Review

Large Language Models in Lung Cancer: Systematic Review

Ruikang Zhong[1], MD; Siyi Chen[1], MD; Zexing Li[1], MD; Tangke Gao[1], MD; Yisha Su[1], MD; Wenzheng Zhang[1], MD; Dianna Liu[2], MD; Lei Gao[2*], MD; Kaiwen Hu[2*], MD

[1]Graduate School, Beijing University of Chinese Medicine, Beijing, China
[2]Oncology Department, Dongfang Hospital, Beijing University of Chinese Medicine, Beijing, China
*these authors contributed equally

Corresponding Author:

Kaiwen Hu, MD
Oncology Department
Dongfang Hospital, Beijing University of Chinese Medicine
No. 6, Fangxingyuan 1st District, Fengtai District
Beijing
China
Phone: 86 13911650713
Email: kaiwenh@163.com

Abstract

Background: In the era of data and intelligence, artificial intelligence has been widely applied in the medical field. As the most cutting-edge technology, the large language model (LLM) has gained popularity due to its extraordinary ability to handle complex tasks and interactive features.

Objective: This study aimed to systematically review current applications of LLMs in lung cancer (LC) care and evaluate their potential across the full-cycle management spectrum.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we conducted a comprehensive literature search across 6 databases up to January 1, 2025. Studies were included if they satisfied the following criteria: (1) journal articles, conference papers, and preprints; (2) studies that reported the content of LLMs in LC; (3) including original data and LC-related data presented separately; and (4) studies published in English. The exclusion criteria were as follows: (1) books and book chapters, letters, reviews, conference proceedings; (2) studies that did not report the content of LLMs in LC; and (3) no original data, and LC-related data that are not presented separately. Studies were screened independently by 2 authors (SC and ZL) and assessed for quality using Quality Assessment of Diagnostic Accuracy Studies-2, Prediction Model Risk of Bias Assessment Tool, and Risk Of Bias in Non-randomized Studies - of Interventions tools, selected based on study type. Key data items extracted included model type, application scenario, prompt method, input and output format, outcome measures, and safety considerations. Data analysis was conducted using descriptive statistics.

Results: Out of 706 studies screened, 28 were included (published between 2023 and 2024). The ability of LLMs to automatically extract medical records, popularize general knowledge about LC, and assist clinical diagnosis and treatment has been demonstrated through the systematic review, emerging visual ability, and multimodal potential. Prompt engineering was a critical component, with varying degrees of sophistication from zero-shot to fine-tuned approaches. Quality assessments revealed overall acceptable methodological rigor but noted limitations in bias control and data security reporting.

Conclusions: LLMs show considerable potential in improving LC diagnosis, communication, and decision-making. However, their responsible use requires attention to privacy, interpretability, and human oversight.

Trial Registration: PROSPERO CRD42024612388; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024612388

J Med Internet Res 2025;27:e74177; doi: 10.2196/74177

Keywords: lung cancer; LC; large language modeling; LLM; artificial intelligence; full-cycle management; clinical practice; systematic review; diagnosis; treatment

Introduction

Lung cancer (LC) is one of the leading causes of cancer incidence and mortality worldwide [1,2]. Early detection and accurate treatment are essential to improving survival [3,4], and low-dose computed tomography (CT) screening has been shown to reduce mortality [5,6]. In recent years, integrated full-cycle management—covering prevention, screening, diagnosis, treatment, and supportive care—has been promoted to improve both survival and quality of life [7,8]. However, this approach requires complex workflows and large-scale data processing, placing heavy demands on medical resources and personnel.

Artificial intelligence, particularly large language models (LLMs), offers a potential solution. LLMs can process complex clinical data, support decision-making, and enable personalized communication between patients and health care providers [9-11]. At the same time, they face limitations such as bias [12] and hallucinations [13]. These issues highlight the need for a systematic evaluation of their role in clinical practice.

