cell text gives the corresponding numerator and denominator. Domains were considered separately and were not combined into a composite readiness score. *One additional EGC study family included limited or demonstration-only video evaluation and was not included in the numerator. †Patient-disjoint internal testing was not applicable to a prospective clinical cohort without a model-development split evaluated in that cohort. EGC, early gastric cancer; GIM, gastric intestinal metaplasia; CAG, chronic atrophic gastritis.

Supplementary Files

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

- [SupplementaryInformation.pdf](#)