Vol.7 Iss.7 2026 Communication & Education Review
https://doi.org/10.37420/j.cer.2026.108

Translating artificial intelligence into precision gastric cancer care

Qiang Zhi[1,†], Zhui Ke[2,†], Tian Wang[1], Jiaojiao Ma[1,*], Chunhua Zheng[1,*]

[1] Department of Gastrointestinal and Hernia Surgery, Guihang Guiyang Hospital, Guiyang 550025, Guizhou Province, China

[2] Xiangshui County People's Hospital, Jiangsu 224600, China

† These authors contributed equally to this work.

*** Corresponding authors, E-mail: qfldru@163.com(Jiaojiao Ma), qfldrh@163.com(Chunhua Zheng)**

Abstract

Artificial intelligence (AI) is being investigated across the gastric cancer care pathway, but the evidence required for clinical adoption extends beyond predictive accuracy. This narrative review critically examines AI applications in endoscopy, computed tomography, digital pathology, multimodal modeling, tumor microenvironment assessment, molecular profiling, and surgical assistance. Selected studies demonstrate substantial progress in lesion detection, staging, response prediction, and prognostic stratification. However, retrospective discrimination, prospective validation, and improved outcomes under AI-guided care represent distinct evidentiary achievements. A large randomized endoscopy trial illustrates this distinction: AI reduced examination blind spots but did not significantly improve the primary endpoint of gastric neoplasm detection after pathology review. Elsewhere, externally validated imaging and pathology models remain vulnerable to differences in acquisition protocols, patient selection, treatment context, and data availability. Associations between model scores and treatment outcomes cannot establish that a score identifies differential treatment benefit. Multimodal and molecular approaches introduce additional challenges related to missing inputs, assay reproducibility, and implementation costs. We propose a staged translational framework linking a clearly defined clinical decision to independent validation, prospective workflow evaluation, comparative impact assessment, and monitored deployment. Priorities include clinically appropriate reference standards, calibration and threshold-specific performance, subgroup evaluation, transparent reporting, human factors testing, and economic assessment. Clinical adoption should depend on reproducible incremental value within a specified care pathway, with safeguards proportionate to the consequences of error. These requirements provide a practical basis for converting promising gastric cancer AI models into reliable clinical tools.

Keywords:

gastric cancer; artificial intelligence; clinical translation; multimodal learning; precision oncology; clinical utility

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

1 Introduction

Gastric cancer remains a major global health burden. GLOBOCAN 2022 estimated approximately 968,000 new cases and 660,000 deaths worldwide, ranking gastric cancer fifth for both incidence and mortality [1]. Its molecular and histological heterogeneity, variation in stage at diagnosis, and differences in treatment response complicate individualized management. Artificial intelligence (AI) offers methods for extracting information from endoscopic images, computed tomography (CT), histopathological slides, and molecular measurements, and for integrating these data with clinical variables [2].

The resulting applications address different decisions. Endoscopic systems may help locate suspicious lesions or improve examination coverage. Imaging models may support staging and risk assessment, while pathology and multimodal models may estimate recurrence risk or response to a particular treatment. These tasks require different reference standards, validation strategies, and acceptable error rates. A model that distinguishes cancer from selected controls cannot be assumed to improve population screening, and a model that predicts response among treated patients does not necessarily identify the most effective treatment for an individual.

The central translational question is therefore whether using an AI system improves a specified clinical decision and its consequences. This distinction is illustrated by a large multicenter endoscopy trial in which improvements in examination quality did not translate into a significant improvement in the primary pathology-reviewed detection endpoint [3]. Progress also depends on reproducibility, integration with clinical workflows, access to required data, and the capacity to detect failure after deployment. This review examines these issues across seven areas of gastric cancer management and proposes an evidence pathway for clinically responsible adoption.

2 Scope and interpretive framework

This article is a narrative review organized around clinical decisions rather than algorithm families. Representative gastric cancer studies are discussed to illustrate the relationship between reported performance and the evidence needed for implementation. Primary reports are prioritized for numerical results; methodological guidance informs the discussion of study design and reporting. The synthesis is selective and does not constitute an exhaustive systematic review, quantitative meta-analysis, or formal grading of evidence certainty.

Four distinctions guide interpretation. First, retrospective model development is separated from independent geographic or temporal validation. Second, prospective data collection is distinguished from an intervention in which model outputs actually influence care. Third, diagnostic or prognostic discrimination is separated from comparative evidence that AI improves clinical outcomes. Finally, the unit of analysis is specified where it affects interpretation: patients, CT examinations, slides, and tissue measurement sites are not interchangeable. TRIPOD+AI provides a framework for transparent prediction-model reporting, whereas PROBAST+AI addresses quality, risk of bias, and applicability; reporting compliance alone does not establish clinical utility [4,5].

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

3 Applications across the gastric cancer care pathway

3.1 Endoscopic detection and characterization

Computer-aided detection and diagnosis serve related but distinct purposes. Detection systems flag potentially abnormal areas, whereas diagnostic systems classify lesions or estimate invasion depth. Real-time development and validation work, including the clinical decision support system reported by Gong and colleagues, demonstrates technical feasibility. Such studies should be interpreted according to their actual validation design rather than treated as evidence from randomized deployment trials [6].