Numerous studies have been conducted on LLMs in the field of LC. Some scholars have carried out a systematic review on the potential of LLMs and natural language processing in LC diagnosis [14]. However, it was limited to diagnostic applications, relied on outdated evidence, and lacked a comprehensive scope. This study aims to address these gaps by systematically reviewing the latest applications of LLMs in LC. We summarize current use cases, model types, fine-tuning strategies, limitations, and future directions. Our goal is to help clinicians and researchers better understand how to integrate LLMs into LC management while recognizing their potential and constraints.

Methods

Overview

This study was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [15]. The PRISMA checklist is presented in Checklist 1.

Eligibility Criteria

We established clear inclusion and exclusion criteria based on the research objectives, as summarized in Table 1. No time restrictions were applied during the selection of studies.

Table 1. Inclusion and exclusion criteria.

Criterion	Inclusion	Exclusion
Types of studies	Journal articles, conference papers, and preprints	Books and book chapters, letters, reviews, and conference proceedings
Content	Content involves LLMs[a] and LC[b]	Neither LLMs nor LC
Outcomes	Including original data, and LC-related data are presented separately	No original data, and LC-related data are not presented separately
Language	English	Non-English

[a]LLM: large language model.
[b]LC: lung cancer.

Data Sources

Eligible studies were identified by searching 6 electronic databases: PubMed, Web of Science, IEEE, Embase, Cochrane Library, and Scopus. The final search was run up to January 1, 2025.

Search Strategy

The search strategy was structured as follows: ((“large language model”) OR (“LLM”) OR (“ChatGPT”) OR (“chatGPT”)) AND ((“lung cancer”) OR (“lung tumor”) OR (“pulmonary ground-glass”) OR (“lung malignancy”) OR (“lung carcinoma”) OR (“lung metastasis”) OR (“lung metastatic”) OR (“pulmonary metastatic”) OR (“pulmonary metastasis”)).

Selection Process

EndNote X9.3.3 (build 13966; Clarivate) was used to manage references and remove duplicates. Two authors (RZ and SC) independently screened the titles and abstracts, followed by full-text screening based on the predefined inclusion and exclusion criteria. Discrepancies were resolved through discussion, with arbitration by a third author (ZL) when necessary. The consistency degree of the 2 authors was verified using the kappa consistency test.

Data Collection Process

Two authors (RZ and SC) carried out the data collection process. All extracted data from the main text, tables, figures, and appendices were annotated using WPS Office Excel (version 12.1.0.18608; Kingsoft Office Software).

Data Items

The data extraction form included the following items: title, first author, year of publication, study design, LLM model used, application scenario, intervention, prompt engineering approach, input and output formats, and outcome measures. The consistency rate of the 2 authors was calculated.

Quality Appraisal

To ensure a rigorous evaluation of study quality, we adopted a mixed methods approach based on the framework by Omar and Levkovich [16]. Appropriate quality assessment tools were selected based on the specific application of LLMs in each study. QUADAS-2 (Quality Assessment of Diagnostic

Accuracy Studies-2) [17] is a validated and widely accepted tool for evaluating the quality of diagnostic tests. For studies where LLMs were primarily applied to LC diagnosis or staging, the QUADAS-2 tool was used. PROBAST (Prediction Model Risk of Bias Assessment Tool) [18] is specifically designed to assess the risk of bias in studies involving predictive modeling and was applied accordingly. The ROBINS-I (Risk Of Bias in Non-randomized Studies - of Interventions) [19] tool is commonly used to assess bias in observational studies. For research on information extraction and knowledge-based tasks, these were considered observational in nature, and thus, the ROBINS-I tool was applied. Given that studies involving LLMs differ in format and content from conventional clinical trials, 2 oncology experts at the chief physician level (LG and KH) adapted the criteria of each tool accordingly to better reflect the nature and objectives of the included studies.

