frontiers | Frontiers in Medicine

TYPE Systematic Review
PUBLISHED 22 January 2025
DOI 10.3389/fmed.2024.1512824

OPEN ACCESS

EDITED BY
Raffaele Pellegrino,
University of Campania Luigi Vanvitelli, Italy

REVIEWED BY
Sheng-Chieh Lu,
University of Texas MD Anderson Cancer Center, United States
Angus Roberts,
King's College London, United Kingdom

*CORRESPONDENCE
Mahmud Omar
✉ Mahmudomar70@gmail.com

RECEIVED 17 October 2024
ACCEPTED 09 December 2024
PUBLISHED 22 January 2025

CITATION
Omar M, Nassar S, SharIf K, Glicksberg BS, Nadkarni GN and Klang E (2025) Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review.
Front. Med. 11:1512824.
doi: 10.3389/fmed.2024.1512824

Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review

Mahmud Omar[1,2]*, Salih Nassar[3], Kassem SharIf[4], Benjamin S. Glicksberg[2], Girish N. Nadkarni[2] and Eyal Klang[2]

[1]Maccabi Health Services, Tel Aviv, Israel, [2]Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY, United States, [3]Edith Wolfson Medical Center, Holon, Israel, [4]Department of Gastroenterology, Sheba Medical Center, Tel HaShomer, Israel

Background and aim: In the last years, natural language processing (NLP) has transformed significantly with the introduction of large language models (LLM). This review updates on NLP and LLM applications and challenges in gastroenterology and hepatology.

Methods: Registered with PROSPERO (CRD42024542275) and adhering to PRISMA guidelines, we searched six databases for relevant studies published from 2003 to 2024, ultimately including 57 studies.

Results: Our review of 57 studies notes an increase in relevant publications in 2023–2024 compared to previous years, reflecting growing interest in newer models such as GPT-3 and GPT-4. The results demonstrate that NLP models have enhanced data extraction from electronic health records and other unstructured medical data sources. Key findings include high precision in identifying disease characteristics from unstructured reports and ongoing improvement in clinical decision-making. Risk of bias assessments using ROBINS-I, QUADAS-2, and PROBAST tools confirmed the methodological robustness of the included studies.

Conclusion: NLP and LLMs can enhance diagnosis and treatment in gastroenterology and hepatology. They enable extraction of data from unstructured medical records, such as endoscopy reports and patient notes, and for enhancing clinical decision-making. Despite these advancements, integrating these tools into routine practice is still challenging. Future work should prospectively demonstrate real-world value.

KEYWORDS

natural language processing, large language models, gastroenterology, hepatology, electronic health records

Introduction

Recent advances in Natural Language Processing (NLP) show potential for being integrated in the field of gastroenterology and hepatology (1, 2). Since the last review in 2014 by Hou et al.—which highlighted NLP's growing utility in gastroenterology, particularly for extracting structured data from colonoscopy and pathology reports to track quality metrics and improve disease detection (2)—the field has evolved considerably. The earlier work by Hou et al. demonstrated promising performance in relatively focused domains, such as colonoscopy quality measure extraction and improving case-finding for inflammatory bowel disease, yet it largely described proof-of-concept implementations and noted challenges with integration into routine clinical workflows and data heterogeneity across settings.

In contrast, significant strides in technology, including the advent of Large Language Models (LLMs) such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) (3), have expanded the scope of NLP applications. While Hou et al.'s era of NLP research centered on rule-based or traditional machine learning methods optimized for specific tasks, newer LLMs can handle a broader range of complex and context-rich functions, from automating routine documentation tasks to supporting sophisticated diagnostic reasoning and therapeutic decision-making (4). These contemporary models may better address scalability and integration challenges, moving beyond static data extraction toward dynamic interactions with unstructured clinical narratives.

NLP and LLMs extract and interpret data from patient records, notes, and reports (5–7). In gastroenterology and hepatology, they streamline the review of endoscopy, radiology, and pathology reports. This technology can help create research cohorts for clinical trials, flag complications, and support decision-making systems. Examples include managing complex conditions like IBD and hepatocellular carcinoma (5, 7, 8).

This review discusses the current applications and challenges of NLP and LLMs in gastroenterology and hepatology.

Methods

Registration and protocol

This systematic literature review was registered with the International Prospective Register of Systematic Reviews, PROSPERO, under the registration code CRD42024542275 (9). Our methodology adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (10).

Search strategy

We conducted a systematic search of six key databases (PubMed, Embase, Web of Science, and Scopus, Cochrane library and IEEE Xplore) for studies published until April 2024. Our focus was on the outcomes of integrating NLP and LLM models in gastroenterology and hepatology. We designed Boolean search strings tailored to each database. To maximize coverage, we supplemented our search with a manual reference screening of included studies and targeted searches on Google Scholar. Details of the specific Boolean strings used are provided in the Supplementary materials.

Study screening and selection

Our review encompasses original research articles, and full conference papers (11). The exclusion criteria were confined to preprints, review papers, case reports, commentaries, protocol studies, editorials, and non-English publications. For the initial screening, we used the Rayyan web application (12). The initial screening and study selection, which were conducted according to predefined criteria, were independently performed by two reviewers (MO and EK). Discrepancies were resolved through discussion. Fleiss' kappa was calculated for the agreement between the two independent reviewers.

Data extraction

Data extraction was conducted by researchers MO and EK using a standardized form to ensure consistent and accurate data capture. This included details such as author, publication year, sample size, data type, task type, specific field, model used, results, numeric metrics, conclusions, and limitations. Any discrepancies in data extraction were resolved through discussion and a third reviewer was consulted when necessary.

Risk of bias assessment

To ensure a thorough evaluation of the included studies, we used three tools, each tailored to a specific study design within our review. The Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool has been employed in interventional studies assessing NLP in applications such as management, prescription guidance, and clinical inquiry responses (13). For diagnostic studies where NLP models were compared with physicians or a reference standard for diagnosing and detection, the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool was used (14). Finally, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) tool was utilized for the remaining studies, which involved NLP models prediction, without direct comparison to reference standards (15). This multitool approach allowed us to appropriately address the diverse methodologies and applications presented in the reviewed studies.

Results

Search results and study selection

A total of 720 articles were identified through initial screening. After the removal of 114 duplicates, 606 articles remained for further evaluation. Title and abstract screening led to the exclusion of 524 articles, leaving 82 articles for full-text review. Of these, the reasons for exclusion and the number of articles excluded for each reason remain the same as described earlier. Ultimately, 55 studies met all inclusion criteria. By employing reference checking and snowballing techniques, two additional studies were identified, resulting in a final tally of 57 studies (16–72). A PRISMA flowchart visually represents the screening process in Figure 1. Fleiss' kappa for the agreement between screeners was calculated as 0.957, which is considered very high (73).

An overview of the included studies

Our systematic review incorporates a total of 57 studies (16–72). Among these, a substantial majority, 49 studies, are centered on