Dong and colleagues randomized patients undergoing esophagogastroduodenoscopy at 24 Chinese hospitals. Among 29,514 participants in the intention-to-treat analysis, gastric neoplasm detection after pathology review was 1.42% with AI assistance and 1.25% without it (risk ratio 1.13, 95% confidence interval 0.92–1.38; P = 0.25). Detection based on the original pathology reports was higher with AI (4.06% versus 3.57%; P = 0.03). The mean number of blind spots decreased from 2.52 to 1.07, while procedure and inspection times increased [3]. The primary result does not establish equivalence, but it does not support a claim of improved pathology-reviewed detection. Reduced blind spots are a process outcome and cannot, by themselves, establish fewer missed cancers. Exploratory findings concerning operator experience or fatigue require confirmation before they determine deployment policy.

Emerging approaches extend beyond conventional image recognition. Yan and colleagues combined surface-enhanced Raman scattering with AI to map mucosal acidity. The work included 50 patient specimens and 1,516 measurement sites, with reported external-validation sensitivity of 87.79%, specificity of 85.04%, and accuracy of 86.89% [7]. Much of this evidence concerns analysis of excised specimens; it should not be described as a validated real-time clinical screening intervention. Multiple measurements from the same specimen also require patient-level separation and appropriate handling of within-patient dependence. Watanabe and colleagues investigated AI-based endoscopy combined with miR148a methylation for lesions classified as indefinite for dysplasia, illustrating the potential and limitations of adding molecular information to morphological assessment [8].

Future endoscopy trials should prespecify the reference pathology process, clinically consequential lesion categories, false-positive biopsies, examination time, and follow-up for missed disease. Endoscopist-plus-AI should be compared with endoscopist-alone performance in representative practice. Quality metrics can explain how an intervention works, but favorable secondary outcomes should not replace an unsuccessful primary endpoint after the results are known.

3.2 CT radiomics and deep learning

CT-based AI encompasses lesion detection, staging, prediction of residual nodal disease, and response or recurrence assessment. These applications operate at different points in care and should not be ranked by area under the receiver operating characteristic curve (AUC) alone. In the GANDA study, Chen and colleagues reported a study population of 4,606 patients and a separate real-world test dataset of 7,695 consecutive CT examinations. In the latter dataset, sensitivity and specificity for neoplasm detection were 83.2% and 93.1%, respectively [9]. These denominators describe different components of the investigation and should not be

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

combined or used interchangeably. Large-scale retrospective testing strengthens evidence of diagnostic validity but does not show that deploying the system improves referral or patient outcomes.

For preoperative T staging, Zheng and colleagues developed GTRNet using 1,792 patients, with reported AUCs of 0.86–0.95 and accuracies of 81%–85%. The method used the largest axial tumor slice, which limits the anatomical information available to the model [10]. He and colleagues evaluated a 2.5-dimensional approach for identifying pathological T1N0 disease in 3,164 surgically treated patients [11]. Such prediction may inform risk assessment, but pathological T1N0 status is not synonymous with eligibility for endoscopic resection; lesion characteristics and established curative-resection criteria remain necessary.

Pretreatment response models provide another example of task specificity. A transformer-based network evaluated by Zheng and colleagues in 1,205 patients predicted lymph node metastasis after neoadjuvant chemotherapy, with validation AUCs of 0.748–0.788 [12]. Cui and colleagues reported a CT-based deep learning radiomics nomogram in 719 patients, with AUCs of 0.829, 0.804, and 0.827 in the internal and two external validation cohorts [13]. By contrast, Huang and colleagues studied 268 patients using CT information that included changes between pretreatment and post-treatment examinations [14]. A model incorporating post-treatment information can assess response at that later time point, but cannot be presented as a baseline treatment-selection tool.

Jiang and colleagues evaluated a multitask model in 2,320 patients. External-validation AUC for peritoneal recurrence was 0.843, whereas the external-validation concordance index for disease-free survival was 0.610 [15]. Reporting both outcomes avoids implying that strong recurrence discrimination establishes equally strong survival prediction. Clinical-variable models have also been investigated for lymph node metastasis in early gastric cancer; these should be distinguished from imaging models and assessed for whether every predictor is available at the intended decision time [16].

Reproducibility remains a central requirement. A review of gastric cancer radiomics reported a median radiomics quality score of 44.4% of the maximum, reflecting limitations in that reviewed literature rather than a score applicable to every subsequent study [17]. Acquisition protocols, segmentation, preprocessing, feature definitions, and software versions should be reported and tested. The Image Biomarker Standardization Initiative provides standardized feature definitions and reference values that can support this process [18].

3.3 Digital pathology and pathomics

Whole-slide imaging permits computational analysis of tumor architecture, stromal composition, and immune-cell patterns. Diagnostic models and quantitative biomarker scoring may support pathologists, while prognostic models attempt to extract information beyond conventional staging [19]. Huang and colleagues demonstrated multicenter diagnostic and prognostic modeling from digital pathological images, providing an early example of multiple-instance learning in gastric cancer [20]. The relevant comparator for adoption is the complete pathology workflow, including review time, missed lesions, and discordant interpretations.