The quality assessment was carried out back-to-back by 2 researchers (SC and ZL) and, in the case of controversial content, by a third researcher (RZ) in order to deliberate jointly on the decision. The final results are reviewed by 2 experts (LG and KH). The consistency degree of the 2 authors was verified using the kappa consistency test.

Synthesis Methods

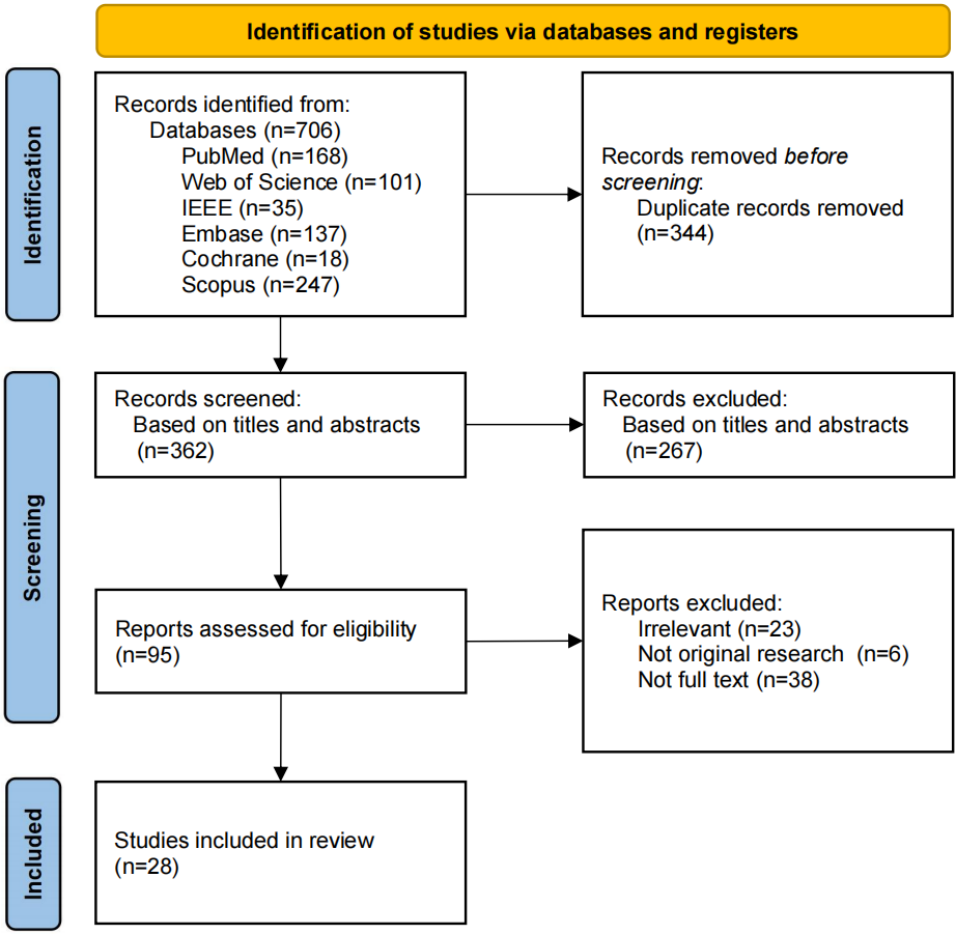
Meta-analysis was not planned in this review. We conducted data analysis using descriptive statistics. Frequencies were used to summarize the application scenarios, prompt strategies, and other relevant characteristics of LLMs. Narrative synthesis was conducted due to the heterogeneity in the specified aims and methodologies across the included studies. We primarily used WPS and the BioRender website for figure generation. We used IBM SPSS (version 29.0.2.0) to calculate the kappa value.

Results

Search Results

In this study, a total of 706 studies were retrieved, and 28 studies [20-47] were finally included after screening. The kappa values of the 2 researchers during the screening stage were 0.87, indicating good consistency. The specific screening process is presented in Figure 1.

Figure 1. Study flowchart (produced according to the PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] 2020 flow diagram).