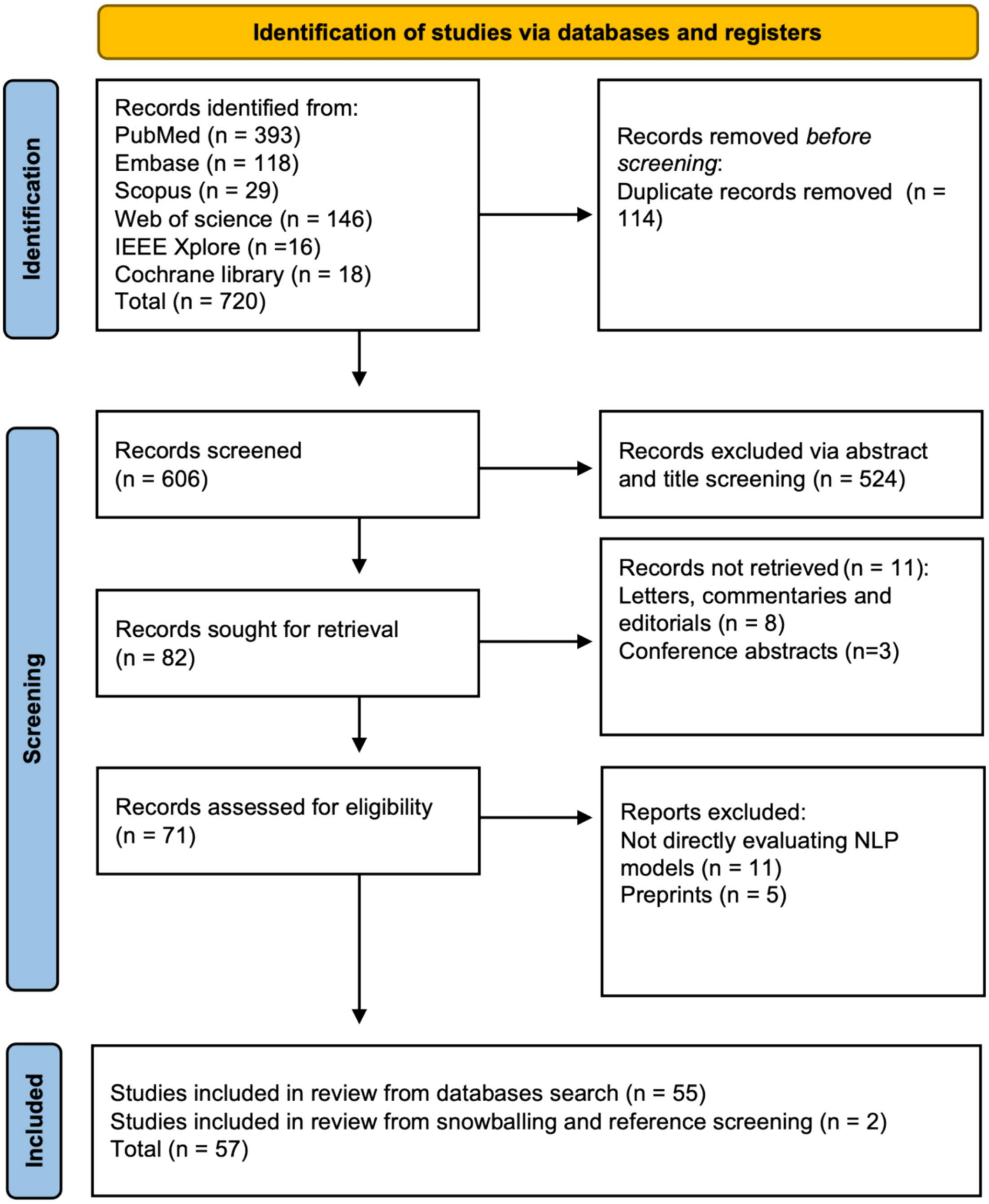

Identification of studies via databases and registers
Identification
Records identified from:
PubMed (n = 393)
Embase (n = 118)
Scopus (n = 29)
Web of science (n = 146)
IEEE Xplore (n =16)
Cochrane library (n = 18)
Total (n = 720)
Records removed *before screening*:
Duplicate records removed (n = 114)
Screening
Records screened (n = 606)
Records excluded via abstract and title screening (n = 524)
Records sought for retrieval (n = 82)
Records not retrieved (n = 11):
Letters, commentaries and editorials (n = 8)
Conference abstracts (n=3)
Records assessed for eligibility (n = 71)
Reports excluded:
Not directly evaluating NLP models (n = 11)
Preprints (n = 5)
Included
Studies included in review from databases search (n = 55)
Studies included in review from snowballing and reference screening (n = 2)
Total (n = 57)

FIGURE 1
PRISMA flowchart.

gastroenterology, while hepatology is the focus of 8 studies. These studies span from 2018 to 2024, with a notable increase in publications in the last 2 years, particularly between 2023 and 2024, which collectively account for 28 of the total included studies. This uptick highlights a growing interest in advanced NLP models like GPT-3 and GPT-4.

The models employed in these studies vary widely, with traditional NLP methods and more recent LLMs like GPT-3 and GPT-4. For instance, Kong et al. (2024) utilized GPT-4 among other versions for medical counseling (38), while Schneider et al. (2023) employed rule-based NLP algorithms for detecting undiagnosed hepatic steatosis (54).

Sample sizes in these studies range from very small datasets to large-scale analyses involving millions of data points, such as in the study by Schneider et al., which analyzed data from over 2.7 million imaging reports (54). The type of data analyzed also varies significantly, encompassing electronic health records (EHRs), pathology reports, and data generated from AI models responding to preset medical queries.

Tasks performed by these models are equally diverse, from diagnostic assistance and disease monitoring to providing patient education and supporting clinical decision-making. Specific examples include the work by Truhn et al. (2024), which focused on extracting structured data from colorectal cancer reports (49), and Lahat et al.

(2023), who evaluated the utility of GPT models in answering patient questions related to gastroenterology (47).

Risk of bias

We used ROBINS-I, QUADAS-2, and PROBAST to map potential biases. Notably, most of the included studies were published in Q1 journals, affirming their scholarly impact and supported by strong SCImago Journal Rank (SJR) scores (Figure 2).

PROBAST results (Supplementary Table S1)

This assessment mostly highlighted low-risk ratings in outcome and analysis domains. However, several studies encountered issues with high participant-related applicability biases, influencing the generalizability of their findings.

QUADAS-2 results (Supplementary Table S2)

A synthesis of QUADAS-2 results revealed that most studies (20 out of 32) exhibited low risk of bias across all four assessed domains. This underscores their methodological robustness and reliability. However, three studies were identified as having a high risk of bias in one of the four categories. Patient selection applicability concerns were notable, primarily due to the reliance on single-center data with specific documentation styles, which may limit the broader applicability of these findings.

ROBINS-I results (Supplementary Table S3)

Analysis of ROBINS-I revealed that 14 studies displayed a moderate risk of bias overall, while one study exhibited a high risk. This was largely due to biases in the selection of participants into the study and confounding factors, particularly because many studies utilized specific questions, queries, or fictional vignettes and case scenarios. Despite these concerns, the other assessment categories predominantly showed low risk. Nonetheless, six studies demonstrated low risk across all evaluated domains.

NLP applications

We categorized the applications of the NLP and LLM models under three main categories for a synthesized analysis of the results: Disease Detection and Diagnosis ($n = 30$), Patient Care ($n = 22$), and Education and Research applications ($n = 5$). Disease Detection and Diagnosis was further divided into Colonoscopy Reports and Other Diagnostic Applications, recognizing that digitized pathology reports—though ultimately part of the broader EHR—were considered separately to better capture unique NLP tasks. Patient Care was subdivided into Management and Communication and Clinical Decision Support, focusing on patient-centered and healthcare professional–oriented applications, respectively (Tables 1, 2 and Figure 3).

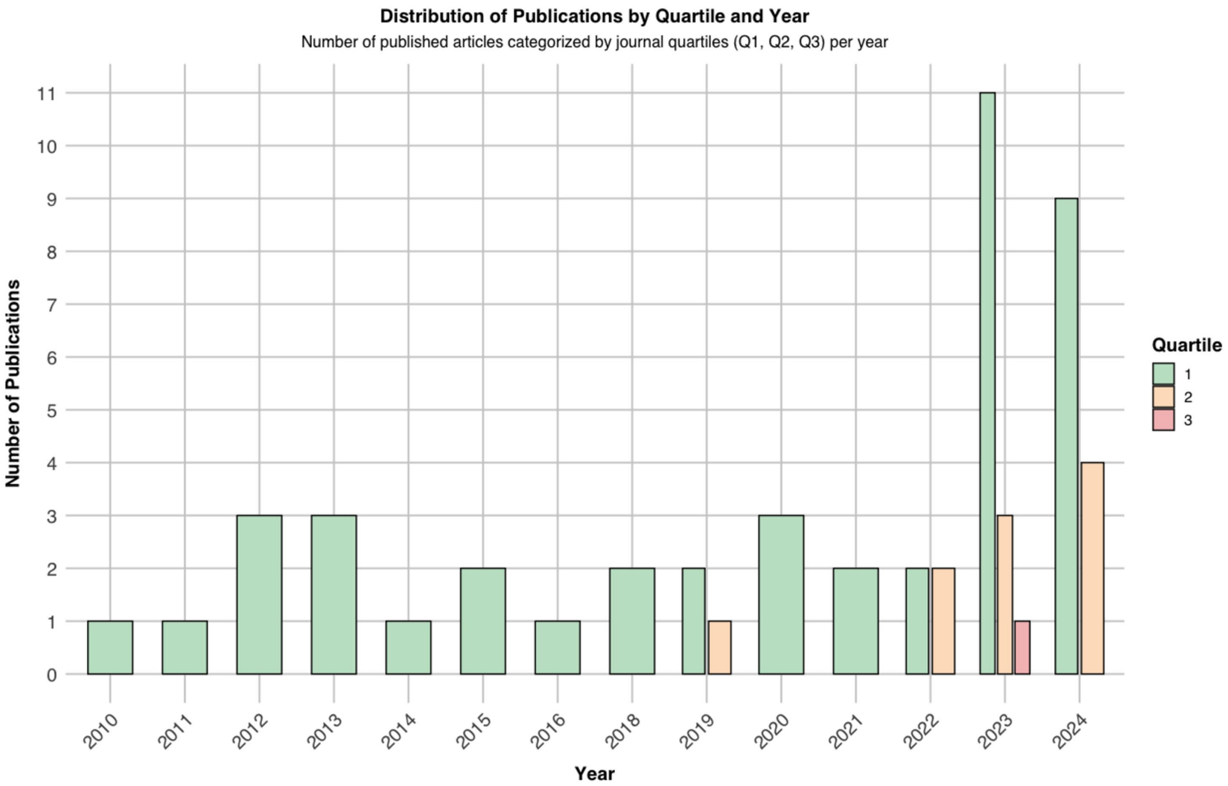


FIGURE 2
Trends of the included studies.

Disease detection and diagnosis

Most of the studies evaluated NLP models in extracting data from colonoscopy reports (n = 17) (Figure 4). Nonetheless, there were many unique applications.

Colonoscopy reports

This category, which includes 17 studies, primarily explored NLP's role in enhancing the interpretation of unstructured colonoscopy reports. Various quality and diagnostic measures were evaluated, such as the adenoma detection rate (ADR), a frequent subject of investigation. For instance, Nayor et al. reported that their NLP pipeline achieved high precision and recall in the automated calculation of ADR (57). Other assessments included polyp detection and sizing, with Imler et al. demonstrating accuracies of 98% for pathology level identification and 96% for size estimation (58). Additionally, Raju et al. noted that NLP matched or exceeded manual methods in identifying and categorizing adenomas with a detection rate of 43% (59). Overall, NLP models showed a broad range of accuracies from 84 to 100%, consistently outperforming manual review methods. Despite needing GPUs, these models reduce the time and effort of manual evaluations.

Other diagnostic applications

Beyond colonoscopy, NLP was applied to a diverse array of diagnostic contexts in gastroenterology and hepatology.

In gastroenterology, several innovative NLP applications have emerged. For example, Wenker et al. utilized NLP to identify dysplasia in Barrett's Esophagus from esophagogastroduodenoscopy (EGD) reports with a high accuracy of 98.7% (69). Song et al. developed a model to extract detailed clinical information such as disease presence, location, size, and stage from unstructured EGD reports, achieving high sensitivity, precision, and accuracy scores (61). Denny et al. applied NLP to enhance colorectal cancer screening by identifying references to four CRC tests within electronic clinical documentation, demonstrating superior recall compared to traditional manual and billing record reviews (63). Additionally, Blumenthal et al. and Parthasarathy et al. used NLP for patient monitoring, with the former detecting non-adherence to follow-up colonoscopies with an AUC of 70.2%, and the latter identifying patients meeting WHO criteria for serrated polyposis syndrome with 93% accuracy (18, 65).