Chen and colleagues developed the PSGC pathomics signature and reported associations with survival and adjuvant chemotherapy outcomes [21]. These findings support further evaluation of risk stratification. They do not justify withholding chemotherapy from a model-defined subgroup: treatment allocation in observation-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

al cohorts can be confounded, and statistical significance in one subgroup but not another does not establish a treatment-by-marker interaction. A claim of treatment selection requires appropriate comparative evidence and prospective testing of the proposed decision rule.

Ding and colleagues developed a recurrence model integrating histopathological features with clinical variables in a multicenter cohort of 1,763 patients, with reported validation AUCs of 0.843–0.887. Evaluation included an exploratory post hoc analysis of a prospective trial population [22]. This use of trial-derived data should be distinguished from a trial of AI-guided care. For immunotherapy response, Han and colleagues reported AUCs of 0.985 in training, 0.921 in internal validation, and 0.914 and 0.927 in two external validation cohorts [23]. The training estimate should not be included in a range labeled as external-validation performance.

Pathology models must remain reliable across staining protocols, scanners, specimen types, and tumor sampling strategies. Patient-level and institution-level separation are essential to avoid leakage through related slides or shared processing patterns. Performance should be assessed on poor-quality and low-tumor-content material as well as selected high-quality slides. Explanatory overlays can help inspect model behavior, but cannot establish that the learned features are causal biological determinants. Clinical value also depends on the costs of slide digitization, storage, quality control, and specialist review.

3.4 Multimodal integration

Multimodal models combine complementary information, but their benefit depends on whether that information is available and useful at the clinical decision point. Chen and colleagues developed MuMo using radiological, pathological, and clinical data from 429 patients with HER2-positive gastric cancer. Reported AUCs were 0.821 for anti-HER2 treatment response and 0.914 for response to anti-HER2 treatment combined with immunotherapy [24]. These estimates arose from different treatment cohorts and do not show that the model can determine which regimen provides greater benefit to the same patient.

Gao and colleagues integrated pretreatment CT and hematoxylin-and-eosin-stained biopsy images in an incremental supervised contrastive learning model. Development and testing used retrospective data from 2,387 patients across 10 centers, followed by evaluation in 132 prospectively enrolled patients; AUCs across test cohorts were 0.846–0.876 [25]. The prospective component advances validation, but does not demonstrate that acting on model predictions improves outcomes. Related radiopathomics studies evaluated immunotherapy-based combination treatment in 298 patients and prognosis or immunotherapy response in a multicohort series of 1,133 patients [26,27]. Their results require interpretation in relation to treatment regimen, prediction horizon, and the availability of matched imaging and pathology.

Multimodal virtual biopsy has also been studied for predicting positive peritoneal lavage cytology. Ding and colleagues reported an AUC of 0.835 in a prospective validation cohort of 152 patients [28]. This supports further evaluation as a risk-assessment tool; it does not establish that staging laparoscopy or cytological confirmation can be omitted.

A practical limitation is missing or unusable input data. Requiring complete CT, digital pathology, and molecular information may exclude the patients or hospitals in which decision support is most needed. Validation

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

should report the proportion of eligible patients who can actually receive a prediction, performance for each available modality combination, and the behavior of the system when an input is absent or corrupted. Modality dropout and alternative single-modality pathways are candidate solutions, but need testing under realistic missingness patterns. Any incremental accuracy should be weighed against acquisition costs, processing time, and delays in treatment decisions.

3.5 Tumor microenvironment assessment and immunotherapy response

Imaging-based models have been developed to estimate immune and stromal features that usually require tissue assessment. Jiang and colleagues proposed biology-guided deep learning that links CT-derived representations with tumor microenvironment states and treatment outcomes. The study also identified model-defined heterogeneity among mismatch repair-deficient tumors receiving immunotherapy [29]. This is an exploratory stratification finding rather than a reason to deny an otherwise indicated treatment. A related deep learning radiomics study evaluated noninvasive microenvironment assessment and its association with prognosis and treatment response [30].

Sun and colleagues evaluated lymphoid and myeloid radiomic signatures in a study encompassing 2,600 patients. Among immunotherapy recipients, the objective response rate was 53.3% in the high-lymphoid, low-myeloid imaging subtype and 10.2% in the low-lymphoid, high-myeloid subtype [31]. These treatment-cohort results should not be interpreted as response rates for all 2,600 participants. Li and colleagues reported CT radiomics AUCs of 0.835 and 0.879 for microsatellite instability classification in two external test sets [32]. Classification of an established biomarker, association with treatment response, and improvement in treatment selection are separate claims requiring separate validation.

Spatial and single-cell analyses may help interpret these associations. Wang and colleagues identified macrophage–fibroblast–tumor co-infiltration patterns and developed a framework for prognostic assessment, while other integrative studies linked fibroblast-associated metabolic signatures with survival [33,34]. Such work supplies hypotheses about the biological context of model predictions, but does not establish that an imaging feature directly measures a specific immune-cell population.

CT-derived immune estimates remain indirect measurements of macroscopic tissue characteristics. Their relationship with tissue assays may vary with sampling location, treatment, inflammation, or necrosis. Reliable longitudinal use therefore requires paired imaging and tissue assessment at relevant time points, together with evaluation of prediction errors and uncertainty. Models should be tested for incremental value over clinically available biomarkers within the same patients and treatment setting.