Basic Information of Included Sources

During the data extraction stage, the consistency rate of the 2 authors reached 0.97. All included studies were published between 2023 and 2024, with 7 published in 2023 [21,23,26,29,31,32,40] and 21 in 2024 [20,22,24,25,27,28,30,33-39,41-47]. Of these, 13 studies originated from the United States [20,23-27,29,30,32,36,43,45,46], followed by 3 each from South Korea [33,42,44], Germany [21,31,34], and China [22,35,39]. The remaining studies were conducted in India [38,47], Turkey [28], Japan [37], Greece [40], and the Netherlands [41]. Publication types included 5 conference papers [23,35,38,40,47] and 4 preprints [24,25,27,39]. The most commonly used LC type was non–small cell lung cancer (NSCLC). Most studies focused on knowledge-based question answering, information extraction, and diagnostic support. The LLMs used varied widely, with frequent use of OpenAI's GPT-3.5, GPT-4, and GPT-4V, Meta AI's LLaMA-2, and Google AI's Bard. A summary of these details is provided in Table 2.

Table 2. Summary of included sources.

Study	Title	Country	Device	Best performance
Information extraction				
Bhattarai et al [20]	Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning–based methods	United States	GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, spaCy	GPT-4
Fink et al [21]	Potential of ChatGPT and GPT-4 for data mining of free-text CT[a] reports on lung cancer	Germany	ChatGPT, GPT-4	GPT-4
Hu et al [22]	Zero-shot information extraction from radiological reports using ChatGPT	China	ChatGPT	—[b]
Naik et al [23]	Applying large language models for causal structure learning in non–small cell lung cancer	United States	NR[c]	—
Niu et al [24]	Cross-institutional structured radiology reporting for lung cancer screening using a dynamic template-constrained large language model	United States	Llama-3.1 (8B, 70B, 405B), Qwen-2 (72B), Mistral-Large (123B)	Llama-3.1 (8B, 70B, 405B)
Lee et al [25]	SEETrials: leveraging large language models for safety and efficacy extraction in oncology clinical trials	United States	GPT-4	—
Lyu et al [26]	Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential	United States	ChatGPT	—
Knowledge-based question and answer evaluation				
Ferrari-Light et al [27]	Evaluating ChatGPT as a patient resource for frequently asked questions about lung cancer surgery–a pilot study	United States	GPT-3.5	—
Gencer [28]	Readability analysis of ChatGPT's responses on lung cancer	Turkey	GPT-3.5-turbo	—
Haver et al [29]	Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening	United States	ChatGPT, GPT 4, Bard	Bard
Janopaul-Naylor et al [30]	Physician assessment of ChatGPT and Bing answers to American Cancer Society's questions to ask about your cancer	United States	GPT-3.5, Bing AI	GPT-3.5
Rogasch et al [31]	ChatGPT: can you prepare my patients for [18F]FDG PET/CT and explain my reports?	Germany	ChatGPT	—
Rahsepar et al [32]	How AI responds to common lung cancer questions: ChatGPT versus Google Bard	United States	GPT-3.5, Google Bard experimental version	GPT-3.5
Auxiliary diagnosis				
Cho et al [33]	Extracting lung cancer staging descriptors from pathology reports: a generative language model approach	Korea	Llama-2-7B, Mistral-7B, Deductive Llama-2-7B (Orca-2), Deductive Mistral-7B (Dolphin),	Deductive Mistral-7B

