

Review

Open Access



# Artificial intelligence for post-polypectomy surveillance: a scoping review of emerging tools in colorectal cancer prevention

Stefania Ferrari<sup>1,#</sup>, Silvia Negro<sup>1,#</sup>, Francesco Celotto<sup>1</sup>, Quoc Riccardo Bao<sup>1</sup>, Giovanni Madeo<sup>2</sup>, Alessandra Pulvirenti<sup>1,†</sup>, Gaya Spolverato<sup>1,†</sup>

<sup>1</sup>General Surgery 3, Department of Surgical, Oncological and Gastroenterological Sciences (DiSCOG), University of Padova, Padova 35128, Italy.

<sup>2</sup>Department of Surgical, Oncological and Gastroenterological Sciences (DiSCOG), University of Padova, Padova 35128, Italy.

#Authors should be considered as co-first authors.

†Authors should be considered as co-last authors.

**Correspondence to:** Dr. Stefania Ferrari, General Surgery 3, Department of Surgical, Oncological and Gastroenterological Sciences (DiSCOG), University of Padova, Padova 35128, Italy. E-mail: stefania.ferrari10@gmail.com

**How to cite this article:** Ferrari S, Negro S, Celotto F, Bao QR, Madeo G, Pulvirenti A, Spolverato G. Artificial intelligence for post-polypectomy surveillance: a scoping review of emerging tools in colorectal cancer prevention. *Art Int Surg*. 2025;5:490-504. <https://dx.doi.org/10.20517/ais.2025.65>

**Received:** 12 Jul 2025 **First Decision:** 3 Sep 2025 **Revised:** 16 Sep 2025 **Accepted:** 9 Oct 2025 **Published:** 26 Nov 2025

**Academic Editor:** Andrew Gumbs **Copy Editor:** Xing-Yue Zhang **Production Editor:** Xing-Yue Zhang

## Abstract

**Aim:** This scoping review aimed to synthesize current evidence on the application of artificial intelligence (AI), including natural language processing (NLP) and large language models (LLMs), in post-polypectomy surveillance for colorectal cancer (CRC). Specific objectives were to assess technological advances, evaluate their impact on guideline adherence, and identify gaps for future research.

**Methods:** We conducted a scoping review following the Arksey and O'Malley framework and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Searches across PubMed, EMBASE, Scopus, and Web of Science identified studies applying AI to CRC surveillance after polypectomy. Eligible studies investigated AI models for interval assignment or risk stratification using colonoscopy and pathology data.

**Results:** Of 950 screened articles, seven met the inclusion criteria. Five studies used NLP-based decision support tools, achieving concordance rates of 81.7%-99.9% with guideline-recommended surveillance intervals,



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



consistently outperforming clinician recommendations. Two studies evaluated ChatGPT-4 in clinical decision making; fine-tuned models demonstrated an accuracy of up to 85.7%, surpassing that of physicians in retrospective and simulated scenarios. NLP systems demonstrated technical maturity and scalability, while LLMs offered flexible, user-friendly interfaces but were less reliable in complex clinical scenarios.

**Conclusion:** AI tools, particularly NLP-enhanced systems, demonstrate strong potential to standardize post-polypectomy surveillance and improve guideline adherence. LLMs are promising but remain under validation. Future research should assess clinical implementation, long-term outcomes, and integration within electronic health records.

**Keywords:** Colorectal cancer, colonoscopy, surveillance, artificial intelligence, natural language processing, large language models

## INTRODUCTION

Colorectal cancer (CRC) represents a major public health challenge, being one of the most commonly diagnosed malignancies and a leading cause of cancer-related deaths globally<sup>[1]</sup>. In recent decades, population-level screening and the widespread use of colonoscopy have led to a reduction in CRC incidence of more than 30% among adults aged 50 and older<sup>[2-5]</sup>. However, individuals with a history of advanced adenomas or CRC remain at increased risk of recurrence or metachronous neoplasia<sup>[6]</sup>. For this reason, they are enrolled in surveillance colonoscopy programs aimed at early lesion detection<sup>[7,8]</sup>. Post-polypectomy colonoscopy surveillance is an essential component of secondary prevention, particularly in patients with advanced neoplasia, multiple synchronous lesions, or suboptimal baseline examinations<sup>[9,10]</sup>. Despite recommendations, surveillance practices show considerable variability in terms of timing and quality. A key limitation is the interval between procedures, during which precancerous lesions may emerge or progress. Despite the existence of evidence-based guidelines that provide recommendations for surveillance intervals based on individual risk stratification, studies consistently report significant discrepancies between these guidelines and clinical practice, with both overuse and underuse of colonoscopy<sup>[11,12]</sup>. These patterns carry clinical and economic consequences. Overuse exposes patients to unnecessary procedural risks, such as bleeding, perforation, and, rarely, death, and is estimated to cost the U.S. healthcare system up to \$3 billion annually<sup>[9,13-18]</sup>. Conversely, underuse in high-risk individuals may result in missed opportunities for early CRC detection. For instance, only ~42% of patients with high-risk polyps undergo follow-up colonoscopy within the recommended 3-year intervals<sup>[15,19]</sup>. In this context, artificial intelligence (AI) has emerged as a promising innovation to support guideline adherence and personalize surveillance intervals based on patient-specific risk. Encompassing machine learning (ML), deep learning (DL), and natural language processing (NLP), AI enables computers to perform complex cognitive tasks with consistency and precision. In gastrointestinal endoscopy, the most extensively validated AI applications are computer vision systems based on convolutional neural networks (CNNs), which enable real-time polyp detection (computer-aided detection system, CAdE) and characterization (computer-aided diagnosis system, CAdx). These tools have been shown to increase adenoma detection rate (ADR) by 40%-60% relative risk and to reduce miss rate across multiple randomized trials and meta-analyses<sup>[20-22]</sup>. Their ability to consistently identify diminutive and flat lesions addresses one of the major limitations of conventional colonoscopy, namely inter-operator variability. More recently, AI has been used to support clinical decision making by analyzing colonoscopy reports, pathology data, and patient histories to inform optimal surveillance intervals<sup>[22-24]</sup>.

These innovations hold considerable promise for reducing variability in practice, minimizing inappropriate use of resources, and improving the quality and efficiency of post-polypectomy care. This scoping review

aims to synthesize the current literature on the use of AI in post-polypectomy surveillance for CRC, highlight key technological advances, and identify gaps and priorities for future research.

## MATERIALS AND METHODS

### Study design

We conducted a scoping review following the methodological framework proposed by Arksey and O'Malley. The findings were reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, and the corresponding checklist is provided in the [Supplementary Materials \[Supplementary Appendix 1\]](#)<sup>[25-27]</sup>. Given that previous reviews have noted a lack of focused research on the implementation of AI in the context of surveillance colonoscopy, the scoping review approach was deemed appropriate to map the extent, range, and nature of research activity in this emerging and rapidly evolving field.

### Search strategy

The literature search aimed to identify studies reporting on AI models applied to endoscopic surveillance in CRC. A comprehensive search was conducted across multiple databases, including Scopus, EMBASE, PubMed, and Web of Science. Medical Subject Headings (MeSH) terms such as “colorectal cancer”, “artificial intelligence”, “natural language processing”, “deep learning”, “machine learning”, “colonoscopy surveillance”, “colonoscopy surveillance interval”, and “colonoscopy” were utilized in various combinations (see [Supplementary Appendix 2](#) for search). The search included articles published until 31<sup>st</sup> May 2025. Additionally, reference lists of identified articles were scrutinized to detect any additional relevant studies. Institutional review board approval was waived as the study used previously published data for analysis.