3.6 Molecular profiling and liquid biopsy

Machine learning can combine molecular measurements into diagnostic or prognostic signatures. Chen and colleagues analyzed 702 plasma samples and developed a 10-metabolite diagnostic model, reporting sensitivity of 0.905 in an external test set compared with less than 0.40 for conventional protein-marker approaches in that study [35]. Cai and colleagues analyzed serum lipidomics from three cohorts totaling 944 participants and developed diagnostic and prognostic signatures [36]. These studies support analytical and clinical investigation, but their results do not by themselves establish the effectiveness of screening an asymptomatic popu-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

lation.

Case–control discrimination can be inflated by differences between selected cancer cases and healthy controls. Screening evaluation requires representative participants, prespecified thresholds, stage-specific sensitivity, specificity, predictive values at the intended prevalence, and follow-up of negative results. Dietary patterns, comorbidities, medication use, sample handling, and laboratory platforms can alter molecular measurements. Cross-platform assay reproducibility and the effect of these variables should be established before clinical implementation.

Other molecular studies address disease biology rather than immediate diagnosis. Vo and colleagues used AI-guided analyses to characterize the gastric cancer continuum, providing a framework for studying progression and candidate intervention targets [37]. A molecular trajectory reconstructed from cross-sectional data should not be equated with a validated longitudinal prediction of an individual's progression. Likewise, fibroblast-associated signatures derived from public multiomics datasets require independent assay development and prospective clinical evaluation before use as treatment-selection tests [34].

A plausible use for a blood-based test is triage before endoscopy, particularly where endoscopic capacity is constrained. This proposed pathway requires direct assessment of cancer yield, unnecessary procedures, missed disease, and cost. Endoscopy provides localization and tissue sampling, and may enable treatment of suitable early lesions; a circulating biomarker cannot be presumed to replace those functions. Comparative studies should evaluate a complete screening strategy rather than compare isolated sensitivity estimates from unrelated populations.

Selected studies of endoscopy, imaging, pathology, multimodal modeling, immune assessment, and molecular profiling are summarized in Table 1, with study design and analysis population distinguished from reported performance and clinical interpretation.

Table 1. Selected gastric cancer AI studies and implications for clinical translation

Study and application	Design and analysis population	Selected reported finding	Clinical interpretation
Dong et al. (2026) [3] Endoscopy	Multicenter randomized trial; 29,514 participants in the intention-to-treat analysis.	Pathology-reviewed detection: 1.42% versus 1.25%; RR 1.13 (95% CI 0.92–1.38). Blind spots: 1.07 versus 2.52.	No statistically significant improvement in the primary endpoint. Better examination coverage does not establish fewer missed cancers.
Yan et al. (2025) [7] Mucosal acidity	Translational specimen study; 50 patient specimens and 1,516 measurement sites.	Reported external-validation sensitivity 87.79%, specificity 85.04%, and accuracy 86.89%.	Supports further technical evaluation. Specimen-based and site-level results are not evidence of patient-level screening effectiveness.
Chen et al. (2026) [9] CT detection	Retrospective development/testing population of 4,606 patients; separate real-world test dataset of 7,695 CT examinations.	Real-world test sensitivity 83.2% and specificity 93.1% for neoplasm detection.	Supports diagnostic validity across study settings. The different denominators should be distinguished; clinical impact remains to be tested.
Zheng et al. (2026) [10] CT T staging	Retrospective multicenter study; 1,792 patients; independent external test cohorts.	Reported AUCs 0.86–0.95; accuracy 81%–85%.	Largest axial tumor slice used. Full anatomical assessment and the consequences of staging errors require evaluation.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

Study and application	Design and analysis population	Selected reported finding	Clinical interpretation
Cui et al. (2022) [13] Chemotherapy response	Retrospective four-center cohort; 719 patients.	AUC 0.829 internally; 0.804 and 0.827 in two external validation cohorts.	Predicts response within the studied context. Does not establish superiority of an AI-guided treatment-selection strategy.
Jiang et al. (2022) [15] Peritoneal recurrence	Retrospective multicenter study; 2,320 patients.	External AUC 0.843 for peritoneal recurrence; external C-index 0.610 for disease-free survival.	Task-specific performance differs. Recurrence discrimination should not be presented as equivalent survival-prediction accuracy.
Ding et al. (2026) [22] Pathology and recurrence	Retrospective multicenter cohort; 1,763 patients; additional exploratory analysis of a prospective trial population.	Validation AUCs 0.843–0.887 for early recurrence.	Trial-derived validation is not a randomized evaluation of AI-guided follow-up or treatment.
Han et al. (2024) [23] Pathology and immunotherapy	Multicenter retrospective study; 584 patients.	AUC 0.921 internally; 0.914 and 0.927 in two external validation cohorts.	Training AUC 0.985 is excluded from the validation range. Independent intervention studies are needed before treatment selection.
Chen et al. (2024) [24] MuMo	Multimodal study; 429 HER2-positive patients: 310 receiving anti-HER2 therapy and 119 receiving combination immunotherapy.	AUC 0.821 for anti-HER2 response and 0.914 for combination-treatment response.	Different treatment cohorts cannot establish which regimen benefits an individual patient more.
Gao et al. (2024) [25] Multimodal response	Retrospective data from 2,387 patients at 10 centers; prospective evaluation in 132 patients.	AUCs 0.846–0.876 across test cohorts.	Prospective discrimination strengthens validity but does not establish improved outcomes from acting on predictions.
Ding et al. (2025) [28] Peritoneal cytology	Multimodal model with prospective validation in 152 patients.	Prospective-validation AUC 0.835 for cytology positivity.	Potential risk-assessment support. Does not establish safe omission of staging laparoscopy or cytological confirmation.
Sun et al. (2024) [31] Immune context	Multicohort study encompassing 2,600 patients; response evaluated in immunotherapy cohorts.	Response rates of 53.3% and 10.2% in selected contrasting imaging immune subtypes.	Response results apply to treated subsets. Imaging estimates are indirect and do not establish causal immune mechanisms.
Chen et al. (2024) [35] Plasma metabolomics	Multicenter analysis of 702 plasma samples with an external test set.	External sensitivity 0.905 for the 10-metabolite diagnostic model.	Screening performance requires specificity, predictive values at target prevalence, and follow-up of negative results.
Cai et al. (2024) [36] Serum lipidomics	Retrospective analysis of three cohorts totaling 944 participants.	Diagnostic and prognostic lipid signatures reported.	Assay transferability and prospective pathway-level value remain to be established.