Study	Title	Country	Device	Best performance
			AWS Llama-2-70B, AWS Titan express	
Dehdab et al [34]	Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment	Germany	GPT-4V	—
Hu et al [35]	The power of combining data and knowledge: GPT-4o is an effective interpreter of machine learning models in predicting lymph node metastasis of lung cancer	China	GPT-4	—
Huang et al [36]	A critical assessment of using ChatGPT for extracting structured data from clinical notes	United States	GPT-3.5-Turbo-16k	—
Yasaka et al [37]	Fine-tuned large language model for extracting patients on pretreatment for lung cancer from a picture archiving and communication system based on radiological reports	Japan	Transformers Japanese model	—
Vallabhaneni et al [38]	Improved lung cancer detection through use of large language systems with graphical attributes	India	NR	—
Qu et al [39]	The rise of AI language pathologists: exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification	China	GPT-4	—
Panagoulias et al [40]	Evaluation of ChatGPT-supported diagnosis, staging and treatment planning for the case of lung cancer	Greece	ChatGPT	—
Mithun et al [41]	Transfer learning with BERT and ClinicalBERT models for multiclass classification of radiology imaging reports	Netherlands	BERT, ClinicalBERT	ClinicalBERT
Lee et al [42]	Lung cancer staging using chest CT and FDG PET/CT free-text reports: comparison among three ChatGPT large-language models and six human readers of varying experience	Korea	GPT-4o, GPT-4, GPT-3.5	GPT-4o
Treatment decision-making				
Dong et al [43]	Large-language-model empowered 3D dose prediction for intensity-modulated radiotherapy	United States	Llama-2	—
Jeong et al [44]	The prediction of stress in radiation therapy: integrating artificial intelligence with biological signals	Korea	Decision tree, random forest, support vector machine, LSTM[d], GPT-4, GPT-3.5	LSTM (limited information); GPT-4 (complex and diverse information)
Aided nursing				
Dos Santos et al [45]	An example of leveraging AI for documentation: ChatGPT-generated nursing care plan for an older adult with lung cancer	United States	ChatGPT	—
Scientific research				
Wang et al [46]	Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception	United States	GPT-4V	—
Devi et al [47]	Automating clinical trial eligibility screening: quantitative analysis of GPT models versus human expertise	India	GPT-3.5-turbo	—

[a]CT: computed tomography.
[b]Not available.
[c]NR: not reported.
[d]LSTM: long short-term memory.

Notably, many studies used multiple LLMs or conducted comparative evaluations, and some explored multimodal capabilities such as image interpretation. The best-performing models identified in these comparative studies are summarized in Table 2. The results indicate that the ChatGPT (OpenAI) series models are the most comprehensive and widely applicable, exhibiting strong performance in both information extraction and auxiliary diagnosis, highlighting the improvements achieved through version updates. However, for a limited number of tasks or under constrained information conditions, lightweight models, such as Bard or architectures like long short-term memory networks may

perform better. In addition, LLMs specialized in the medical domain, such as Deductive Mistral-7B and ClinicalBERT, demonstrate superior performance compared with general-purpose pretrained models.

Prompt Engineering and Model Training

Prompt engineering plays a critical role in the development and application of LLMs and is a frequent topic of discussion in related studies. Therefore, we synthesized and summarized the prompt engineering strategies, model inputs and outputs, and evaluation metrics used in the included studies (Table 3). In total, 12 (43%) studies [24,27-30,32,34,37,38,41,44,47] did not explicitly describe their prompting strategies, which were generally basic queries, primarily intended for educational use. Furthermore, 16 (57%) studies [20-23,25,26,31,33,35,36,39,40,42,43,45,46] clearly described their prompting methods. These methods included prompt templates, instructional prompts, zero-shot or few-shot learning, and other fine-tuning techniques. Regarding the types of training data, a total of 22 (79%) studies [20-23,25,26,31,33,35,36,39,40,42,43,45,46] focused on text, 3 (11%) studies [24,34,43] on images, and 3 (11%) studies [38,42,46] on a combination of images and text. Outcome metrics commonly included confusion matrices, rating scales, and comparisons against gold-standard references or expert consensus. Some studies also reported on the time efficiency and cost-effectiveness of LLM-generated outputs.

Table 3. Prompt engineering and model training.