### Eligibility criteria and study selection

Eligibility criteria were defined using the Population, Concept, and Context (PCC) framework in accordance with the Joanna Briggs Institute's methodology for scoping reviews<sup>[27]</sup>. Studies were eligible for inclusion if they directly investigated the application of AI models to endoscopic surveillance following post-polypectomy colonoscopy, specifically for CRC prevention. In line with the Effective Practice and Organization of Care (EPOC) guidance, we considered a wide range of study designs, including randomized controlled trials, controlled clinical trials, prospective and retrospective observational studies, cohort studies, population-based studies, cross-sectional and case-control studies, as well as reviews, editorials, invited commentaries, and case reports. No language or publication date restrictions were applied.

During screening, studies were excluded if they focused exclusively on real-time polyp detection or histopathology without reference to surveillance or if they investigated AI applications in colonoscopy but did not report surveillance-related outcomes (e.g., models predicting ADR, bowel preparation quality, or cecal intubation time).

### Data extraction, synthesis, and analysis

Study selection and reporting followed the PRISMA-ScR (2018) guidelines. Screening was conducted according to predefined inclusion and exclusion criteria. Two independent reviewers (Ferrari S and Negro S) performed the literature screening and study selection using Rayyan®, with disagreements resolved by consensus or by consultation with two additional reviewers (Spolverato G and Pulvirenti A). A standardized data charting form was developed and piloted by the review team to ensure consistent and comprehensive extraction of relevant information from each included study. Extracted variables included: authorship, year of publication, country, study setting, target population, AI methodology (e.g., ML, DL, NLP), data sources [e.g., colonoscopy reports, pathology, electronic health records (EHRs)], type of surveillance application (e.g., interval recommendation, risk stratification), and key outcomes reported. A dedicated database was

used to organize and manage the extracted data. Any discrepancies in data charting were resolved through discussion among the reviewers to ensure accuracy and consistency. Data synthesis was conducted using a descriptive and thematic approach. Results were synthesized narratively and summarized in structured tables. The studies were grouped based on the type of AI application and the clinical purpose, allowing for comparison across different methodological approaches and healthcare contexts. In line with the objectives of a scoping review, no formal critical appraisal of methodological quality or risk of bias was performed.

## RESULTS

The literature search identified a total of 950 articles. After removing duplicates, 946 titles and abstracts were screened, resulting in 223 full-text articles assessed for eligibility. Of these, 216 were excluded for the following reasons. Many studies ( $n = 101$ ) evaluated AI tools for real-time polyp detection or histopathology, without addressing post-polypectomy surveillance. A substantial number ( $n = 88$ ) reported AI applications in colonoscopy but did not provide surveillance-related outcomes. The other studies were excluded for low methodological quality ( $n = 19$ ) and redundant data ( $n = 8$ ). Ultimately, seven studies met all predefined inclusion criteria and were included in the final review. The study selection process is illustrated in [Figure 1](#).

The majority of the included studies were published within the last five years (2020-2025), reflecting the recent and growing interest in this area. Most of the studies primarily focused on the application of AI models to support the determination of appropriate surveillance intervals following post-polypectomy colonoscopy. Notably, all of the included studies focused on the use of NLP techniques or large language models (LLMs). A summary of the key characteristics and findings of the included studies is provided in [Table 1](#).

### Natural language processing-based AI models for surveillance colonoscopy

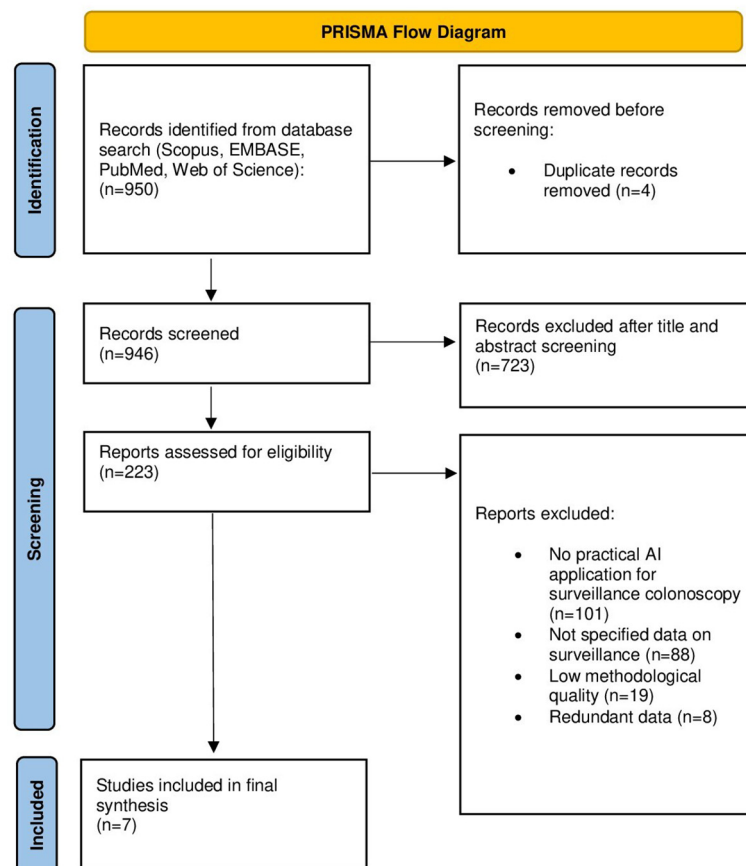
Five of the included studies applied NLP-based AI models to optimize post-polypectomy surveillance<sup>[28-32]</sup>. The methodological pipeline is illustrated in [Figure 2](#).

Imler *et al.*<sup>[28]</sup> were pioneers in this field, developing an NLP-enhanced Clinical Decision Support (CDS) system designed to automatically generate surveillance recommendations based on 2012 United States Multi-Society Task Force (USMSTF) guidelines, which offer standardized, evidence-based protocols for surveillance based on polyp characteristics and procedural quality<sup>[35]</sup>. In a retrospective analysis of 10,789 colonoscopy reports, including 6,379 linked to corresponding pathology reports, the authors utilized an open-source NLP platform (Apache cTAKES, Clinical Text Analysis and Knowledge Extraction System), which employs named entity recognition (NER) to map clinical concepts to the Unified Medical Language System Metathesaurus. The NLP pipeline also included negation detection and context analysis to differentiate true findings from negated or uncertain mentions, thereby improving robustness beyond simple keyword-based text mining. Extracted features were then fed into a rule-based CDS algorithm aligned with USMSTF surveillance intervals. The system's performance was validated against a manually annotated gold standard created from a random subset of 300 cases, reviewed by expert clinicians. The NLP-CDS tool demonstrated 81.7% concordance with expert adjudication (Cohen's  $\kappa = 0.74$ ;  $P < 0.001$ ) and showed excellent accuracy in assigning 10-year surveillance intervals, with a 99.1% agreement rate. Performance varied across different intervals, with agreement rates of 85.7% for 5-10 years, 74.3% for 3 years, and 50% for 1-3 years. Karwa *et al.*<sup>[29]</sup> advanced the integration of AI into post-polypectomy surveillance by developing an NLP-based CDS system capable of synthesizing both structured and unstructured data from electronic medical records (EMRs). The NLP pipeline was implemented in Prolog, a general-purpose programming language, and processed colonoscopy and pathology reports through