Abbreviations: AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; C-index, concordance index; CI, confidence interval; CT, computed tomography; HER2, human epidermal growth factor receptor 2; RR, risk ratio.

For Dong et al., paired values denote AI-assisted versus unassisted endoscopy. Estimates are study-specific and are not directly comparable. Population sizes refer to the indicated dataset or study component and are not necessarily the denominator for every reported endpoint. Study descriptions are not formal evidence-certainty ratings.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

3.7 Surgical assistance

AI applications in gastric cancer surgery include recognition of procedural phases, identification of anatomical structures, video annotation, and assessment of technical performance. These functions differ substantially from autonomous instrument control. The review by Kinoshita and Komatsu describes the opportunities for surgical assistance and the limitations of translating technical demonstrations into operative practice [38].

Near-term evaluations should prioritize assistance under surgeon oversight. Relevant outcomes include recognition of critical anatomy, false or delayed alerts, operative interruptions, complications, and the learning curve. Models trained on curated video segments must be tested in the presence of bleeding, smoke, tissue deformation, anatomical variation, and incomplete views. Surgical-phase recognition accuracy is insufficient evidence of safer surgery. Equally, errors in diagnostic AI can cause serious downstream harm; risk should be judged by the clinical consequences of the intended use, rather than by a categorical assumption that diagnosis is readily reversible whereas surgery is not.

4 Why predictive performance does not ensure clinical benefit

4.1 Study design and the claim being tested

Model validation and clinical impact evaluation answer different questions. Geographic validation assesses transfer to another institution, temporal validation assesses performance in a later population, and prospective evaluation can test data availability and workflow feasibility. None necessarily demonstrates that model-guided management is superior to usual care. The use of a prospective or randomized source cohort does not convert a post hoc prediction analysis into a randomized test of the AI intervention [39,40].

Treatment-response models require particular caution. A model trained among recipients of one treatment estimates outcomes under that observed treatment and case mix. Selecting between treatments requires evidence concerning differential benefit, ideally from appropriately designed randomized data and a prespecified interaction analysis. Retrospective adjustment or propensity methods can address measured confounding but cannot eliminate unmeasured confounding. Apparent differences in subgroup P values are not sufficient evidence of effect modification. These issues are directly relevant to claims that a pathology or imaging score can identify patients who should omit chemotherapy or receive an alternative immunotherapy regimen [21,29].

Clinical endpoints should follow the intended decision. For an endoscopic detection tool, pathology-confirmed clinically relevant lesions and follow-up for missed cancers may be more informative than frame-level accuracy. For a treatment-selection tool, outcomes can include survival, toxicity, or quality of life. Changes in management are intermediate outcomes: a changed decision may be appropriate or harmful. Composite endpoints should be used only when their components are clinically coherent and independently reported. Negative primary results should remain central to interpretation, with secondary and subgroup findings used to generate appropriately testable hypotheses.

4.2 Bias and transportability

Large datasets do not guarantee representative evidence. Many gastric cancer studies draw on special-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

ist-center cohorts, including substantial contributions from East Asia. Their findings may be relevant elsewhere, but transportability must be demonstrated across healthcare systems, disease prevalence, histological subtypes, and treatment practices. A surgical cohort can also be a poor proxy for a population being assessed before any treatment, because selection for surgery changes the spectrum of disease [5,11].

Validation should separate patients before image patches, slides, or repeated scans are generated. Feature selection, imputation, harmonization, and threshold tuning must be confined to development data. Reusing public datasets across multiple publications can create apparent replication without independent evidence. External validation should evaluate calibration as well as discrimination, report clinically relevant operating thresholds, and include confidence intervals. Subgroup assessment should consider age, sex, histological subtype, anatomical site, stage, institution, and acquisition platform where sample sizes permit. Small subgroup samples require explicit uncertainty rather than claims of equivalent performance [4,5].

4.3 Interpretability and human factors

Feature attribution and saliency maps can reveal unexpected associations and assist error review. They do not prove that a prediction is biologically correct or that the model will remain reliable after a change in scanner, stain, or patient population. The value of an explanation depends on its stability, fidelity, and usefulness to the clinical user. A visually plausible heatmap should not substitute for an error analysis or an independent validation study.