Study	Prompt method or content	Model input	Model output	Outcome indicators
Information extraction				
Bhattarai et al [20]	Zero-shot prompt	Segmented text and zero-shot prompt	Phenotypic information (cancer staging, cancer treatment), evidence of cancer recurrence, and organs affected by cancer recurrence	Accuracy, recall rate, F_1-score, generation time, operating costs
Fink et al [21]	25 original lung cancer CT[a] reports used to prompt training	Original lung cancer CT reports	Tumor information includes tumor lesions, metastatic sites, tumor impression assessment (deterioration, stability, improvement), and interpretation	McNemar test, accuracy, 5-point Likert scale
Hu et al [22]	Prompt template, including an information extraction command, a question form, extraction requirements, and some relevant medical knowledge	CT reports and prompt template	Answers to the question form	Accuracy, precision, recall rate, and F_1-score
Naik et al [23]	Code interpreter plugin (developed by OpenAI)	Electronic medical records, genomic data	Directed acyclic graph	Bdeu score
Niu et al [24]	Not mentioned	CT imaging	Standardized and structured radiological reports	F_1-score, CI, McNemar test, and z test
Lee et al [25]	Prompt templates	Journal abstract	Details of clinical trials in the article	Accuracy, recall rate, F_1-score
Lyu et al [26]	Instruction	Radiological reports	Report translation and suggestions	Self score, report completeness and accuracy
Knowledge-based question and answer evaluation				
Ferrari-Light et al [27]	Not mentioned	Questions	Answers	5-point Likert scale
Gencer [28]	Not mentioned	Questions	Answers	Flesch Reading Ease (FRE) formula, Flesch-Kincaid Grade level (FKGL), Gunning FOG formula, SMOG index, Automated readability index (ARI), Coleman-Liau index, Linsear write formula, Dale-Chall readability score, Spache readability formula

Study	Prompt method or content	Model input	Model output	Outcome indicators
Haver et al [29]	Not mentioned	Questions	Baseline responses and simplified responses	Reading Ease Score, readability, clinical appropriateness
Janopaul-Naylor et al [30]	Not mentioned	Questions	Answers	Self rating
Rogasch et al [31]	Regeneration-response function repeated three times for training	Questions	Answers	Self rating
Rahsepar et al [32]	Not mentioned	Questions	Answers	Accuracy, consistency
Auxiliary diagnosis				
Cho et al [33]	Morphology group	Segmented pathological report	42 lung cancer staging descriptors; tumor node classification	Macro F_1-score, accurate matching ratio, accuracy
Dehdab et al [34]	Not mentioned	CT images of lung window	Diagnosis of lung cancer (yes or no)	Accuracy, sensitivity, specificity
Hu et al [35]	Prompt templates, including roles, tasks, patient data, machine learning model results and instructions	Prompt templates	Prediction results of lymph node metastasis in lung cancer	AUC[b], AP (average precision of 3 repetitions)
Huang et al [36]	Prompt templates, including clinical staging introduction and instructions	Pathology reports and prompt templates	Tumor size, tumor characteristics, lymph node involvement, histological classification, clinical staging	Accuracy, average precision, F_1-score, Kappa, recall rate
Yasaka et al [37]	Not mentioned	Clinical indications and diagnosis of radiological reports	Patient grouping (Group 0: no lung cancer, Group 1: lung cancer pre-treatment present, Group 2: after lung cancer treatment, Group 3: planned radiotherapy)	Overall accuracy, sensitivity, consistency, AUC, classification time
Vallabhaneni et al [38]	Not mentioned	Images, symptoms, clinical prescriptions	Diagnosis of lung cancer (yes or no)	Accuracy, recall rate, F_1-score, AUC
Qu et al [39]	Guide GPT-4 to visually describe complex medical concepts	Questions (text)	Answers	AUC
Panagoulias et al [40]	Build and refine prompts based on the returned answers	Symptom description	Diagnosis and treatment plan for lung cancer	Self-drafted standards
Mithun et al [41]	Not mentioned	Radiological reports	Classification results of lung cancer	AUC, F_1-score, accuracy, recall rate, precision
Lee et al [42]	Instruction	Chest CT and FDG PET[c] or CT reports	The maximum size of the primary tumor, local invasion, satellite lesions, metastatic lymph nodes, intrathoracic and extrathoracic metastases, and TNM[d] staging diagnosis	Accuracy, recall rate, F_1-score, average task completion time, misreading rate
Treatment decision-making				
Dong et al [43]	Clinical physician commands (findings, treatment goals, and precautions)	CT images	DVH (Radiation dose volume histogram)	Mean absolute error (MAE) of Dmax, Dmean, D95, and D1 between actual and predicted plans
Jeong et al [44]	Not mentioned	Biological signals before radiotherapy and instructions	Prediction results of biological signals and stress response during radiotherapy	Accuracy, recall rate, precision, F_1-score
Aided nursing				