For IBDs, Stidham et al. utilized NLP to detect and infer the activity status of extraintestinal manifestations from clinical notes, enhancing detection accuracy to 94.1% and specificity to 95% (52). Ananthakrishnan et al. explored improving case definitions for Crohn's disease and ulcerative colitis by combining codified data with narrative clinical texts, which identified 6–12% more patients than models using codified data alone, with AUCs of 0.95 for Crohn's disease and 0.94 for ulcerative colitis (53).

In hepatology, NLP has facilitated significant advancements in disease identification and progression monitoring. Sada et al. combined NLP with ICD-9 codes to improve the identification of hepatocellular carcinoma cases from EHR data, significantly enhancing sensitivity and specificity, with an F2 score of 0.92 (71). Van Vleck et al. employed NLP to track disease progression in patients with non-alcoholic fatty liver disease (NAFLD), demonstrating superior sensitivity and F2 scores compared to traditional methods, effectively identifying disease progression from NAFLD to NASH or cirrhosis with sensitivity of 0.93 and an F2 score of 0.92 (30). Furthermore, Sherman et al. developed an NLP model capable of automatically scoring and classifying histological features found in pathology reports related to metabolic associated steatohepatitis (17). The goal was to estimate the risk of progression towards cirrhosis. The model demonstrated high positive and negative predictive values, ranging from 93.5 to 100%, across various histological features (17). Importantly, this NLP model facilitated the creation of a large and quality-controlled cohort of MASLD patients (17).

Patient care

The patient care section is subdivided into two categories: patient management and communication, which comprises 13 studies, and clinical decision support, encompassing 9 studies.

Management and communication

This category explores the use of NLP and LLMs in facilitating communication and management.

In gastroenterology, studies like Lahat et al. evaluated ChatGPT's ability to answer real-life gastroenterology-related patient queries, achieving moderate effectiveness with accuracy scores ranging from 3.4 to 3.9 (47). Choo et al. reported an 86.7% concordance rate between ChatGPT's recommendations for managing complex colorectal cancer cases and decisions made by multidisciplinary teams (39). Furthermore, Lim et al. demonstrated that a contextualized GPT-4 model provided accurate colonoscopy interval advice, significantly outperforming standard models by adhering closely to established guidelines (33). Imler et al. used the cTAKES system to achieve an 81.7% agreement with guideline-adherent colonoscopy surveillance intervals, substantially surpassing manual review accuracies (36). However, studies like Huo et al. and Atarere et al. indicated variability in ChatGPT's performance, suggesting the need for enhancements in AI consistency and reliability (25, 44). In the area of IBD, Zand et al. developed an NLP model that categorized electronic dialog data, showing a 95% agreement with physician evaluations and underscoring the potential of automated chatbots in patient interaction (23). Sciberras et al. found ChatGPT to provide highly accurate (84.2%) and moderately complete responses to patient inquiries about IBD, with particular strengths in topics like smoking and medication (20).

In hepatology, Yeo et al. tested GPT's proficiency in delivering emotional support and accurate information on cirrhosis and hepatocellular carcinoma, achieving correct response rates of 79.1% for cirrhosis and 74% for carcinoma (29). Samaan et al. explored GPT's effectiveness in Arabic, noting a 72.5% accuracy rate, though it was less accurate than its English counterpart, indicating disparities in language performance (34).

Clinical decision support

NLP models were tested for their accuracy and effectiveness in decision-making scenarios. For example, Kong et al. evaluated LLMs' capability to provide counseling on *Helicobacter pylori*, noting that while accuracy was generally high (90% acceptable responses), completeness needed improvement (38). Li et al.'s integration of NLP with machine learning for predicting liver metastases showed impressive results with accuracy and F1 scores around 80.4% (24). The study by Becker et al. utilized an NLP pipeline tailored for German,

TABLE 1 Summary of the included studies.

Author	Year	Data Type + Sample size	Model	Model Task	Main Result
Gastroenterology					
Kong et al. (38)	2024	15 questions related to *H. pylori*	ChatGPT 4.0, ChatGPT 3.5, ERNIE Bot 4.0	Counseling on *H. pylori* infection	ChatGPT 4.0 achieved 90% accuracy in responses and 100% comprehensibility but had a lower completeness rate at 45.6%. ChatGPT 3.5 had an accuracy of 88% and a completeness rate of 40.9%, while ERNIE Bot 4.0 showed lower scores across all metrics.
Lahat et al.	2023	110 real-life patient questions	GPT	Answering patient questions	ChatGPT's accuracy varied across question types, with a mean accuracy score ranging from 3.4 to 3.9 out of 5. It performed better in treatment-related questions (average score: 3.9) compared to diagnostic questions (average score: 3.4).
Truhn et al. (49)	2024	100 colorectal cancer reports	GPT-4	Extracting structured information	GPT-4 achieved 99% accuracy in extracting T-stage, 95% for N-stage, and 94% for M-stage from unstructured histopathology reports.
Zhou et al. (48)	2023	23 medical knowledge questions	GPT-3.5 and GPT-4	Gastric cancer consultation and report analysis	GPT-4 achieved 91.3% appropriateness and 95.7% consistency in a gastric cancer knowledge test. GPT-3.5 had 73.9% appropriateness and 82.6% consistency.
Choo et al. (39)	2024	30 patients with Stage IV or recurrent colorectal cancer	GPT	Formulating management plans	ChatGPT achieved an 86.7% concordance with the Multidisciplinary Tumor Board decisions, including a 73.3% level 1 concordance for first-line treatments.
Huo et al. (44)	2024	Responses for 9 patient cases	ChatGPT, Bing Chat, Google Bard, Claude 2	Providing screening recommendations	ChatGPT aligned with guidelines in 77.8% of clinician cases and 55.6% of patient cases. Bing Chat, Google Bard, and Claude 2 had alignment rates ranging from 25 to 66.7%.
Imler et al. (58)	2013	500 colonoscopy and pathology reports	cTAKES NLP engine	Categorizing pathology findings	The NLP engine achieved 98% accuracy in identifying pathology levels, 97% accuracy for location, and 84% accuracy for the number of adenomas.
Lim et al. (33)	2024	62 example case scenarios, tested three times	GPT-4, contextualized and non-contextualized	Providing advice on colonoscopy intervals	The contextualized GPT-4 model identified high-risk features with 79% accuracy and recommended correct colonoscopy intervals 79% of the time, compared to 51% with the standard model.
Imler et al. (36)	2014	10,798 colonoscopy reports, 6,379 linked to pathology	Clinical text analysis and knowledge extraction system (cTAKES)	Determining colonoscopy surveillance intervals	Achieved an agreement level of 81.7% with manual review, with a Pearson R of 0.813.
Bae et al. (60)	2022	2,425 colonoscopy and pathology reports	Regular expressions and smartTA	Assessing quality indicators	The NLP pipeline achieved 99–100% accuracy for identifying polyp subtypes, anatomical locations, and neoplastic polyps.
Denny et al. (63)	2012	200 patients	KnowledgeMap Concept Identifier	Identifying colorectal cancer tests in EMRs	Achieved 93% recall and 94% precision in identifying CRC tests, outperforming manual reviews (74% recall).
Lahat et al.	2023	20 research questions	GPT	Generating gastroenterology research questions	The model generated relevant questions with a mean clarity score of 4.6 but had low originality (1.5 out of 5).

(Continued)

TABLE 1 (Continued)

Author	Year	Data Type + Sample size	Model	Model Task	Main Result
Laique et al. (26)	2021	35,914 colonoscopy reports	Optical Character Recognition (OCR) and NLP	Extracting quality metrics	Achieved over 95% accuracy for various clinical variables, with some metrics exceeding 99%.
Blumenthal et al. (65)	2015	1,531 patients	NLP tool called QPID	Predicting non-adherence to colonoscopy	Achieved an AUC of 70.2%, with 92% specificity and a PPV of 26%.
Harkema et al. (42)	2011	679 colonoscopy and pathology reports	Rule-based NLP engine	Quality measurement in colonoscopy	Achieved an F-measure of 0.74 and accuracy of 0.89 for various quality metrics.
Raju et al. (59)	2015	12,748 colonoscopy patients	Custom NLP software	Reporting colonoscopy quality metrics	Achieved 91.3% accuracy in identifying screening colonoscopies and 99.4% accuracy in adenoma identification.
Nayor et al. (57)	2018	8,032 screening colonoscopies	NLP pipeline	Calculating adenoma and serrated polyp detection rates	Achieved 100% precision and recall for both adenomas and serrated polyps.
Atarere et al. (25)	2024	20 questions using AI models	ChatGPT, BingChat, and YouChat™	CRC screening advice	Achieved 89.2% inter-rater reliability with variable alignment to clinical guidelines.
Seong et al. (40)	2023	280,668 colonoscopy reports	LSTM, BioBERT, Bi-LSTM-CRF	Extracting information from reports	Bi-LSTM-CRF achieved F1 scores from 0.9564 to 0.9862 across various findings.
Lee et al. (21)	2019	800 colonoscopy reports	Commercial NLP tool	Identifying quality and large polyps	Achieved 100% sensitivity and a PPV of 90.6% for identifying large polyps.
Denny et al. (50)	2010	200 patients	KnowledgeMap concept identifier	Detecting colonoscopy timing and status	Achieved a recall of 0.91 and precision of 0.95 for timing descriptors.
Parthasarathy et al. (18)	2020	323,494 colonoscopy patients	NLP	Diagnosing serrated polyposis syndrome	Achieved 93% accuracy in identifying correct SPS diagnoses.
Rammohan et al. (68)	2024	NR	GPT-4 and Bard	Answering standard gastroenterology questions	ChatGPT 4.0 achieved a mean reliability score of 6.23, while Bard had a mean of 2.04.
Pereyra et al. (37)	2024	238 physicians	GPT-3.5	Assessing CRC screening recommendations	ChatGPT had a mean score of 4.5/10, compared to 7.71/10 for physicians with the app.
Song et al. (61)	2022	1,000 validation, 248,966 application EGD reports	Custom NLP pipeline	Extracting information from EGD reports	Achieved sensitivity, PPV, accuracy, and F1 scores above 0.966 for various conditions.
Peng et al. (45)	2024	131 colorectal cancer questions	GPT-3.5	Answering CRC-related questions	Achieved a mean accuracy score of 0.91, but lower comprehensiveness (0.85).
Tinmouth et al. (70)	2023	1,450 pathology reports	NLP	Identifying adenomas for ADR	Achieved sensitivity of 99.60% and specificity of 99.01%.
Mehrotra et al. (27)	2012	24,157 colonoscopy reports	NLP (C-QUAL)	Assessing colonoscopy quality measures	Achieved kappa >0.7 for nine out of 20 measures.
Becker et al. (64)	2019	2,513 German clinical notes from 500 patients	German-specific NLP pipeline	Guideline-based treatment evaluation	Achieved 96.64% precision and 94.89% recall for tumor stage detection.
Hou et al. (31)	2013	575 colonoscopy pathology reports	Automated Retrieval Console (ARC)	Identifying surveillance colonoscopy	Achieved 77% recall and 80% precision for surveillance reports.