Clinical evaluation should measure how users respond to correct and incorrect predictions, whether alerts are ignored, and whether confidence in AI causes inappropriate overrides of other evidence. Latency, image acquisition failures, staff training, and downtime can change the effective intervention. Early live evaluations should document these features, including user learning curves and modifications to the system, in line with DECIDE-AI [39]. A technically accurate model may still fail if its output arrives too late or cannot be acted upon within the existing care pathway.

4.4 Regulation and economic value

Regulation of medical AI should be discussed in relation to a specific jurisdiction, intended use, and product lifecycle. Existing frameworks should not be described as a universal regulatory vacuum. For example, the US Food and Drug Administration provides guidance on predetermined change control plans for AI-enabled device software functions [41]. A system should not be introduced into patient care under a quality-improvement label to avoid requirements applicable to its medical purpose. Research governance, authorization where required, cybersecurity, version control, and post-deployment surveillance need to be planned before implementation.

Economic value depends on the entire pathway. Relevant costs include hardware, licensing, integration, slide digitization, storage, assay performance, staff training, false-positive investigations, and maintenance. Benefits may include improved outcomes or more efficient use of specialist capacity, but these must be measured. Cost-effectiveness, budget impact, and affordability are related but distinct considerations. A small increase in AUC does not establish economic value, while a modest accuracy gain could still be useful if it safely reduces delays or workload. Access across hospitals with different resources should be assessed along-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

side average performance.

5 A practical translational pathway

5.1 Define the decision and validate the model

Development should begin with an intended-use statement specifying the patient population, clinical setting, available inputs, time of prediction, user, and action supported by the output. A realistic comparator and the consequences of false-positive and false-negative results should determine performance targets. An endoscopy quality tool, a staging assistant, and a chemotherapy-selection model require different evidence plans. Technical objectives should be linked to a clinical objective before model training begins.

Independent validation should then test a locked model with prespecified thresholds and transparent handling of missing data. Reports should include calibration, uncertainty, subgroup performance, and comparison with feasible clinical alternatives. Multimodal systems should be assessed when inputs are missing or degraded. Radiomics pipelines should use reproducible processing and feature definitions, supported where applicable by IBSI [18]. Multicenter collaboration can improve representation. Federated learning is one possible training approach, but it does not itself guarantee privacy, remove institutional bias, or replace independent evaluation.

5.2 Evaluate the complete clinical intervention

Prospective silent evaluation, in which predictions do not influence care, can test operational feasibility and reveal drift before an intervention study. Early live evaluation should then assess usability, safety, errors, and clinician interaction [39]. Depending on the clinical question, comparative evaluation may use individual or cluster randomization, an appropriate stepped-wedge design, or another rigorously justified design. Protocols and reports should specify the model version, user training, workflow, comparator, input failures, and handling of disagreements; SPIRIT-AI and CONSORT-AI provide relevant reporting guidance [40,42].

Trials should evaluate the end-to-end pathway. For screening, this includes diagnostic confirmation and follow-up; for treatment selection, it includes the consequences of acting on the prediction. Clinical benefit, harms, workload, and economic outcomes should be evaluated together. For safety-critical decisions, the evidence threshold should reflect the severity and reversibility of an error. Technical success should not be used to infer an unmeasured patient benefit. Figure 1 summarizes the clinical applications (panel A), the proposed evidence pathway (panel B), and the requirements that apply throughout the lifecycle (panel C).

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)



Figure 1. A framework for translating artificial intelligence into gastric cancer care

(A) Clinical applications considered in this review. (B) Five stages link a defined clinical decision to independent validation, prospective workflow evaluation, comparative impact assessment, and monitored deployment. Solid arrows indicate progression; the dashed feedback arrow indicates reassessment after material changes to the model or clinical setting. Progression depends on the intended claim and accumulated evidence. (C) Requirements that apply throughout the lifecycle. This conceptual synthesis draws on prediction-model reporting and appraisal guidance, radiomics standardization, clinical evaluation guidance, and regulatory principles [4,5,18,39–42]. It is not a validated scoring system or a set of regulatory phases. AI, artificial intelligence; CT, computed tomography.

5.3 Monitor performance after deployment

Implementation requires named clinical and technical responsibilities, an audit trail, and a defined response to failure. Monitoring should examine shifts in input distributions, missingness, calibration, subgroup performance, and clinical outcomes when these become available. Changes to scanners, stains, assays, or treatment protocols can alter performance even if the software is unchanged. Model updates should be versioned, assessed against prespecified criteria, and governed by the applicable regulatory pathway [41].

A fallback to usual clinical assessment is necessary when data quality is inadequate or the system is un-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

available. Periodic reassessment should determine whether the tool continues to add value and whether unequal access or performance is worsening disparities. Progression between stages depends on the intended claim and accumulated evidence; it is not equivalent to assigning drug-development phases to AI systems.