**Table 1. Summary of studies included in the evaluation of AI-based tools for post-polypectomy surveillance**

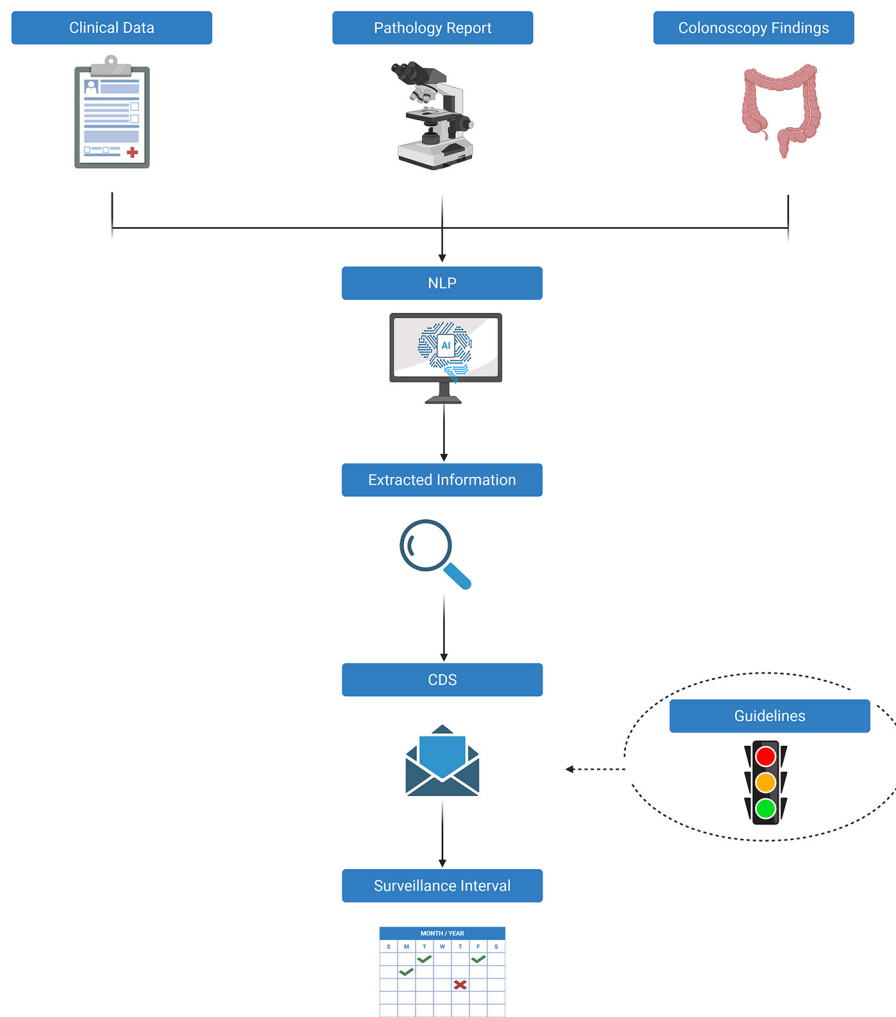
| Study (year)                                  | AI approach         | Data input                                                                                                       | Output                                                                        |
|-----------------------------------------------|---------------------|------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Imler <i>et al.</i> (2014) <sup>[28]</sup>    | NLP + CDS           | Colonoscopy reports, pathology reports (polyp size, location, histology)                                         | Identification of appropriate surveillance intervals                          |
| Karwa <i>et al.</i> (2020) <sup>[29]</sup>    | NLP + CDS           | Patient variables, endoscopic procedure data, pathological findings from EMRs                                    | Automated surveillance interval recommendations                               |
| Peterson <i>et al.</i> (2021) <sup>[30]</sup> | NLP pipeline        | Colonoscopy and pathology reports (polyp characteristics such as location, size, histology)                      | Polyp classification and follow-up determination                              |
| Bae <i>et al.</i> (2022) <sup>[31]</sup>      | NLP pipeline        | Colonoscopy and pathology reports (polyp type, location, size; free-text in English, Korean, mixed formats)      | Automated extraction of quality indicators (ADR, SDR, surveillance intervals) |
| Wu <i>et al.</i> (2023) <sup>[32]</sup>       | NLP + deep learning | Endoscopic and pathological reports from three hospitals (large multicenter dataset)                             | Risk stratification and decision support                                      |
| Lim <i>et al.</i> (2024) <sup>[33]</sup>      | ChatGPT-4           | Clinical scenarios with patient history, colonoscopy findings, pathology reports                                 | Evaluation of LLM utility for surveillance recommendations                    |
| Chang <i>et al.</i> (2024) <sup>[34]</sup>    | ChatGPT-4           | De-identified clinical data: history of present illness, colonoscopy findings, pathology reports, family history | Simulation of clinical decision making in surveillance                        |

AI: Artificial intelligence; NLP: Natural language processing; CDS: clinical decision support; EMRs: electronic medical records; ADR: adenoma detection rate; SDR: sessile serrated lesion detection rate; LLM: large language model (e.g., ChatGPT-4); ChatGPT-4: Chat Generative Pre-trained Transformer 4.



**Figure 1.** Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart of the included studies. Created in Covidence (<https://www.covidence.org>).

tokenization of paragraphs into sentences and words. Each sentence was parsed to identify polyp findings, which were then tabulated (none or one-to-many per procedure). Polyp locations were normalized to



**Figure 2.** Natural language processing (NLP)-based AI models for surveillance colonoscopy. NLP: Natural language processing; CDS: clinical decision support; AI: artificial intelligence. Created in BioRender. Ferrari S (2025) <https://BioRender.com/lvprlw9>.