(Continued)

TABLE 1 (Continued)

Author	Year	Data Type + Sample size	Model	Model Task	Main Result
Gorelik et al. (51)	2023	20 clinical scenarios	GPT-4	Post-colonoscopy patient management	Achieved 90% compliance with guidelines and an 85% accuracy in recommendations.
Samaan et al. (34)	2023	91 questions on liver cirrhosis	GPT	Answering cirrhosis-related questions in Arabic	Achieved 72.5% accuracy in Arabic, with comprehensive responses only in 24.2% of cases.
Cankurtaran et al. (67)	2023	20 questions on Crohn's disease and ulcerative colitis	GPT	Responding to IBD queries	Scored higher for professional queries (mean reliability: 6/7) than for patient queries (mean: 4/7).
Nguyen Wenker et al. (69)	2023	1,000 patients for NLP validation	CLAMP NLP software	Identifying dysplasia in Barrett's Esophagus	Achieved 98.7% accuracy, 100% precision, and 92.3% recall.
Imler et al. (66)	2018	23,674 ERCP procedures	NLP	Quality measurement for ERCP	Achieved accuracy of 90–100% and precision of 84–100%.
Li et al. (62)	2021	5,570 patients	NLP	Identifying Lynch Syndrome for MMR screening	Achieved 100% sensitivity, specificity, PPV, and NPV.
Li et al. (16)	2022	22,206 patients across various tests	ENDOANGEL-AS NLP and deep learning	Identifying high-risk patients for surveillance	Achieved 100% accuracy in internal testing and 99.91% in external testing.
Wagholikar et al. (35)	2012	53 patients	NLP	Providing colonoscopy surveillance guidance	Made optimal recommendations in 90.6% of cases.
Sciberras et al. (20)	2024	38 questions from IBD patients	GPT-3.5	Generating responses to IBD patient queries	Achieved 84.2% accuracy with a median score of 4.0 for completeness.
Stidham et al. (52)	2023	1,240 patients with IBD	NLP	Detecting and inferring EIM activity status	Achieved 94.1% accuracy, with sensitivity of 0.92 and specificity of 0.95.
Ganguly et al. (22)	2023	2,276 colonoscopy procedures	NLP	Adenoma detection and report card generation	Achieved 100% sensitivity, specificity, and accuracy.
Ma et al. (43)	2024	165 esophageal ESD cases	GPT-3.5	Post-procedural quality control for esophageal ESD	Achieved accuracy of 92.5–100% across different factors.
Gravina et al. (32)	2024	Questions from 2023 Italian medical exam	GPT 3.5 and Perplexity AI	Answering medical residency exam questions	GPT 3.5 achieved 94.11% correct responses in the latest exam.
Fevrier et al.	2020	401,566 colonoscopy linked with pathology reports	SAS® PERL NLP tool	Extracting data from colonoscopy reports	Achieved Cohen's κ between 93 and 99% and PPV of 97–100% for common categories.
Benson et al. (55)	2023	24,584 pathology reports	NLP pipeline	Extracting features of colorectal polyps	Achieved 98.9% precision and 98.0% recall, with an F1-score of 98.4%.
Zand et al. (23)	2020	16,453 lines of dialog from 424 patients	NLP model	Developing a chatbot for IBD patient support	Achieved 95% agreement with physician evaluations in categorizing dialogs.
Ananthakrishnan et al. (53)	2013	1,200 patients for Crohn's and UC	NLP techniques	Improving EMR case definitions for IBD	Achieved AUC of 0.95 for CD and 0.94 for UC.

(Continued)

TABLE 1 (Continued)

Author	Year	Data Type + Sample size	Model	Model Task	Main Result
Wang et al. (72)	2024	200 medical discharge summaries	GPT-4	Classifying GI bleeding events	GPT-4 showed high accuracy (94.4% for identifying GI bleeding), outperforming ICD codes significantly and demonstrating comparable or slightly lower accuracy to human reviewers
Hepatology					
Benedicenti et al. (56)	2023	56 gastroenterologists, 25 residents, 31 specialists	GPT-3	Answering clinical vignettes on Hepatology and Gastroenterology	Demonstrated improvement over time, underperformed vs. humans
Li et al. (24)	2023	1,463 postoperative colorectal cancer patients	NLP and machine learning integration	Predicting liver metastases	High accuracy in risk prediction
Wang et al. (41)	2022	LiverTox database	DeepCausality framework	Causal inference for drug-induced liver injury	Achieved 92% accuracy and an F1-score of 0.84 for DILI predictions.
Yeo et al. (29)	2023	164 questions about cirrhosis and hepatocellular carcinoma	GPT	Providing answers on cirrhosis and HCC	GPT provided accurate knowledge on cirrhosis (79.1% correct) and hepatocellular carcinoma (74% correct), although only a small proportion were considered comprehensive (cirrhosis 47.3%, HCC 41.1%).
Sherman et al. (17)	2024	3,134 patients with liver disease	NLP	Classifying liver disease pathology	The NLP model achieved high positive and negative predictive values (93.5–100%) across different histological features
Van Vleck et al. (30)	2019	38,575 patients	CLiX clinical NLP engine	Identifying NAFLD patients and disease progression	The NLP model demonstrated superior sensitivity and F2 scores compared to ICD codes and text searches. Sensitivity of 0.93 and an F2 score of 0.92 in identifying NAFLD
Sada et al. (71)	2016	1,138 patients identified from ICD-9 codes	Automated Retrieval Console (ARC)	Improving identification of hepatocellular cancer	Combining ICD-9 codes with NLP improved HCC identification: pathology (PPV 0.96, sensitivity 0.96, specificity 0.97), radiology (PPV 0.75, sensitivity 0.94, specificity 0.68)
Pradhan et al. (28)	2024	22 patients/caregivers and transplant hepatologists	Multiple LLMs	Generating patient educational materials about cirrhosis	AI materials matched human readability but were rated less actionable.
Schneider et al. (54)	2023	2.15 million pathology and 2.7 million imaging reports	Rule-based NLP algorithm	Identifying hepatic steatosis	Identified 3,007 biopsy-proven NAFLD cases and 42,083 imaging-proven cases, with a PPV of 99.7%.

AI, artificial intelligence; AUC, area under the curve; BARD, Google's Generative AI Model; BioBERT, biomedical bidirectional encoder representations from transformers; cTAKES, clinical text analysis and knowledge extraction system; CD, Crohn's disease; CDSS, clinical decision support system; CLAMP, clinical language annotation modeling and processing; CLiX, clinical information extraction; CRC, colorectal cancer; DILI, drug-induced liver injury; EMR, electronic medical record; ENDOSC, endoscopic submucosal dissection; EIM, extraintestinal manifestations; ERNIE, enhanced representation through knowledge integration; F1, F1-score; GI, gastrointestinal; GPT, generative pre-trained transformer; HCC, hepatocellular carcinoma; IBD, inflammatory Bowel disease; ICD, international classification of diseases; LSTM, long short-term memory; MMR, mismatch repair; NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic steatohepatitis; NLP, natural language processing; NPV, negative predictive value; OCR, optical character recognition; PPV, positive predictive value; SPS, serrated polyposis syndrome; UC, ulcerative colitis.

TABLE 2 Included studies designs, comparisons and validations methods.