6 Limitations of the evidence synthesis

This narrative review emphasizes illustrative studies and cannot quantify the prevalence of bias or establish the comparative effectiveness of different algorithms. Reported AUCs concern different populations, outcomes, validation designs, and prediction times, and should not be pooled informally or used as a league table. Some studies may draw on overlapping institutional or public datasets, while model updates and implementation results may not be fully represented in publications. Formal risk-of-bias assessment and a reproducible systematic search would be required for an exhaustive evidence appraisal. The proposed translational pathway is a synthesis of methodological principles and clinical considerations, rather than an empirically validated scoring system.

7 Conclusions

AI has demonstrated technical potential across gastric cancer detection, staging, pathological assessment, molecular characterization, and outcome prediction. Clinical adoption requires evidence that using a specific system improves a defined care pathway with acceptable harms and resource requirements. Retrospective performance, external validation, prospective observation, and randomized impact evaluation should be interpreted separately. Priorities are independent and representative validation, appropriate reference standards, reliable performance with imperfect inputs, and prospective assessment of the clinician–AI intervention. Treatment-selection claims require evidence of differential benefit, and deployment requires ongoing monitoring and governance. These steps can align AI development with the clinical needs of patients with gastric cancer.

Funding

This study was supported by the 2025 Yancheng Municipal Health Commission Scientific Research Project (Grant No. YK2025177).

Author contributions

Qiang Zhi was responsible for study design, data collection, data analysis, and drafting the manuscript. Zhui Ke was responsible for study design, data collection, statistical analysis, and drafting the manuscript. Tian Wang was responsible for data curation, literature search, figure preparation, and manuscript revision. Jiaojiao Ma was responsible for study conception and design, research supervision, quality control, and manuscript review and finalization. Chunhua Zheng was responsible for study design guidance, data interpretation, and manuscript review and revision. All authors have read and approved the final manuscript.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

References

[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263. doi:10.3322/caac.21834.

[2] Fu J, Fang M, Wu L, et al. Development, advancement, and clinical integration of artificial intelligence technology in gastric cancer. Chin Med J (Engl). 2025;138(24):3332-3350. doi:10.1097/CM9.0000000000003922.

[3] Dong Z, Wu L, Du H, et al. Effect of a computer-aided device for detecting gastric neoplasms: a multicenter, randomized controlled trial. Gastroenterology. 2026;170(7):1518-1532. doi:10.1053/j.gastro.2025.12.020.

[4] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

[5] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.

[6] Gong EJ, Bang CS, Lee JJ, et al. Deep learning-based clinical decision support system for gastric neoplasms in real-time endoscopy: development and validation study. Endoscopy. 2023;55(8):701-708. doi:10.1055/a-2031-0691.

[7] Yan H, Li Z, Zhao J, et al. AI-assisted detection of early gastric cancer via visualization of mucosal acidity compromise during endoscopy. Adv Sci (Weinh). 2025;12(46):e04932. doi:10.1002/advs.202504932.

[8] Watanabe Y, Oikawa R, Agawa S, et al. Combination of artificial intelligence-based endoscopy and miR148a methylation for gastric indefinite dysplasia diagnosis. J Clin Lab Anal. 2022;36(1):e24122. doi:10.1002/jcla.24122.

[9] Chen X, Xia Y, Yao L, et al. Gastric neoplasm detection at contrast-enhanced CT with deep learning. Radiol Artif Intell. 2026;8(1):e250145. doi:10.1148/ryai.250145.

[10] Zheng G, Wang H, Chai X, et al. Interpretable deep learning for multicenter gastric cancer T staging from CT images. NPJ Digit Med. 2026;9(1):2. doi:10.1038/s41746-025-02002-5.

[11] He J, Xu J, Chen W, et al. Development of a deep learning model for T1N0 gastric cancer diagnosis using 2.5D radiomic data in preoperative CT images. NPJ Precis Oncol. 2025;9(1):249. doi:10.1038/s41698-025-01055-9.

[12] Zheng Y, Qiu B, Liu S, et al. A transformer-based deep learning model for early prediction of lymph node metastasis in locally advanced gastric cancer after neoadjuvant chemotherapy using pretreatment CT images. EClinicalMedicine. 2024;75:102805. doi:10.1016/j.eclinm.2024.102805.

[13] Cui Y, Zhang J, Li Z, et al. A CT-based deep learning radiomics nomogram for predicting the response to neoadjuvant chemotherapy in patients with locally advanced gastric cancer: a multicenter cohort study. EClinicalMedicine. 2022;46:101348. doi:10.1016/j.eclinm.2022.101348.

[14] Huang ZN, Zhang HX, Sun YQ, et al. Multi-cohort study in gastric cancer to develop CT-based radiomic models to predict pathological response to neoadjuvant immunotherapy. J Transl Med. 2025;23(1):362. doi:10.1186/s12967-025-06363-z.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

[15] Jiang Y, Zhang Z, Yuan Q, et al. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. Lancet Digit Health. 2022;4(5):e340-e350. doi:10.1016/S2589-7500(22)00040-1.

[16] Chen Y, Bai K, Yang M, et al. Machine learning for lymph node metastasis prediction in early gastric cancer: a comparative analysis. Int J Med Sci. 2026;23(3):1033-1045. doi:10.7150/ijms.124229.

[17] Chen Q, Zhang L, Liu S, et al. Radiomics in precision medicine for gastric cancer: opportunities and challenges. Eur Radiol. 2022;32(9):5852-5868. doi:10.1007/s00330-022-08704-8.