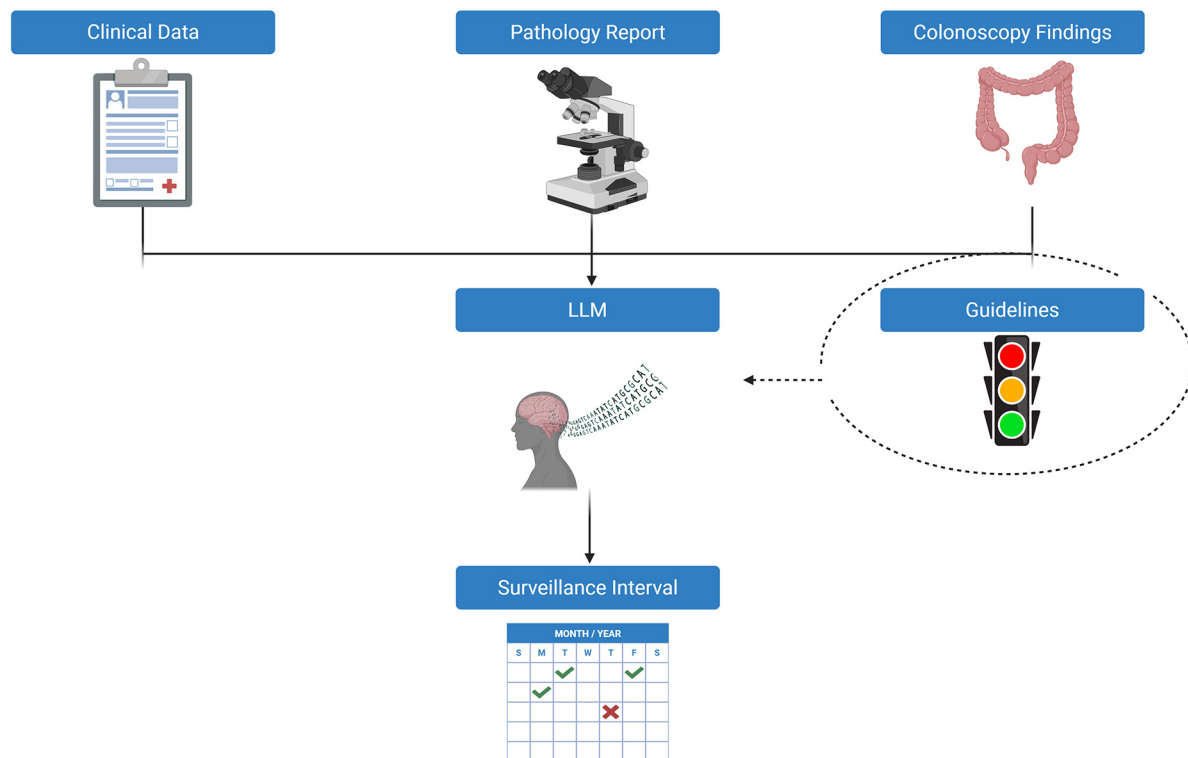
standardized colonic segments to ensure accurate linkage with pathology reports, where histological specimens were tagged by site. A complementary pathology parser extracted semi-structured diagnostic data, and mapped pathology results back to each polyp by location. Extracted features were fed into a rule-based CDS algorithm aligned with the 2020 USMSTF guidelines<sup>[36]</sup>. In performance testing, the system achieved 100% concordance with guideline-based recommendations in a development cohort of 1,961 cases and 99% concordance in a separate validation cohort of 263 cases, indicating robust generalizability across datasets. In contrast, surveillance intervals determined by endoscopists aligned with guidelines in only 62% of cases. Importantly, 94% of the discordant clinical decisions involved premature surveillance scheduling, which may expose patients to unnecessary procedural risks and contribute to healthcare overutilization. Peterson *et al.*<sup>[30]</sup> developed an automated pipeline using NLP to extract, synthesize, and interpret critical endoscopic and pathological variables for the purpose of assigning post-polypectomy surveillance intervals. The system was specifically designed to capture polyp-level details, such as size, location, and histologic subtype, from colonoscopy and pathology reports, aggregate them at the patient level, and classify patients into one of six risk-based categories corresponding to the USMSTF surveillance intervals<sup>[36]</sup>. The pipeline was trained on a cohort of 324 patients (546 reports) and evaluated on an independent test set of 100 patients. The NLP modules were validated against a manually curated gold standard created by chart

review, with performance assessed at both the variable-extraction level and the final surveillance interval assignment. Module-level validation used metrics such as positive predictive value (PPV), sensitivity, and F1-score. Performance was consistently high across these metrics, with PPV ranging from 0.91 to 1.00, sensitivity from 0.85 to 1.00, and F1-scores between 0.88 and 1.00. Overall, the model correctly assigned guideline-concordant surveillance intervals in 92% of cases, demonstrating its strong potential for scalable, automated implementation in clinical workflows. Bae *et al.*<sup>[31]</sup> developed an NLP pipeline to automatically assign post-colonoscopy surveillance intervals, based on colonoscopy reports matched with pathology data, following the 2020 USMSTF guidelines<sup>[36]</sup>. The pipeline was implemented using Python regular expressions and smartTA, incorporating a custom Korean-English lexicon of medical terms and endoscopic abbreviations, with extraction rules refined interactively based on input from gastroenterologists. Reports underwent processing to link biopsy phrases in colonoscopy notes with specimen-level pathology diagnoses, followed by information extraction of key variables (polyp size, location, histology) and summarization for surveillance assignment. A development dataset of 3,000 colonoscopies (paired with 2,168 pathology reports) was randomly selected and divided into a 2,000-report training set and a 1,000-report testing set, with manual annotation by five gastroenterologists providing a reference standard. In this testing dataset, the NLP system and manual review showed similar performance, with NLP achieving higher accuracy for the 3-year (100% *vs.* 93.6%) and 3-5-year (98.6% *vs.* 94.2%) intervals. When applied to 54,562 screening and surveillance colonoscopies, endoscopists with a lower ADR (< 30%) assigned a larger proportion of patients to the longest 10-year interval (77.8%) compared to endoscopists with an ADR of 46.1%. Wu *et al.*<sup>[32]</sup> proposed an advanced AI-based Automatic Surveillance (AS) system that combines NLP with DL techniques to assign post-colonoscopy surveillance intervals in accordance with five major international guidelines, including those from the USMSTF<sup>[9,36-39]</sup>. The system architecture consisted of four models: (1) a regular expression-based classifier to identify patients after polypectomy, (2) a universal information extraction (UIE) model that applied a text-to-structure framework for parsing endoscopic and pathological reports, (3) a rule-based classifier to stratify patients into guideline-defined risk categories, and (4) a text-to-speech and automated calling module for patient notifications and follow-up. The UIE model represented the core DL component, enabling the extraction and alignment of polyp-level variables (size, histology, and location) with pathology results. The system was trained and validated using an extensive multicenter dataset comprising 47,544 endoscopic and pathological reports from three tertiary care hospitals, reflecting a wide range of real-world clinical scenarios. The AS system demonstrated exceptional performance across multiple validation sets. In the internal test cohort ( $n = 9,583$ ), the model achieved an accuracy of 99.30% [95%confidence interval (CI), 98.67%-99.63%]. Similar levels of performance were observed in two external validation cohorts: 98.89% accuracy (95%CI, 98.33%-99.27%) in a 4,792-patient set, and 98.56% accuracy (95%CI, 95.86%-99.51%) in a 1,731-patient set. Notably, when benchmarked against clinical practice, the AI system vastly outperformed human decision making: in a comparative analysis, the AS tool correctly assigned guideline-concordant intervals in 99.9% of cases, compared to only 15.7% accuracy in routine physician assessments. Moreover, in a prospective multireader, multicase study, the AS system significantly improved endoscopists' adherence to surveillance guidelines (from 78.1% to 98.7%) and substantially reduced the burden of manual review.

### Large language models for post-colonoscopy surveillance interval assignment

Two studies explored the use of LLMs, specifically ChatGPT-4, to support decision making in the assignment of post-polypectomy surveillance intervals. The methodological workflow is illustrated in Figure 3.

Lim *et al.*<sup>[33]</sup> conducted a comparative evaluation of two ChatGPT configurations: a contextualized version, fine-tuned with U.S. CRC surveillance guidelines and supplementary medical resources, and the standard base model without domain-specific adaptation. Both models were tested using a set of 62 simulated clinical



**Figure 3.** Large language models (LLMs) for post-colonoscopy surveillance interval assignment. LLM: Large language models. Created in BioRender. Ferrari S (2025) <https://BioRender.com/kcdbftw>.

scenarios that mirrored real-world post-colonoscopy surveillance decision points. The contextualized model achieved a guideline-concordant recommendation rate of 79%, significantly outperforming the standard model, which reached only 50.5% accuracy. However, both models demonstrated reduced performance in complex cases requiring surveillance after abnormal findings: the contextualized model yielded 63% accuracy, compared to 37% for the standard version.