Study	Prompt method or content	Model input	Model output	Outcome indicators
Dos Santos et al [45]	Patient's needs framework (Situation or Background, Physical, Safety, Psychosocial, Spiritual or Culture, Nursing Recommendation)	Medical records, needs framework, problem prompts	Care plan	The number of items that match the gold standard (16 tags including NANDA, NOC, and NIC)
Scientific research				
Wang et al [46]	Instruction	K-M[e] curves generated based on gene expression data and survival information	Analysis and Interpretation of K-M curves	Overall accuracy, Accuracy under each category
Devi et al [47]	Not mentioned	Unprocessed raw dataset	Whether the patient is qualified for enrollment (yes or no)	Accuracy compared with manual classification

[a]CT: computed tomography.
[b]AUC: area under the curve.
[c]FDG PET: Fluorodeoxyglucose positron emission tomography.
[d]TNM: tumor, nodes, metastasis.
[e]K-M: Kaplan Meier.

Quality Appraisal

The included studies were categorized based on their research objectives, and quality was assessed using corresponding appraisal tools (Multimedia Appendix 1). The kappa values of the 2 researchers were 0.84. Furthermore, 3 predictive modeling studies [35,43,44] were evaluated using the PROBAST tool (Figure 2A). These studies showed low risk of bias regarding data sources, populations, and methodologies but exhibited a potentially high risk in predictor and outcome domains. In total, 10 diagnostic studies [33,34,36-39,41,42,47] were assessed using QUADAS-2 (Figure 2B). While most demonstrated good applicability, the overall risk of bias remained unclear. Furthermore, 16 intervention studies [20-32,40,45,46] were appraised using the ROBINS-I tool (Figure 2C), showing low risk of bias in participant selection and intervention assignment, but unclear or high risk in other domains. Among them, 29% (18/63) of conference papers and preprints have a high-risk or unclear bias risk, while 26% (34/133) of journal papers have a high risk or unclear bias risk.

Figure 2. (A) The quality appraisal for 3 predictive studies with PROBAST (Prediction model Risk Of Bias Assessment Tool). (B) The quality appraisal for 9 diagnostic studies with QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2). (C) The quality appraisal for 16 intervention trials with ROBINS-I (Risk Of Bias In Non-randomized Studies - of Interventions).