[18] Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020;295(2):328-338. doi:10.1148/radiol.2020191145.

[19] Choi S, Kim S. Artificial intelligence in the pathology of gastric cancer. J Gastric Cancer. 2023;23(3):410-427. doi:10.5230/jgc.2023.23.e25.

[20] Huang B, Tian S, Zhan N, et al. Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: a retrospective multicentre study. EBioMedicine. 2021;73:103631. doi:10.1016/j.ebiom.2021.103631.

[21] Chen D, Fu M, Chi L, et al. Prognostic and predictive value of a pathomics signature in gastric cancer. Nat Commun. 2022;13(1):6903. doi:10.1038/s41467-022-34703-w.

[22] Ding P, Chen S, Guo H, et al. A deep learning-based digital biopsy for predicting early recurrence in gastric cancer. Nat Commun. 2026;17(1):5220. doi:10.1038/s41467-026-71347-6.

[23] Han Z, Zhang Z, Yang X, et al. Development and interpretation of a pathomics-driven ensemble model for predicting the response to immunotherapy in gastric cancer. J Immunother Cancer. 2024;12(5):e008927. doi:10.1136/jitc-2024-008927.

[24] Chen Z, Chen Y, Sun Y, et al. Predicting gastric cancer response to anti-HER2 therapy or anti-HER2 combined immunotherapy based on multi-modal data. Signal Transduct Target Ther. 2024;9(1):222. doi:10.1038/s41392-024-01932-y.

[25] Gao P, Xiao Q, Tan H, et al. Interpretable multi-modal artificial intelligence model for predicting gastric cancer response to neoadjuvant chemotherapy. Cell Rep Med. 2024;5(12):101848. doi:10.1016/j.xcrm.2024.101848.

[26] Huang W, Wang X, Zhong R, et al. Multimodal radiopathomics signature for prediction of response to immunotherapy-based combination therapy in gastric cancer using interpretable machine learning. Cancer Lett. 2025;631:217930. doi:10.1016/j.canlet.2025.217930.

[27] Wang W, Huang W, Yang X, et al. Multimodal radiopathomics approach for predictions of prognosis and immunotherapy response in patients with gastric cancer: a multicohort retrospective study. Int J Surg. 2025;111(11):7606-7621. doi:10.1097/JS9.0000000000002939.

[28] Ding P, Yang J, Guo H, et al. Multimodal artificial intelligence-based virtual biopsy for diagnosing abdominal lavage cytology-positive gastric cancer. Adv Sci (Weinh). 2025;12(15):e2411490. doi:10.1002/advs.202411490.

[29] Jiang Y, Zhang Z, Wang W, et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. Nat Commun. 2023;14(1):5135. doi:10.1038/s41467-023-40890-x.

[30] Jiang Y, Zhou K, Sun Z, et al. Non-invasive tumor microenvironment evaluation and treatment response prediction in gastric cancer using deep learning radiomics. Cell Rep Med. 2023;4(8):101146. doi:10.1016/

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

j.xcrm.2023.101146.

[31] Sun Z, Zhang T, Ahmad MU, et al. Comprehensive assessment of immune context and immunotherapy response via noninvasive imaging in gastric cancer. J Clin Invest. 2024;134(6):e175834. doi:10.1172/JCI175834.

[32] Li Z, Ding Z, Lian Y, et al. Computed tomography radiomics to predict microsatellite instability status and immunotherapy response in gastric cancer. Insights Imaging. 2025;16(1):177. doi:10.1186/s13244-025-02050-1.

[33] Wang Q, Ni Y, Lu S, et al. Multi-dimensional omics integrated machine learning framework identifies macrophage-fibroblast-tumor co-infiltration patterns to predict prognosis in gastric cancer. NPJ Digit Med. 2026;9(1):10. doi:10.1038/s41746-025-02179-9.

[34] Ngoc Tram VT, Minh Xuan DT, Quy Nguyen DP, et al. Integrative multi-omics and machine-learning approaches uncover a novel metabolic-related signature associated with cancer-associated fibroblasts in gastric cancer development. Comput Biol Med. 2025;195:110653. doi:10.1016/j.compbiomed.2025.110653.

[35] Chen Y, Wang B, Zhao Y, et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. Nat Commun. 2024;15(1):1657. doi:10.1038/s41467-024-46043-y.

[36] Cai ZR, Wang W, Chen D, et al. Diagnosis and prognosis prediction of gastric cancer by high-performance serum lipidome fingerprints. EMBO Mol Med. 2024;16(12):3089-3112. doi:10.1038/s44321-024-00169-0.

[37] Vo D, Ghosh P, Sahoo D. Artificial intelligence-guided discovery of gastric cancer continuum. Gastric Cancer. 2023;26(2):286-297. doi:10.1007/s10120-022-01360-3.

[38] Kinoshita T, Komatsu M. Artificial intelligence in surgery and its potential for gastric cancer. J Gastric Cancer. 2023;23(3):400-409. doi:10.5230/jgc.2023.23.e27.

[39] Vasey B, Nagendran M, Campbell B, 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):924-933. doi:10.1038/s41591-022-01772-9.

[40] Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x.

[41] US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. Guidance for industry and Food and Drug Administration staff. August 2025. Accessed October 27, 2026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence.

[42] Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)