Chang *et al.*<sup>[34]</sup> further investigated the real-world applicability of ChatGPT-4 by retrospectively applying it to a cohort of 505 patients, using de-identified clinical inputs that included history of present illness, colonoscopy and pathology findings, and family history. ChatGPT-4's interval recommendations were compared with those derived from the 2020 USMSTF guidelines and with those documented by practicing gastroenterologists in routine clinical practice. The model demonstrated high concordance with guideline-based recommendations in 85.7% of cases, significantly exceeding the 75.4% concordance rate achieved by clinicians ( $P < 0.001$ ). A summary of guideline discordance rates reported in the included studies is provided in Table 2.

## DISCUSSION

This scoping review identified and synthesized the current evidence on the application of AI, particularly NLP and LLMs, in guiding surveillance colonoscopy intervals following polypectomy. Seven studies met the inclusion criteria overall, highlighting a nascent but growing body of literature that aims to address persistent gaps between guideline recommendations and clinical practice. Most studies have demonstrated a high level of agreement between AI-generated surveillance recommendations and established guidelines, suggesting that AI tools may offer substantial support in standardizing post-polypectomy follow-up and

**Table 2. Concordance and discordance rates with post-polypectomy surveillance guidelines: comparison between physicians and AI-based systems**

| Study (year)                           | Guideline(s)                                | Discordance (AI)                                  | Discordance (physicians) |
|----------------------------------------|---------------------------------------------|---------------------------------------------------|--------------------------|
| Imler et al. (2014) <sup>[28]</sup>    | USMSTF                                      | 18.3%                                             | 18.3%                    |
| Karwa et al. (2020) <sup>[29]</sup>    | USMSTF                                      | 0%-1%                                             | 31%-38%                  |
| Peterson et al. (2021) <sup>[30]</sup> | USMSTF                                      | 8%                                                | Not reported             |
| Bae et al. (2022) <sup>[31]</sup>      | USMSTF                                      | 0%-1.4%                                           | 34.3%                    |
| Wu et al. (2023) <sup>[32]</sup>       | USMSTF, BSG, ESGE, Asia-Pacific, Chinese    | 0.1%-1.4%                                         | > 50%                    |
| Lim et al. (2024) <sup>[33]</sup>      | U.S. guidelines (USMSTF, ACG, ASGE, USPSTF) | 21% (contextualized GPT-4) vs. 49.5% (base GPT-4) | Not reported             |
| Chang et al. (2024) <sup>[34]</sup>    | USMSTF                                      | 14.3%                                             | 24.6%                    |

AI: Artificial intelligence; USMSTF: United States Multi-Society Task Force on Colorectal Cancer; ESGE: European Society of Gastrointestinal Endoscopy; BSG: British Society of Gastroenterology; ACG: American College of Gastroenterology; ASGE: American Society for Gastrointestinal Endoscopy; USPSTF: United States Preventive Services Task Force; GPT-4: Generative Pre-trained Transformer 4.

mitigating inappropriate utilization of colonoscopy. Key differences between the two models are summarized in [Table 3](#).

### NLP-based decision support: promise and maturity

NLP is a promising interdisciplinary field that combines computer science, AI, and computational linguistics. It focuses on enabling computers to understand human language in unstructured text. In healthcare, NLP shows great potential for addressing information overload by facilitating the aggregation and summarization of clinical notes, extracting key data from discharge summaries, analyzing treatment patterns, and interpreting patient queries semantically. NLP is also increasingly used to support medical decision making by identifying patterns and anomalies in large text corpora and generating recommendations aligned with experts' opinions<sup>[40,41]</sup>. NLP in healthcare plays a pivotal role in extracting, structuring, and analyzing different clinical text data. It enhances decision support systems, improves care delivery, and promotes evidence-based practice<sup>[42-44]</sup>. Of the studies reviewed, five incorporated NLP-augmented decision support systems into clinical workflows. These systems were trained using structured and unstructured data from colonoscopy and pathology reports. Overall, the systems demonstrated high accuracy in determining guideline-concordant surveillance intervals. Notably, Wu *et al.*<sup>[32]</sup> reported accuracy rates exceeding 98% across internal and external validation cohorts. This performance markedly outperformed physician recommendations, which were guideline-concordant in only 15.7% of cases. These results highlight the technical maturity and clinical usefulness of NLP-based tools, especially their ability to reduce inter-clinician variability and avoid unnecessary procedures. While these results are impressive, the reported 99.9% concordance should be interpreted with caution, as the validation relied primarily on simple percent agreement with guideline-derived standards. The absence of complementary measures such as Cohen's kappa or confidence intervals limits the robustness of the statistical evaluation and may overestimate performance. Nevertheless, these findings provide a valuable proof of concept, while current guidelines still present structural gaps. A major limitation of the current surveillance guidelines is their failure to account for variability in endoscopist performance. The current recommendations are based exclusively on polyp characteristics and do not consider individual ADRs. According to the European Society of Gastrointestinal Endoscopy (ESGE) guidelines, ADR rates of 40%-45% are preferable<sup>[45]</sup>. This gap can result in suboptimal surveillance recommendations, especially when colonoscopies are performed by low-performing endoscopists. Recent data illustrate this discrepancy. Endoscopists with high ADRs (greater than 45%) recommended 10-year surveillance intervals in 46.1% of cases (6,397 out of 13,883), compared to 77.8% (2,231 out of 2,873) among those with ADRs below 30%. This variance suggests that lower-

**Table 3. Comparison between natural language processing (NLP) and large language models (LLMs)**

| Feature                 | NLP                                                                          | LLMs                                                                                            |
|-------------------------|------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| <b>Definition</b>       | Traditional AI techniques used to process and analyze human language in text | Advanced deep learning models trained on large text corpora to understand and generate language |
| <b>Architecture</b>     | Often rule-based or uses classic machine learning                            | Based on transformer architectures                                                              |
| <b>Training data</b>    | Requires task-specific annotated datasets                                    | Trained on billions of words from general and domain-specific sources                           |
| <b>Flexibility</b>      | Built for specific tasks (e.g., entity recognition, parsing)                 | Capable of performing multiple tasks with minimal customization                                 |
| <b>Ease of use</b>      | Requires expert design and manual tuning                                     | User-friendly; responds to prompts with minimal input                                           |
| <b>Interpretability</b> | More interpretable due to modular design                                     | Less interpretable; functions as a black box                                                    |
| <b>Limitations</b>      | Less effective in complex language tasks; needs expert maintenance           | Can generate inaccurate or biased output; still under clinical validation                       |

AI: Artificial intelligence.