Author	Study design	Field + Specific interest	Model	Human comparator	Validation	Limitations
Gastroenterology						
Kong et al. (38)	Comparative analysis	Gastroenterology, *H. pylori*	ChatGPT 4.0, ChatGPT 3.5, ERNIE Bot 4.0	None	None	Limited completeness and possible bias in non-English settings.
Lahat et al. (46)	Cross-sectional analysis	Gastroenterology, General	GPT	3 Gastroenterologists	None	Variation in response quality; occasional inaccuracies.
Truhn et al. (49)	Retrospective analysis	Gastroenterology, CRC	GPT-4	Manual data extraction	Internal	Challenges with OCR accuracy; handling of handwritten notes.
Zhou et al. (48)	Retrospective analysis	Gastroenterology, Gastric cancer	GPT-3.5 and GPT-4	None	None	Importance of human oversight and detail accuracy limitations.
Choo et al. (39)	Prospective analysis	Gastroenterology, CRC	GPT	MDT decisions	None	Small sample size; retrospective nature.
Huo et al. (44)	Retrospective analysis	Gastroenterology, CRC screening	ChatGPT, Bing Chat, Google Bard, Claude 2	None	None	Variability in responses between chatbots; static data collection point.
Imler et al. (58)	Retrospective cohort study	Gastroenterology, Colonoscopy (adenoma detection)	cTAKES NLP engine	Manual review	Internal	Single institution study; template-driven reports.
Lim et al. (33)	Retrospective analysis	Gastroenterology, CRC screening	GPT-4	None	Internal	Free-text input limitations; lack of standardized prompts.
Imler et al. (36)	Retrospective analysis	Gastroenterology, Colonoscopy intervals	cTAKES NLP engine	Paired, blinded experts	None	Difficulty with complex cases needing manual review.
Bae et al. (60)	Retrospective analysis	Gastroenterology, Colonoscopy quality	Regular expressions, smartTA	5 Human annotators	Internal	Dataset-specific NLP system; reliance on accurate input data.
Denny et al. (63)	Retrospective analysis	Gastroenterology, CRC screening	KnowledgeMap Concept Identifier	Manual chart review	Internal	Single-center study with small sample size.
Lahat et al. (47)	Retrospective analysis	Gastroenterology, Research questions	GPT	3 Gastroenterologists	None	Small expert panel; subjective ratings.
Laique et al. (26)	Retrospective analysis	Gastroenterology, CRC screening	OCR + NLP	Manual review	Internal & External	Variability in documentation styles; reliance on high-quality scans.
Blumenthal et al. (65)	Retrospective analysis	Gastroenterology, Colonoscopy adherence	QPID NLP tool	None	Internal	Sample representativeness; generalizability to other settings.
Harkema et al. (42)	Retrospective analysis	Gastroenterology, Colonoscopy quality	Rule-based NLP engine	Manual annotations	Internal	Single institution study; mix of report styles.
Raju et al. (59)	Retrospective analysis	Gastroenterology, Colonoscopy ADR	Custom NLP software	Manual review	None	Institution-specific study; not tested on other systems.

(Continued)

TABLE 2 (Continued)

Author	Study design	Field + Specific interest	Model	Human comparator	Validation	Limitations
Nayor et al. (57)	Retrospective analysis	Gastroenterology, Colonoscopy ADR	Custom NLP pipeline	Manual review	Internal	Misclassification risks; complexity of free text.
Atarere et al. (25)	Cross-sectional analysis	Gastroenterology, CRC screening	ChatGPT, BingChat, YouChat™	Board-certified physicians	None	Models not designed for medical use; reliance on training data.
Seong et al. (40)	Retrospective analysis	Gastroenterology, Colonoscopy	Bi-LSTM-CRF	None	Internal	Single institution study; reporting style variations.
Lee et al. (21)	Cross-sectional analysis	Gastroenterology, Colonoscopy quality	Commercial NLP tool	Manual chart review	Internal	Single healthcare system study; dependency on physician reports.
Denny et al. (50)	Retrospective analysis	Gastroenterology, CRC screening	KnowledgeMap Concept Identifier	Manual review	Internal	Focus on colonoscopies; error sources in date references.
Parthasarathy et al. (18)	Retrospective cohort study	Gastroenterology, CRC screening	NLP	Clinicians	None	7% error rate in data extraction.
Rammohan et al. (68)	Prospective analysis	Gastroenterology, General	GPT-4, Bard	None	None	Focus on specific questions; limited to two AI tools.
Pereyra et al. (37)	Prospective observational study	Gastroenterology, CRC	GPT-3.5	Physicians	None	Small number of vignettes; outdated model.
Song et al. (61)	Retrospective analysis	Gastroenterology, Gastroscopy	Custom NLP pipeline	Gastroenterologists	Internal	Specificity to data formatting; need for manual updates.
Peng et al. (45)	Prospective observational study	Gastroenterology, CRC	GPT-3.5	Expert answers	Internal	Limited question scope from a reference book.
Tinmouth et al. (70)	Retrospective analysis	Gastroenterology, Colonoscopy ADR	NLP	Expert review	Internal	Voluntary reporting system; procedure-specimen mismatch.
Mehrotra et al. (27)	Cross-sectional analysis	Gastroenterology, Colonoscopy	NLP (C-QUAL)	Physician manual review	Internal	Single healthcare system study.
Becker et al. (64)	Retrospective analysis	Gastroenterology, CRC	German-specific NLP	Manual review	Internal	Documentation complexity; moderate performance in some areas.
Hou et al. (31)	Retrospective analysis	Gastroenterology, IBD surveillance	ARC	Gastroenterologist	Internal	Pathology reporting variability.
Gorelik et al. (51)	Prospective observational study	Gastroenterology, Postcolonoscopy	GPT-4	Society guidelines	None	Inherent randomness and outdated training data.
Samaan et al. (34)	Cross-sectional analysis	Gastroenterology, Cirrhosis	GPT	Transplant hepatologist	None	Model hallucinations; Arabic response accuracy gap.
Cankurtaran et al. (67)	Retrospective analysis	Gastroenterology, IBD	GPT-4	None	None	Response variability; lack of detail in treatment advice.

(Continued)

TABLE 2 (Continued)

Author	Study design	Field + Specific interest	Model	Human comparator	Validation	Limitations
Wenker et al. (69)	Retrospective analysis	Gastroenterology, Barrett's Esophagus	CLAMP	Manual review	Internal & External	VA sample limits generalizability.
Imler et al. (66)	Retrospective cohort study	Gastroenterology, ERCP	Apache UIMA-based NLP	Gastroenterologist	Internal	Single-center bias; ICD coding assumptions.
Li et al. (62)	Retrospective analysis	Gastroenterology and hepatology	ML and NLP fusion	Two experienced physicians	External	Scale of data integration and complexity; limited interpretability.
Li et al. (16)	Retrospective cohort study	Gastroenterology, Upper GI cancer	ENDOANGEL-AS	Physicians	Internal & External	Annotation variability; semi-structured data limits.
Wagholikar et al. (35)	Retrospective analysis	Gastroenterology, Colonoscopy	NLP	Gastroenterologist	Internal	Single expert; single institution.
Sciberras et al. (20)	Prospective study	Gastroenterology, IBD	GPT-3.5	None	None	Lacks detailed responses; occasional inaccuracies.
Stidham et al. (52)	Retrospective cohort study	Gastroenterology, IBD	NLP	Human reviewers	Internal	Source document variation limits generalizability.
Ganguly et al. (22)	Retrospective analysis	Gastroenterology, Colonoscopy ADR	NLP	Manual review	Internal	Human input reliance; data integration complexity.
Ma et al. (43)	Retrospective analysis	Gastroenterology, Esophageal ESD	GPT-3.5	Human operators	Internal	Single-center dataset limits; small prompt optimization cases.
Gravina et al. (32)	Cross-sectional analysis	Gastroenterology, Education	GPT 3.5, Perplexity AI	None	Internal	AI education lacks oversight; variable performance.
Fevrier et al. (84)	Retrospective cohort study	Gastroenterology, Colonoscopy	NLP tool	Manual review	Internal	Incomplete documentation and specimen data challenges.
Benson et al. (55)	Retrospective analysis	Gastroenterology, Colonoscopy	NLP	Manual annotations	None	Report structure adaptations; evaluation of rare features limited.
Zand et al. (23)	Retrospective cohort study	Gastroenterology, IBD	NLP	3 Physicians	None	Homogeneous patient sample limits generalizability.
Ananthakrishnan et al. (53)	Retrospective analysis	Gastroenterology, IBD	NLP	None	Internal	Single healthcare system; needs broader validation.
Wang et al. (72)	Retrospective analysis	Gastroenterology, GI bleeding	GPT-4	Human reviewers	None	Single clinical scenario focus; model specificity.
Hepatology						
Benedicenti et al. (56)	Cross-sectional analysis	Gastroenterology, Education	GPT-3	Gastroenterologists	None	Performance variability; static format.
Li et al. (24)	Retrospective cohort study	Hepatology, HCC	NLP	Manual review	External	Limited generalizability; physician agreement variability.
Wang et al. (41)	Retrospective analysis	Hepatology, DILI	DeepCausality	None	Internal	Dependency on structured data; specificity to LiverTox.

(Continued)

TABLE 2 (Continued)

Author	Study design	Field + Specific interest	Model	Human comparator	Validation	Limitations
Yeo et al. (29)	Retrospective analysis	Hepatology, Cirrhosis	GPT	Transplant hepatologists	None	Inconsistent comprehensiveness; limited regional guideline knowledge.
Sherman et al. (17)	Retrospective cohort study	Hepatology, MASLD	NLP	Manual review	Internal	Heterogeneous biopsy data; evolving MASLD definitions.
Van Vleck et al. (30)	Retrospective cohort study	Hepatology, NAFLD	CLiX NLP engine	Manual validation	Internal	Generalizability limits; dependency on physician impressions.
Sada et al. (71)	Retrospective analysis	Hepatology, HCC	ARC	Manual classification	Internal and External	VA-specific data limits generalizability.
Pradhan et al. (28)	Retrospective analysis	Hepatology, Cirrhosis	Multiple LLMs	Human materials	None	AI not designed for medical use; lack of visual aids.
Schneider et al. (54)	Retrospective analysis	Hepatology, NAFLD	NLP	Manual review	Internal	Bias in EHR analysis; NLP interpretation challenges.

AI, artificial intelligence; ADR, adenoma detection rate; ARC, automated retrieval console; CDSS, clinical decision support system; CRC, colorectal cancer; cTAKES, clinical text analysis and knowledge extraction system; DILI, drug-induced liver injury; EHR, electronic health records; EMR, electronic medical records; ESD, endoscopic submucosal dissection; GI, gastrointestinal; GPT, generative pre-trained transformer; HCC, hepatocellular carcinoma; IBD, inflammatory Bowel disease; IHC, immunohistochemistry; LLM, large language model; MDT, multidisciplinary tumor board; ML, machine learning; NAR, non-adherence ratio; NASH, non-alcoholic steatohepatitis; NAFLD, non-alcoholic fatty liver disease; NLP, natural language processing; NPV, negative predictive value; OCR, optical character recognition; PPV, positive predictive value; SPS, serrated polyposis syndrome; USMSTF, U.S. multi-society task force.

achieving high precision and recall in guideline-based treatment extraction from clinical notes (64). Further, Wang et al.'s "DeepCausality" framework accurately assessed causal factors for drug-induced liver injuries, aligning well with clinical guidelines (41). Another significant study, Wagholikar et al., demonstrated that an NLP-powered clinical decision support system could assist in making guideline-adherent recommendations for colonoscopy surveillance, as it made optimal recommendations in 48 out of 53 cases (35).