performing endoscopists may miss lesions, underestimate future risk, and recommend unduly extended surveillance intervals. Therefore, regular assessment of individual performance metrics, such as ADR, is essential, and surveillance recommendations must be calibrated accordingly<sup>[46,47]</sup>. In addition to Western recommendations, the Asia-Pacific consensus provides region-specific strategies that broaden the global relevance of surveillance guidelines<sup>[39]</sup>. These recommendations place particular emphasis on the number and size of adenomas, as well as on serrated histology, often suggesting shorter surveillance intervals than those recommended by USMSTF or ESGE. For example, while the USMSTF recommends a 7-10-year interval after removal of 1-2 small adenomas, the Asia-Pacific guidelines propose a 5-year follow-up in the same setting, whereas the ESGE does not recommend surveillance at all. This heterogeneity across international guidelines further illustrates the complexity of defining optimal surveillance strategies and underscores the need for more flexible approaches across diverse practice settings. In this context, NLP-based systems can play a transformative role. With pooled precision rates as high as 99.7% for extracting colonoscopy quality indicators, NLP enables continuous quality monitoring and risk-adjusted decision making<sup>[48]</sup>. These systems can operate in real time or retrospectively to identify instances in which the endoscopist's performance differs from surveillance recommendations. This supports more personalized and appropriate follow-up strategies. Furthermore, Karwa *et al.*<sup>[29]</sup> found that most discrepancies between clinician and AI recommendations were caused by premature surveillance. These discrepancies have significant implications for patient safety and healthcare resource utilization. In the United States alone, an estimated two to three million inappropriate colonoscopies are performed annually, contributing to over nine thousand serious adverse events and up to one and a half million minor complications. These include bleeding (up to 36 cases per 10,000 procedures) and perforation (up to 8.5 cases per 10,000 procedures). Elderly populations are particularly vulnerable to post-procedural morbidity and mortality<sup>[49,50]</sup>. Beyond surveillance planning, NLP is being used across different areas of gastrointestinal endoscopy<sup>[51-54]</sup>. In administrative and operational domains, NLP automates the coding and classification of procedures. This improves documentation consistency and reduces reliance on manual abstraction. These improvements support accurate billing, quality reporting, and scalable analytics. For instance, NLP systems can reliably distinguish between screening and diagnostic colonoscopy indications, which is essential for reimbursement and benchmarking. Additionally, NLP enhances data accessibility for population health initiatives and clinical research by structuring free-text data within EHRs<sup>[54,55]</sup>. In chronic disease management, particularly for inflammatory bowel disease (IBD), NLP can track patients' progress over time and identify those eligible for dysplasia screening using unstructured clinical narratives. These capabilities demonstrate the expanding scope of NLP applications and highlight its value as a technical and clinical enabler in modern gastroenterological practice<sup>[56-58]</sup>.

### Emerging role of large language models

In parallel with advances in traditional NLP techniques, LLMs have undergone rapid development, demonstrating substantial potential in processing unstructured textual data and supporting clinical decision making. Among these, ChatGPT, developed by OpenAI, stands out as a leading example. Based on transformer architecture, ChatGPT can interpret, generate, and contextualize human-like text. Its most recent version, ChatGPT-4, introduces multimodal capabilities, allowing the model to process and reason over both textual and visual inputs<sup>[59,60]</sup>. Within the field of gastroenterology, LLMs have been applied across various domains. ChatGPT has shown the ability to accurately respond to common patient questions regarding colonoscopy and to generate guideline-consistent recommendations for IBD management, in alignment with European Crohn's and Colitis Organization (ECCO) guidelines<sup>[61-64]</sup>. The model has also been evaluated in more acute care settings, such as assessing the clinical severity of ulcerative colitis presentations in emergency departments<sup>[65]</sup>. Furthermore, ChatGPT has demonstrated high accuracy in providing management recommendations for gastroesophageal reflux disease (GERD), with over 90% of its outputs deemed clinically appropriate<sup>[66]</sup>. Despite these promising applications, the integration of LLMs into image-based diagnostic workflows has produced mixed results. Current performance in visual interpretation tasks remains limited, with accuracy rates often falling below 60% in complex imaging scenarios<sup>[67,68]</sup>. Nevertheless, studies by Lim *et al.*<sup>[33]</sup> and Chang *et al.*<sup>[34]</sup> underscore the model's potential when applied in structured or domain-specific contexts. ChatGPT-4, when fine-tuned or augmented with structured inputs, achieved up to 85.7% concordance with USMSTF surveillance guidelines, surpassing the accuracy of practicing gastroenterologists. However, its performance was less reliable in complex clinical cases, particularly those involving abnormal or ambiguous findings. These findings suggest that while LLMs may serve as effective decision-support tools in routine scenarios, further refinement and clinical fine-tuning are necessary before deployment in high-risk or autonomous settings. A particularly promising avenue for future research lies in the development of LLMs as multimodal clinical assistants capable of simultaneously processing textual and imaging data. A recent exploratory study evaluated the feasibility of using ChatGPT-4 to interpret images from a variety of endoscopic modalities, including capsule endoscopy, device-assisted enteroscopy, and endoscopic ultrasound, highlighting both the opportunities and the current technical limitations of such applications<sup>[69]</sup>. The introduction of LLMs with integrated vision capabilities represents a potential inflection point in the application of AI within gastroenterology. Rather than replacing established CNN-based systems, these models may enhance existing technologies by bridging the gap between language and image data. This integration opens the possibility for truly multimodal, patient-specific clinical decision support systems that can contextualize findings within a broader clinical narrative, thus improving both diagnostic precision and care personalization.

### Implementation challenges: privacy and medicolegal implications

Beyond technical performance, the real-world adoption of AI-driven surveillance systems requires addressing critical challenges in data privacy, computational resources, and medicolegal accountability. Data privacy is a central concern, given the sensitivity of endoscopic and pathology records and the variability of regulatory frameworks across jurisdictions. While U.S. regulations emphasize compliance with the Health Insurance Portability and Accountability Act (HIPAA), the European Union enforces stricter standards under the General Data Protection Regulation (GDPR), and Asia-Pacific countries follow heterogeneous policies, complicating multicenter deployment<sup>[70-72]</sup>. Advanced approaches such as homomorphic encryption, federated learning, and decentralized data sharing have been proposed to mitigate these risks, but their clinical implementation remains limited<sup>[71,72]</sup>. Equally important are the unresolved medicolegal implications of AI-guided surveillance. Current evidence supports AI as a "decision support" rather than "decision maker", with physicians retaining ultimate accountability<sup>[71,73]</sup>. However, the absence of a clear framework delineating liability, transparency, and data stewardship may hinder large-scale adoption. Addressing these gaps is essential to ensure that AI-enabled surveillance systems are

implemented safely, equitably, and in compliance with diverse healthcare regulations.

### **Limitations of the current evidence**

Despite the encouraging findings, the current evidence base is limited by the small number of available studies and the heterogeneity of their study designs. In many cases, there is also a lack of external validation. Most models were developed and tested within single health systems or academic centers, raising concerns about generalizability. Moreover, none of the studies reported real-time prospective implementation or patient-centered outcomes such as safety, satisfaction, or CRC incidence reduction. Finally, while NLP-based systems are more mature, LLM applications remain largely experimental and require robust regulatory, ethical, and clinical oversight prior to large-scale deployment.

### **Implications for future research**

Future studies should aim to evaluate AI systems in prospective, multicenter clinical settings, and explore integration into existing EHRs and colonoscopy reporting platforms. Research should also assess AI's impact on long-term outcomes, such as interval cancer rates, healthcare utilization, and cost-effectiveness. In parallel, regulatory frameworks will be essential to ensure algorithmic transparency, accountability, and equity in model performance across diverse populations.

## **CONCLUSION**

In summary, AI-based approaches, particularly NLP and emerging LLMs, show strong potential to improve the accuracy, consistency, and personalization of post-polypectomy surveillance. While promising, further validation in real-world clinical settings is essential before routine implementation. Key research priorities include prospective multicenter evaluations, integration into existing reporting platforms, standardized validation metrics, and regulatory frameworks to ensure safe and equitable translation into practice.

## **DECLARATIONS**

### **Authors' contributions**

Conceptualization: Ferrari S, Spolverato G  
Project administration: Ferrari S, Celotto F, Pulvirenti A, Spolverato G  
Writing - original draft: Ferrari S, Negro S  
Data curation: Ferrari S, Negro S, Celotto F, Madeo G  
Writing - review and editing: Negro S  
Methodology: Celotto F, Madeo G, Pulvirenti A, Spolverato G  
Preparation of figures/images: Bao QR  
Manuscript revision: Bao QR  
Supervision: Pulvirenti A, Spolverato G

### **Availability of data and materials**

All data supporting the findings of this study are contained within the article and its [Supplementary Materials](#).

### **Financial support and sponsorship**

None.

### **Conflicts of interest**

Spolverato G is an Associate Editor of the journal *Artificial Intelligence Surgery*. Spolverato G was not involved in any steps of the editorial process, including, in particular, reviewer selection, manuscript handling, or decision making. The remaining authors declared that there are no conflicts of interest.

**Ethical approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Copyright**

© The Author(s) 2025.

**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;74:229-63. DOI PubMed
2. Bretthauer M, Løberg M, Wieszczyn P, et al; NordICC Study Group. Effect of colonoscopy screening on risks of colorectal cancer and related death. *N Engl J Med.* 2022;387:1547-56. DOI PubMed
3. Brenner H, Heisser T, Cardoso R, Hoffmeister M. Reduction in colorectal cancer incidence by screening endoscopy. *Nat Rev Gastroenterol Hepatol.* 2024;21:125-33. DOI PubMed
4. Zheng S, Schrijvers JJA, Greuter MJW, Kats-Ugurlu G, Lu W, de Bock GH. Effectiveness of colorectal cancer (CRC) screening on all-cause and CRC-specific mortality reduction: a systematic review and meta-analysis. *Cancers.* 2023;15:1948. DOI PubMed PMC
5. Click B, Pinsky PF, Hickey T, Doroudi M, Schoen RE. Association of colonoscopy adenoma findings with long-term colorectal cancer incidence. *JAMA.* 2018;319:2021-31. DOI PubMed PMC
6. Hao Y, Wang Y, Qi M, He X, Zhu Y, Hong J. Risk factors for recurrent colorectal polyps. *Gut Liver.* 2020;14:399-411. DOI PubMed PMC
7. Hassan C, Antonelli G, Dumonceau JM, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - update 2020. *Endoscopy.* 2020;52:687-700. DOI PubMed
8. Issaka RB, Chan AT, Gupta S. AGA clinical practice update on risk stratification for colorectal cancer screening and post-polypectomy surveillance: expert review. *Gastroenterology.* 2023;165:1280-91. DOI PubMed PMC
9. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *N Engl J Med.* 2012;366:687-96. DOI PubMed PMC
10. Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. *N Engl J Med.* 2014;370:2539-41. DOI PubMed
11. Djinbachian R, Dubé AJ, Durand M, et al. Adherence to post-polypectomy surveillance guidelines: a systematic review and meta-analysis. *Endoscopy.* 2019;51:673-83. DOI PubMed
12. Krist AH, Jones RM, Woolf SH, et al. Timing of repeat colonoscopy: disparity between guidelines and endoscopists' recommendation. *Am J Prev Med.* 2007;33:471-8. DOI PubMed
13. Kruse GR, Khan SM, Zaslavsky AM, Ayanian JZ, Sequist TD. Overuse of colonoscopy for colorectal cancer screening and surveillance. *J Gen Intern Med.* 2015;30:277-83. DOI PubMed PMC
14. Fraiman J, Brownlee S, Stoto MA, Lin KW, Huffstetler AN. An estimate of the US rate of overuse of screening colonoscopy: a systematic review. *J Gen Intern Med.* 2022;37:1754-62. DOI PubMed PMC
15. Murphy CC, Sandler RS, Grubber JM, Johnson MR, Fisher DA. Underuse and overuse of colonoscopy for repeat screening and surveillance in the Veterans Health Administration. *Clin Gastroenterol Hepatol.* 2016;14:436-44.e1. DOI PubMed PMC
16. Saini SD, Powell AA, Dominitz JA, et al. Developing and testing an electronic measure of screening colonoscopy overuse in a large integrated healthcare system. *J Gen Intern Med.* 2016;31 Suppl 1:53-60. DOI PubMed PMC
17. Mittal S, Lin YL, Tan A, Kuo YF, El-Serag HB, Goodwin JS. Limited life expectancy among a subgroup of medicare beneficiaries receiving screening colonoscopies. *Clin Gastroenterol Hepatol.* 2014;12:443-50.e1. DOI PubMed PMC
18. Kim SY, Kim HS, Park HJ. Adverse events related to colonoscopy: global trends and future challenges. *World J Gastroenterol.* 2019;25:190-204. DOI PubMed PMC
19. Vader JP, Pache I, Froehlich F, et al. Overuse and underuse of colonoscopy in a European primary care setting. *Gastrointest Endosc.* 2000;52:593-99. DOI PubMed
20. Aziz M, Fatima R, Dong C, Lee-Smith W, Nawras A. The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: a systematic review with meta-analysis. *J Gastroenterol Hepatol.* 2020;35:1676-83. DOI PubMed
21. Huang D, Shen J, Hong J, et al. Effect of artificial intelligence-aided colonoscopy for adenoma and polyp detection: a meta-analysis of randomized clinical trials. *Int J Colorectal Dis.* 2022;37:495-506. DOI PubMed
22. Barua I, Vinsard DG, Jodal HC, et al. Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. *Endoscopy.* 2021;53:277-84. DOI PubMed
23. Taghiakbari M, Mori Y, von Renteln D. Artificial intelligence-assisted colonoscopy: a review of current state of practice and research. *World J Gastroenterol.* 2021;27:8103-22. DOI PubMed PMC
24. Mitsala A, Tsalikidis C, Pitiakoudis M, Simopoulos C, Tsaroucha AK. Artificial intelligence in colorectal cancer screening, diagnosis