Education and research

Five studies focused on this aspect. Generally, NLP and LLMs have demonstrated a promising capacity to enhance learning and knowledge dissemination. Benedicenti et al. explored the accuracy of ChatGPT in solving clinical vignettes against gastroenterologists, noting an initial 40% accuracy that improved to 65% over time, suggesting a potential for future clinical integration with continued advancements (56). Zhou et al. assessed GPT-3.5 and GPT-4 for their ability to provide consultation recommendations and analyze gastroscopy reports related to gastric cancer, with GPT-4 achieving 91.3% appropriateness and 95.7% consistency (48). Lahat et al. utilized GPT to generate research questions in gastroenterology, finding the questions relevant and clear but lacking in originality (46). Meanwhile, Gravina et al. highlighted the efficacy of ChatGPT 3.5 in medical education, as it outperformed Perplexity AI in residency exam questions with a 94.11% accuracy rate (32). Additionally, Pradhan et al. compared AI-generated patient educational materials on cirrhosis with human-derived content, finding no significant differences in readability or accuracy, though human materials were deemed more actionable (28).

Validation and comparisons

Of the 57 studies, only 5 performed external validation using independent datasets. A total of 30 studies used internal validation, with 27 applying classical subsets of the same data for re-testing and validating their main results. Three studies employed a method of running LLM prompts multiple times (2–3 times) to assess the consistency of responses. Meanwhile, 22 studies did not perform any validation. Regarding direct comparisons of NLPs and LLMs with human counterparts, 44 studies compared the model's performance with manual review by physicians or manual data extraction methods. The number of human reviewers varied between studies, ranging from 1 to 5. Thirteen studies did not perform direct comparisons (Table 2).

Discussion

Our systematic review assessed the integration of NLP and LLMS in gastroenterology and hepatology, registering significant advancements. We reviewed 57 studies, highlighting a sharp increase in research over the last 2 years, particularly focusing on newer models like GPT-3 and GPT-4. These studies reflect a shift from traditional tasks, such as report analysis, to more dynamic roles in patient management and research facilitation.

To present the findings in a clear and easily interpretable manner, we opted to categorize the reviewed studies into a minimal set of

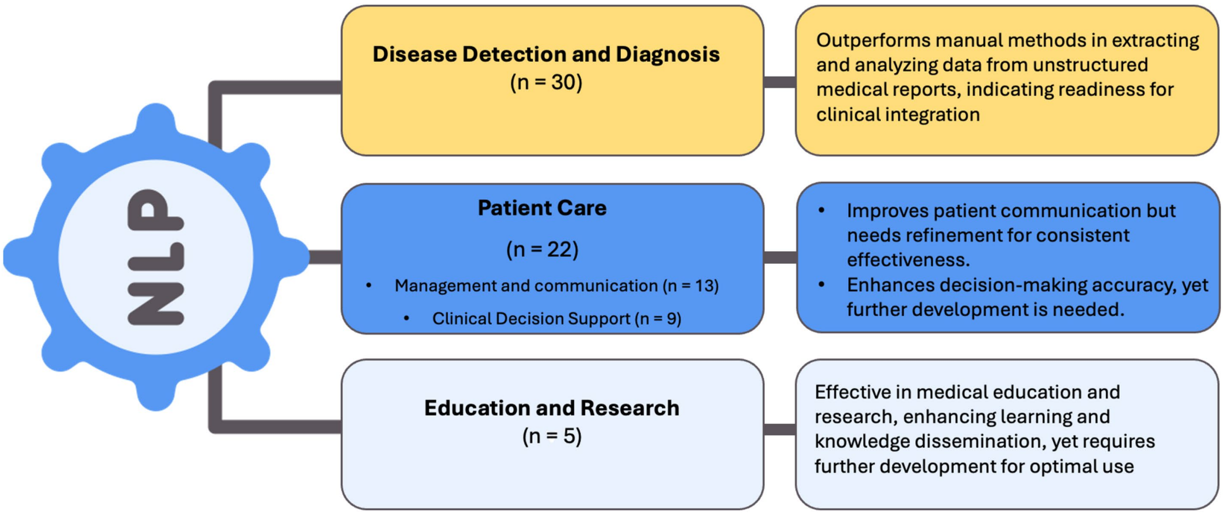
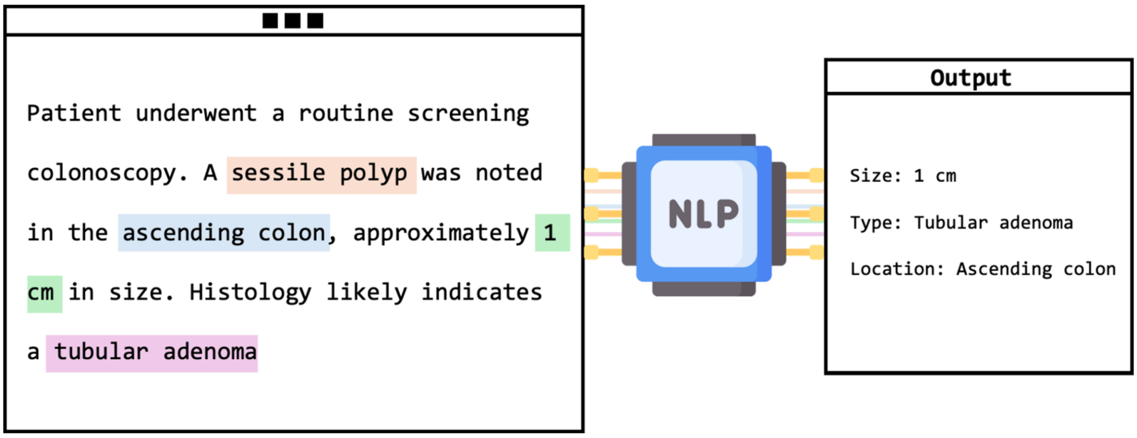

NLP
Disease Detection and Diagnosis
(n = 30)
Outperforms manual methods in extracting and analyzing data from unstructured medical reports, indicating readiness for clinical integration
Patient Care
(n = 22)
• Management and communication (n = 13)
• Clinical Decision Support (n = 9)
• Improves patient communication but needs refinement for consistent effectiveness.
• Enhances decision-making accuracy, yet further development is needed.
Education and Research
(n = 5)
Effective in medical education and research, enhancing learning and knowledge dissemination, yet requires further development for optimal use

FIGURE 3
Summary of NLP applications and outcomes.



```
Patient underwent a routine screening
colonoscopy. A sessile polyp was noted
in the ascending colon, approximately 1
cm in size. Histology likely indicates
a tubular adenoma
```

```
Output

Size: 1 cm
Type: Tubular adenoma
Location: Ascending colon
```

FIGURE 4
Visual framework of NLP extracting adenoma characteristics from unstructured colonoscopy report.

broad application areas. We acknowledge that these categories are not absolute and that certain studies may naturally span multiple domains (for example, colonoscopy surveillance intervals, while placed under one heading, could also be considered a form of clinical decision support). Nonetheless, by grouping the research into broader, more encompassing categories, we aimed to give readers a high-level understanding of where progress is most pronounced, and which areas appear closer to real-world clinical integration.

The results show that certain NLP applications seem ready for immediate clinical use. For example, Schneider et al. (2023) identified 42,000 hepatic steatosis cases using an NLP model on 2.15 million pathology reports and 2.7 million imaging reports. This level of precision (PPV 99.7%) exemplifies NLP's readiness to support diagnostic processes in large-scale healthcare settings. Similarly, Truhn et al. (2024) successfully employed GPT-4 to extract structured data from colorectal cancer reports with a precision of 99% for T-stage identification, suggesting a high reliability of NLP in processing and structuring complex pathological data.

Conversely, the technology's expansion into more dynamic roles such as comprehensive disease management and holistic patient care is still evolving. For instance, Kong et al. (2024) found that while the accuracy and comprehensibility of GPT-4's responses to medical inquiries about *Helicobacter pylori* were high, the completeness of the information was less satisfactory. This indicates ongoing challenges in ensuring that NLP outputs are not only accurate but also fully informative.

Our results suggest that both classic NLP methods and newer models can be effectively integrated to streamline manual tasks such as extracting data and making diagnoses from complex and unstructured reports, with an accuracy that typically surpasses manual screening (16–18, 21, 22, 27, 33). This builds upon and adds on a previous systematic review of NLP in gastroenterology and hepatology conducted by Hou et al. (2). While he found promising results, he emphasized the need for careful consideration of the quality of clinical data within EHRs, and also highlighted the importance of understanding variations and deviations from established clinical practice standards (2). Our updated results indicate that these models consistently demonstrate high accuracies (16–18, 21, 22, 27, 33). This trend is observable in other fields utilizing NLP, such as radiology and infectious diseases (74, 75). However, our research suggests that applying these methods to more complex tasks like patient management, education, and clinical decision-making is still challenging (20, 29, 34, 37). While newer models show promising results, there are significant limitations and variability that require further development (67). This trend is consistent with data and the current findings from other fields (76, 77).