- and treatment. A new era. *Curr Oncol*. 2021;28:1581-607. DOI PubMed PMC
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169:467-73. DOI PubMed
  26. Arksey H, O'malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Method*. 2005;8:19-32. DOI
  27. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc*. 2015;13:141-6. DOI PubMed
  28. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol*. 2014;12:1130-6. DOI PubMed
  29. Karwa A, Patell R, Parthasarathy G, Lopez R, McMichael J, Burke CA. Development of an automated algorithm to generate guideline-based recommendations for follow-up colonoscopy. *Clin Gastroenterol Hepatol*. 2020;18:2038-45.e1. DOI PubMed
  30. Peterson E, May FP, Kachikian O, et al. Automated identification and assignment of colonoscopy surveillance recommendations for individuals with colorectal polyps. *Gastrointest Endosc*. 2021;94:978-87. DOI PubMed
  31. Bae JH, Han HW, Yang SY, et al. Natural language processing for assessing quality indicators in free-text colonoscopy and pathology reports: development and usability study. *JMIR Med Inform*. 2022;10:e35257. DOI PubMed PMC
  32. Wu L, Shi C, Li J, et al. Development and evaluation of a surveillance system for follow-up after colorectal polypectomy. *JAMA Netw Open*. 2023;6:e2334822. DOI PubMed PMC
  33. Lim DYZ, Tan YB, Koh JTE, et al. ChatGPT on guidelines: providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals. *J Gastroenterol Hepatol*. 2024;39:81-106. DOI PubMed
  34. Chang PW, Amini MM, Davis RO, et al. ChatGPT4 outperforms endoscopists for determination of postcolonoscopy rescreening and surveillance recommendations. *Clin Gastroenterol Hepatol*. 2024;22:1917-25.e17. DOI PubMed
  35. Lieberman DA, Rex DK, Winawer SJ, Giardiello FM, Johnson DA, Levin TR. Guidelines for colonoscopy surveillance after screening and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology*. 2012;143:844-57. DOI PubMed
  36. Gupta S, Lieberman D, Anderson JC, et al. Recommendations for follow-up after colonoscopy and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer. *Am J Gastroenterol*. 2020;115:415-34. DOI PubMed PMC
  37. Rutter MD, East J, Rees CJ, et al. British Society of Gastroenterology/Association of Coloproctology of Great Britain and Ireland/Public Health England post-polypectomy and post-colorectal cancer resection surveillance guidelines. *Gut*. 2020;69:201-23. DOI PubMed PMC
  38. Zhao S, Wang S, Pan P, et al. Expert consensus on management strategies for precancerous lesions and conditions of colorectal cancer in China. *Chin J Dig Endosc*. 2022;39:1-18. DOI
  39. Saito Y, Oka S, Kawamura T, et al. Colonoscopy screening and surveillance guidelines. *Dig Endosc*. 2021;33:486-519. DOI PubMed
  40. Chen X, Xie H, Wang FL, Liu Z, Xu J, Hao T. A bibliometric analysis of natural language processing in medical research. *BMC Med Inform Decis Mak*. 2018;18:14. DOI PubMed PMC
  41. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. *J Biomed Inform*. 2014;52:112-20. DOI PubMed PMC
  42. Velupillai S, Mowery D, South BR, Kvist M, Dalanian H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform*. 2015;10:183-93. DOI PubMed PMC
  43. Shen F, Liu S, Fu S, et al. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) competition. *JMIR Med Inform*. 2021;9:e24008. DOI PubMed PMC
  44. Kalyan KS, Sangeetha S. SECNLP: a survey of embeddings in clinical natural language processing. *J Biomed Inform*. 2020;101:103323. DOI PubMed
  45. Rembacken B, Hassan C, Riemann JF, et al. Quality in screening colonoscopy: position statement of the European Society of Gastrointestinal Endoscopy (ESGE). *Endoscopy*. 2012;44:957-68. DOI PubMed
  46. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clin Gastroenterol Hepatol*. 2013;11:689-94. DOI PubMed PMC
  47. Raju GS, Lum PJ, Slack RS, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. *Gastrointest Endosc*. 2015;82:512-9. DOI PubMed PMC
  48. IJspeert JE, Bastiaansen BA, van Leerdam ME, et al; Dutch Workgroup serrated polypS & Polyposis (WASP). Development and validation of the WASP classification system for optical diagnosis of adenomas, hyperplastic polyps and sessile serrated adenomas/polyps. *Gut*. 2016;65:963-70. DOI PubMed
  49. Goodwin JS, Singh A, Reddy N, Riall TS, Kuo YF. Overuse of screening colonoscopy in the Medicare population. *Arch Intern Med*. 2011;171:1335-43. DOI PubMed PMC
  50. Brownlee S, Huffstetler AN, Fraiman J, Lin KW. An estimate of preventable harms associated with screening colonoscopy overuse in the U.S. *AJPM Focus*. 2025;4:100296. DOI PubMed PMC
  51. Sabrie N, Khan R, Jogendran R, et al. Performance of natural language processing in identifying adenomas from colonoscopy reports: a systematic review and meta-analysis. *iGIE*. 2023;2:350-6.e7. DOI
  52. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assoc*. 2011;18 Suppl 1:i150-6. DOI PubMed PMC

53. Gawron AJ, Thompson WK, Keswani RN, Rasmussen LV, Kho AN. Anatomic and advanced adenoma detection rates as quality metrics determined via natural language processing. *Am J Gastroenterol*. 2014;109:1844-9. DOI PubMed
54. Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput Biol Med*. 2023;155:106649. DOI PubMed
55. Seong D, Choi YH, Shin SY, Yi BK. Deep learning approach to detection of colonoscopic information from unstructured reports. *BMC Med Inform Decis Mak*. 2023;23:28. DOI PubMed PMC
56. Stidham RW, Yu D, Zhao X, et al. Identifying the presence, activity, and status of extraintestinal manifestations of inflammatory bowel disease using natural language processing of clinical notes. *Inflamm Bowel Dis*. 2023;29:503-10. DOI PubMed
57. Montoto C, Gisbert JP, Guerra I, et al; PREMONITION-CD Study Group. Evaluation of natural language processing for the identification of crohn disease-related variables in Spanish electronic health records: a validation study for the PREMONITION-CD project. *JMIR Med Inform*. 2022;10:e30345. DOI PubMed PMC
58. Gomollón F, Gisbert JP, Guerra I, et al; Premonition-CD Study Group. Clinical characteristics and prognostic factors for Crohn's disease relapses using natural language processing and machine learning: a pilot study. *Eur J Gastroenterol Hepatol*. 2022;34:389-97. DOI PubMed PMC
59. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. *arXiv* 2024; arXiv:2303.08774. Available from <https://doi.org/10.48550/arXiv.2303.08774> [accessed 5 November 2025].
60. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv* 2022; arXiv:2108.07258. Available from <https://doi.org/10.48550/arXiv.2108.07258> [accessed 5 November 2025].
61. Ghersin I, Weisshof R, Koifman E, et al. Comparative evaluation of a language model and human specialists in the application of European guidelines for the management of inflammatory bowel diseases and malignancies. *Endoscopy*. 2024;56:706-9. DOI PubMed
62. Lusetti F, Maimaris S, La Rosa GP, et al. Applications of generative artificial intelligence in inflammatory bowel disease: a systematic review. *Dig Liver Dis*. 2025;57:1883-9. DOI PubMed
63. Laoveeravat P, Simonetto DA. AI (artificial intelligence) as an IA (intelligent assistant): ChatGPT for surveillance colonoscopy questions. *Gastro Hep Adv*. 2023;2:1138-9. DOI PubMed PMC
64. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT answers common patient questions about colonoscopy. *Gastroenterology*. 2023;165:509-11.e7. DOI PubMed
65. Levartovsky A, Ben-Horin S, Kopylov U, Klang E, Barash Y. Towards AI-augmented clinical decision-making: an examination of ChatGPT's utility in acute ulcerative colitis presentations. *Am J Gastroenterol*. 2023;118:2283-9. DOI PubMed
66. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA. Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management. *Am J Gastroenterol*. 2023;118:2276-9. DOI PubMed PMC
67. Sattler SS, Chetla N, Chen M, et al. Evaluating the diagnostic accuracy of ChatGPT-4 Omni and ChatGPT-4 Turbo in identifying melanoma: comparative study. *JMIR Dermatol*. 2025;8:e67551. DOI PubMed PMC
68. Saraiva MM, Ribeiro T, Agudo B, et al. Evaluating ChatGPT-4 for the interpretation of images from several diagnostic techniques in gastroenterology. *J Clin Med*. 2025;14:572. DOI PubMed PMC
69. Qin Y, Chang J, Li L, Wu M. Enhancing gastroenterology with multimodal learning: the role of large language model chatbots in digestive endoscopy. *Front Med*. 2025;12:1583514. DOI PubMed PMC
70. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: artificial intelligence in healthcare. *Cureus*. 2023;15:e43262. DOI PubMed PMC
71. Elendu C, Amaechi DC, Elendu TC, et al. Ethical implications of AI and robotics in healthcare: a review. *Medicine*. 2023;102:e36671. DOI PubMed PMC
72. Chen Y, Esmailzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. *J Med Internet Res*. 2024;26:e53008. DOI PubMed PMC
73. Alowais SA, Alghamdi SS, Alsuhbany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023;23:689. DOI PubMed PMC