Several limitations of our review must be acknowledged. Many studies utilize single-institution datasets, which could affect the generalizability of the findings. This is important especially because only 5 studies (8.7%) reported performing an external validation. The accuracy of NLP outputs is heavily dependent on the quality of the input data, with errors or inconsistencies in medical records potentially leading to inaccurate results (78). The opaque nature of AI decision-making processes (‘black box’) raises concerns about the transparency and trustworthiness of these models in clinical settings (79). Ethical considerations around potential biases in training data and algorithmic outputs underscore the necessity for careful implementation to ensure fairness and equity in healthcare delivery (80). Moreover, the accuracy and reliability of NLP and LLM outputs are directly tied to the quality of the input data. EHRs, clinical notes, and imaging reports often contain incomplete, ambiguous, or inaccurately recorded information. These data imperfections can lead to propagation of errors and compound biases within the model’s output, potentially influencing clinical decision-making and patient care. Additionally, while many NLP and LLM models show promise in structured tasks like disease detection or data extraction, they remain susceptible to “hallucinations”—generating plausible-sounding but factually incorrect statements (81). Such errors, if undetected, may result in misguided clinical judgments, suboptimal patient management, and delayed interventions. An additional critical dimension of these limitations involves the potential for algorithmic biases, including those related to sociodemographic factors such as race, ethnicity, gender, language proficiency, and socioeconomic status (82, 83). Models trained on unrepresentative or historically biased data risk perpetuating systemic inequalities in healthcare. Despite the promising accuracy of some NLP applications, they are not yet widely integrated into day-to-day clinical workflows, particularly for patient care and decision-making; current limitations and the need for thorough testing and validation—especially for newer, less researched techniques—have thus far hindered their routine implementation in practice.

In conclusion, our systematic review highlights the impact of NLP and LLMs in gastroenterology and hepatology. On one hand, NLP has already proven its utility in screening and analyzing medical reports, facilitating streamlined screening policies with impressive outcomes. On the other hand, the capabilities of newer LLMs are still unfolding, with their full potential in complex management and research roles yet to be fully realized. The results demonstrate that while some applications of NLP are well-established and highly effective, newer LLMs offer exciting, emerging applications that promise to further enhance clinical practice. Moving forward, research focus should be on refining these models, and externally validating the results to ensure prospectively they meet real-world clinical needs.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MO: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SN: Supervision, Validation, Writing – review & editing. KS: Supervision, Validation, Writing – review & editing. BG: Conceptualization, Formal analysis, Writing – review & editing. GN: Supervision, Writing – review & editing. EK: Conceptualization, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1512824/full#supplementary-material

References

1. Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of chat GPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. *Ther Adv Gastroenterol*. (2023) 16:17562848231218618. doi: 10.1177/17562848231218618

2. Hou JK, Imler TD, Imperiale TF. Current and future applications of natural language processing in the field of digestive diseases. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. (2014) 12:1257–61. doi: 10.1016/j.cgh.2014.05.013

3. Dave T, Athaluri SA, Singh S. Chat GPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. (2023) 6:1169595. doi: 10.3389/frai.2023.1169595

4. Nehme F, Feldman K. Evolving role and future directions of natural language processing in gastroenterology. *Dig Dis Sci*. (2021) 66:29–40. doi: 10.1007/s10620-020-06156-y

5. Tavanapong W, Oh J, Riegler MA, Khaleel M, Mittal B, de Groen PC. Artificial intelligence for colonoscopy: past, present, and future. *IEEE J Biomed Health Inform*. (2022) 26:3950–65. doi: 10.1109/JBHI.2022.3160098

6. Stidham RW. Artificial intelligence for understanding imaging, text, and data in gastroenterology. *Gastroenterol Hepatol*. (2020) 16:341–9.

7. Zaver HB, Patel T. Opportunities for the use of large language models in hepatology. *Clin Liver Dis*. (2023) 22:171–6. doi: 10.1097/CLD.0000000000000075

8. Shahab O, El Kurdi B, Shaukat A, Nadkarni G, Soroush A. Large language models: a primer and gastroenterology applications. *Ther Adv Gastroenterol*. (2024) 17:17562848241227031. doi: 10.1177/17562848241227031

9. Schiavo JH. PROSPERO: An international register of systematic review protocols. *Med Ref Serv Q*. (2019) 38:171–80. doi: 10.1080/02763869.2019.1588072

10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. (2021) 372:n71. doi: 10.1136/bmj.n71

11. Brietzke E, Gomes FA, Gerchman F, Freire RCR. Should systematic reviews and meta-analyses include data from preprints? *Trends Psychiatry Psychother*. (2023) 45:e20210324.

12. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. (2016) 5:210. doi: 10.1186/s13643-016-0384-4

13. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. (2016) 355:i4919. doi: 10.1136/bmj.i4919

14. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. (2011) 155:529–36. doi: 10.7326/0003-4819-155-8-201110180-00009

15. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of Bias and applicability of prediction model studies. *Ann Intern Med*. (2019) 170:51–8. doi: 10.7326/M18-1376

16. Li J, Hu S, Shi C, Dong Z, Pan J, Ai Y, et al. A deep learning and natural language processing-based system for automatic identification and surveillance of high-risk patients undergoing upper endoscopy: a multicenter study. *EClinicalMedicine*. (2022) 53:101704. doi: 10.1016/j.eclinm.2022.101704

17. Sherman MS, Challa PK, Przybyszewski EM, Wilechansky RM, Uche-Anya EN, Ott AT, et al. A natural language processing algorithm accurately classifies steatotic liver disease pathology to estimate the risk of cirrhosis. *Hepatol Commun*. (2024) 8:e 0403. doi: 10.1097/HC9.0000000000000403

18. Parthasarathy G, Lopez R, McMichael J, Burke CA. A natural language-based tool for diagnosis of serrated polyposis syndrome. *Gastrointest Endosc*. (2020) 92:886–90. doi: 10.1016/j.gie.2020.04.077

19. A Transparent and Adaptable Method to Extract Colonoscopy and Pathology Data Using Natural Language Processing-Pub Med. (2023) Available at: https://pubmed.ncbi.nlm.nih.gov/32737597/

20. Sciberras M, Farrugia Y, Gordon H, Furfaro F, Allocca M, Torres J, et al. Accuracy of information given by chat GPT for patients with inflammatory bowel disease in relation to ECCO guidelines. *J Crohns Colitis*. (2024) 18:1215–21. doi: 10.1093/ecco-jcc/jjae040

21. Lee JK, Jensen CD, Levin TR, Zauber AG, Doubeni CA, Zhao WK, et al. Accurate identification of colonoscopy quality and polyp findings using natural language processing. *J Clin Gastroenterol*. (2019) 53:e25–30. doi: 10.1097/MCG.0000000000000929

22. Ganguly EK, Purvis L, Reynolds N, Akram S, Lidofsky SD, Zubarik R. An accurate and automated method for adenoma detection rate and report card generation utilizing common electronic health records. *J Clin Gastroenterol*. (2023) 58:656–60. doi: 10.1097/MCG.0000000000001915

23. Zand A, Sharma A, Stokes Z, Reynolds C, Montilla A, Sauk J, et al. An exploration into the use of a Chatbot for patients with inflammatory bowel diseases: retrospective cohort study. *J Med Internet Res*. (2020) 22:e15589. doi: 10.2196/15589

24. Li J, Wang X, Cai L, Sun J, Yang Z, Liu W, et al. An interpretable deep learning framework for predicting liver metastases in postoperative colorectal cancer patients using natural language processing and clinical data integration. *Cancer Med*. (2023) 12:19337–51. doi: 10.1002/cam4.6523

25. Atarere J, Naqvi H, Haas C, Adewunmi C, Bandaru S, Allamneni R, et al. Applicability of online chat-based artificial intelligence models to colorectal Cancer screening. *Dig Dis Sci*. (2024) 69:791–7. doi: 10.1007/s10620-024-08274-3

26. Laique SN, Hayat U, Sarvepalli S, Vaughn B, Ibrahim M, McMichael J, et al. Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports. *Gastrointest Endosc*. (2021) 93:750–7. doi: 10.1016/j.gie.2020.08.038

27. Mehrotra A, Dellon ES, Schoen RE, Saul M, Bishehsari F, Farmer C, et al. Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. *Gastrointest Endosc*. (2012) 75:1233–1239.e14. doi: 10.1016/j.gie.2012.01.045

28. Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. *Hepatol Commun*. (2024) 8:e 0367. doi: 10.1097/HC9.0000000000000367

29. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of chat GPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. (2023) 29:721–32. doi: 10.3350/cmh.2023.0089

30. Van Vleck TT, Chan L, Coca SG, Craven CK, Do R, Ellis SB, et al. Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression. *Int J Med Inform*. (2019) 129:334–41. doi: 10.1016/j.ijmedinf.2019.06.028

31. Hou JK, Chang M, Nguyen T, Kramer JR, Richardson P, Sansgiry S, et al. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Dig Dis Sci*. (2013) 58:936–41. doi: 10.1007/s10620-012-2433-8

32. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: cross-sectional evaluation of chat GPT and perplexity AI in medical residency exam. *Dig Liver Dis Off J Ital Soc Gastroenterol Ital Assoc Study Liver*. (2024) S1590-8658:00302–5.

33. Lim DYZ, Tan YB, Koh JTE, Tung JYM, Sng GGR, Tan DMY, et al. Chat GPT on guidelines: providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals. *J Gastroenterol Hepatol*. (2024) 39:81–106. doi: 10.1111/jgh.16375

34. Samaan JS, Yeo YH, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Chat GPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab J Gastroenterol Off Publ Pan-Arab Assoc Gastroenterol*. (2023) 24:145–8.

35. Wagholikar K, Sohn S, Wu S, Kaggal V, Buehler S, Greenes R, et al. Clinical decision support for colonoscopy surveillance using natural language processing In: 2012 IEEE second international conference on healthcare informatics, imaging and systems biology (2012). 12–21.

36. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. (2014) 12:1130–6. doi: 10.1016/j.cgh.2013.11.025

37. Pereyra L, Schlottmann F, Steinberg L, Lasa J. Colorectal Cancer prevention: is chat generative Pretrained transformer (chat GPT) ready to assist physicians in determining appropriate screening and surveillance recommendations? *J Clin Gastroenterol*. (2024) 58:1022–7. doi: 10.1097/MCG.0000000000001979

38. Kong Q, Ju K, Wan M, Liu J, Wu X, Li Y, et al. Comparative analysis of large language models in medical counseling: a focus on *Helicobacter pylori* infection. *Helicobacter*. (2024) 29:e13055. doi: 10.1111/hel.13055

39. Choo JM, Ryu HS, Kim JS, Cheong JY, Baek SJ, Kwak JM, et al. Conversational artificial intelligence (chat GPT™) in the management of complex colorectal cancer patients: early experience. *ANZ J Surg*. (2024) 94:356–61. doi: 10.1111/ans.18749

40. 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: 10.1186/s12911-023-02121-7

41. Wang X, Xu X, Tong W, Liu Q, Liu Z. Deep causality: a general AI-powered causal inference framework for free text: a case study of liver Tox. *Front Artif Intell*. (2022) 5:5. doi: 10.3389/frai.2022.999289

42. 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 JAMIA*. (2011) 18:i150–6. doi: 10.1136/amiajnl-2011-000431

43. Ma H, Ma X, Yang C, Niu Q, Gao T, Liu C, et al. Development and evaluation of a program based on a generative pre-trained transformer model from a public natural language processing platform for efficiency enhancement in post-procedural quality control of esophageal endoscopic submucosal dissection. *Surg Endosc*. (2024) 38:1264–72. doi: 10.1007/s00464-023-10620-x

44. Huo B, McKechnie T, Ortenzi M, Lee Y, Antoniou S, Mayol J, et al. GPT will see you now: the ability of large language model-linked chatbots to provide colorectal cancer screening recommendations. *Health Technol*. (2024) 14:463–9. doi: 10.1007/s12553-024-00836-9

45. Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, et al. Evaluating AI in medicine: a comparative analysis of expert and chat GPT responses to colorectal cancer questions. *Sci Rep*. (2024) 14:2840. doi: 10.1038/s41598-024-52853-3

46. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep*. (2023) 13:4164. doi: 10.1038/s41598-023-31412-2

47. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? *Diagnostics*. (2023) 13:1950. doi: 10.3390/diagnostics13111950

48. Zhou J, Li T, Fong SJ, Dey N, González-Crespo R. Exploring chat GPT's potential for consultation, recommendations and report diagnosis: gastric Cancer and gastroscopy reports' case. *Int J Interact Multimed Artif Intell*. (2023) 8:7–13. doi: 10.9781/ijimai.2023.04.007

49. Truhn D, Loeffler CM, Müller-Franzes G, Nebelung S, Hewitt KJ, Brandner S, et al. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). *J Pathol*. (2024) 262:310–9. doi: 10.1002/path.6232

50. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc JAMIA*. (2010) 17:383–8. doi: 10.1136/jamia.2010.004804

51. Gorelik Y, Ghersin I, Maza I, Klein A. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. *Gastrointest Endosc*. (2023) 98:639–641.e4. doi: 10.1016/j.gie.2023.06.025

52. Stidham RW, Yu D, Zhao X, Bishu S, Rice M, Bourque C, 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: 10.1093/ibd/izac109

53. Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflamm Bowel Dis*. (2013) 19:1411–20. doi: 10.1097/MIB.0b013e31828133fd

54. Schneider CV, Li T, Zhang D, Mezina AI, Rattan P, Huang H, et al. Large-scale identification of undiagnosed hepatic steatosis using natural language processing. *eClinicalMedicine*. (2023) 62:102149. doi: 10.1016/j.eclinm.2023.102149

55. Benson R, Winterton C, Winn M, Krick B, Liu M, Abu-el-rub N, et al. Leveraging natural language processing to extract features of colorectal polyps from pathology reports for epidemiologic study. *JCO Clin Cancer Inform*. (2023):7. doi: 10.1200/CCI.22.00131

56. Benedicenti F, Pessarelli T, Corradi M, Michelon M, Nandi N, Lampertico P, et al. Mirror, mirror on the wall, who is the best of them all? Artificial intelligence versus gastroenterologists in solving clinical problems. *Gastroenterol Rep*. (2023) 11:goad052.

57. Nayor J, Borges LF, Goryachev S, Gainer VS, Saltzman JR. Natural language processing accurately calculates adenoma and sessile serrated polyp detection rates. *Dig Dis Sci*. (2018) 63:1794–800. doi: 10.1007/s10620-018-5078-4

58. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. (2013) 11:689–94. doi: 10.1016/j.cgh.2012.11.035

59. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. *Gastrointest Endosc*. (2015) 82:512–9. doi: 10.1016/j.gie.2015.01.049

60. Bae JH, Han HW, Yang SY, Song G, Sa S, Chung GE, 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: 10.2196/35257

61. Song G, Chung SJ, Seo JY, Yang SY, Jin EH, Chung GE, et al. Natural language processing for information extraction of gastric diseases and its application in large-scale clinical research. *J Clin Med*. (2022) 11:2967. doi: 10.3390/jcm11112967

62. Li D, Udaltsova N, Layefsky E, Doan C, Corley DA. Natural language processing for the accurate identification of colorectal Cancer mismatch repair status in Lynch syndrome screening. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. (2021) 19:610–612.e1. doi: 10.1016/j.cgh.2020.01.040

63. Denny JC, Choma NN, Peterson JF, Miller RA, Bastarache L, Li M, et al. Natural language processing improves identification of colorectal cancer testing in the electronic medical record. *Med Decis Mak Int J Soc Med Decis Mak*. (2012) 32:188–97. doi: 10.1177/0272989X11400418

64. Becker M, Kasper S, Böckmann B, Jöckel KH, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. *Int J Med Inform*. (2019) 127:141–6. doi: 10.1016/j.ijmedinf.2019.04.022

65. Blumenthal DM, Singal G, Mangla SS, Macklin EA, Chung DC. Predicting non-adherence with outpatient colonoscopy using a novel electronic tool that measures prior non-adherence. *J Gen Intern Med*. (2015) 30:724–31. doi: 10.1007/s11606-014-3165-6

66. Imler TD, Sherman S, Imperiale TF, Xu H, Ouyang F, Beesley C, et al. Provider-specific quality measurement for ERCP using natural language processing. *Gastrointest Endosc*. (2018) 87:164–173.e2. doi: 10.1016/j.gie.2017.04.030

67. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of chat GPT for inflammatory bowel diseases: An analysis for patients and healthcare professionals. *Cureus*. (2023) 15:e46736.

68. Rammohan R, Joy MV, Magam SG, Natt D, Magam SR, Pannikodu L, et al. Understanding the landscape: the emergence of artificial intelligence (AI), chat GPT, and Google bard in gastroenterology. *Cureus*. (2024) 16

69. Nguyen Wenker T, Natarajan Y, Caskey K, Novoa F, Mansour N, Pham HA, et al. Using natural language processing to automatically identify dysplasia in pathology reports for patients with Barrett's esophagus. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. (2023) 21:1198–204. doi: 10.1016/j.cgh.2022.09.005

70. Tinmouth J, Swain D, Chorneyko K, Lee V, Bowes B, Li Y, et al. Validation of a natural language processing algorithm to identify adenomas and measure adenoma detection rates across a health system: a population-level study. *Gastrointest Endosc*. (2023) 97:121–129.e1. doi: 10.1016/j.gie.2022.07.009

71. Sada Y, Hou J, Richardson P, El-Serag H, Davila J. Validation of case finding algorithms for hepatocellular Cancer from administrative data and electronic health records using natural language processing. *Med Care*. (2016) 54:e9–e14. doi: 10.1097/MLR.0b013e3182a30373

72. Wang Y, Huang Y, Nimma IR, Pang S, Pang M, Cui T, et al. Validation of GPT-4 for clinical event classification: a comparative analysis with ICD codes and human reviewers. *J Gastroenterol Hepatol*. (2024) 39:1535–43. doi: 10.1111/jgh.16561

73. Mandrekar JN. Measures of interrater agreement. *J Thorac Oncol*. (2011) 6:6–7. doi: 10.1097/JTO.0b013e318200f983

74. Omar M, Brin D, Glicksberg B, Klang E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: a systematic review. *Am J Infect Control*. (2024)

75. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak*. (2021) 21:179. doi: 10.1186/s12911-021-01533-7

76. Cheng S, Chang C, Chang W, Wang H, Liang C, Kishimoto T, et al. The now and future of chat GPT and GPT in psychiatry. *Psychiatry Clin Neurosci*. (2023) 77:592–6. doi: 10.1111/pcn.13588

77. Oh N, Choi GS, Lee WY. Chat GPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res*. (2023) 104:269–73. doi: 10.4174/astr.2023.104.5.269

78. Wang J, Deng H, Liu B, Hu A, Liang J, Fan L, et al. Systematic evaluation of research Progress on natural language processing in medicine over the past 20 years: bibliometric study on pub med. *J Med Internet Res*. (2020) 22:e16816. doi: 10.2196/16816

79. Poon AIF, Sung JJY. Opening the black box of AI-medicine. *J Gastroenterol Hepatol*. (2021) 36:581–4. doi: 10.1111/jgh.15384

80. Herington J, McCradden MD, Creel K, Boellaard R, Jones EC, Jha AK, et al. Ethical considerations for artificial intelligence in medical imaging: deployment and governance. *J Nucl Med Off Publ Soc Nucl Med*. (2023) 64:1509–15.

81. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. (2023) 27:120. doi: 10.1186/s13054-023-04393-x

82. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Socio-demographic biases in medical decision-making by large language models: a large-scale multi-model analysis. *med Rxiv*. (2024):2024. doi: 10.1101/2024.10.29.24316368v1

83. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *med Rxiv*. (2024). doi: 10.1101/2024.09.09.24313295v2

84. Fevrier HB, Liu L, Herrinton LJ, Li D. A transparent and adaptable method to extract colonoscopy and pathology data using natural language processing. *J Med Syst*. (2020) 44:151. doi: 10.1007/s10916-020-01604-8