Other Aspects

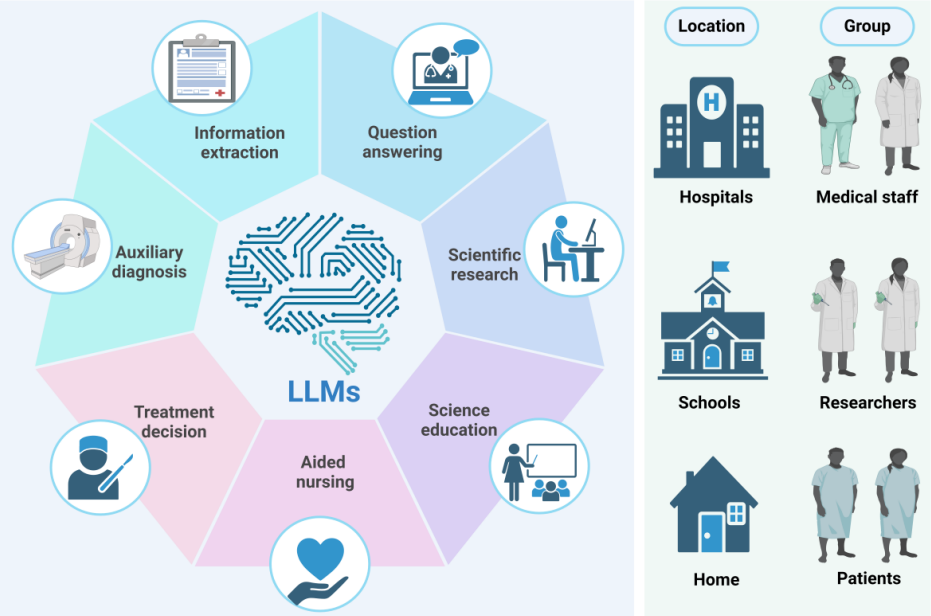
In addition, we examined whether the included studies reported human oversight, addressed safety considerations, and acknowledged limitations. In total, 26 (93%) studies [20-22,24,26-47] reported human involvement in system design, operation, or evaluation. Only 6 (21%) studies [20,31,33,38,42,43] explicitly addressed issues related to information security or data privacy. Furthermore, 20 (71%) studies [20-26,31-33,35,37-43,46,47] clearly stated their limitations.

Discussion

Principal Findings

Through a systematic review of 28 studies [20-47], we identified 7 primary application domains of LLMs in LC: auxiliary diagnosis, information extraction, question answering, scientific research, medical education, nursing support, and treatment decision-making (Figure 3). These domains often overlap in real-world practice—for instance, information extraction frequently supports diagnostic processes, while question-answering is commonly applied in science communication and patient education (Table 2).

Figure 3. Applications of large language models in lung cancer. LLM: large language model.



Applications of LLMs in LC

LLMs can extract clinical features by applying natural language processing methods. Therefore, many studies have used LLMs to extract and analyze information from electronic medical records [48], CT reports [21,22], and pathological reports [33,36] related to LC. This not only enables the diagnosis of clinical staging, histological type, lung-RADS (Reporting and Data System) score, and metastasis sites of LC, but also leverages their reasoning ability for diagnosis and prediction, such as lymph node metastasis [35] and malignancy degree of lung nodules [49]. This highlights the potential of LLMs in LC diagnosis, especially for early screening. Early diagnosis of LC can effectively improve survival rates [50], and mass LC screening achieves a high detection rate of early-stage LC [51], but it is time-consuming and labor-intensive. Ding et al [52] applied ChatGPT to automatically generate medical records during lung nodule screening sessions and integrated it into a WeChat (Tencent Holdings Limited) applet to streamline the consultation process. Singh et al [53] applied ChatGPT and Gemini (Google AI) to generate lung-RADS scores based on low-dose CT reports for LC screening, achieving up to 83.6% accuracy. A systematic review of LLMs in gastroenterology [54] similarly demonstrated the potential applications of LLMs in gastrointestinal endoscopy and the screening of precancerous lesions. Although LLMs still face challenges, such as insufficient extraction performance for complex tasks and hallucinations [55], the results of the study by Jong et al [42] also indicate that using LLMs in place of medical professionals for LC staging is not currently supported. However, with ongoing updates to training data and continuous upgrading and optimization of LLMs, we remain optimistic about their future performance in assisting with LC diagnosis and early screening.

Given the interactive nature and vast data reserves of LLMs, many studies have evaluated their application in

knowledge question answering [27-32]. They have been widely applied in disseminating general knowledge about LC. With the refinement and diversification of training data and the development of multimodal large models, LLMs have shown improved capabilities in processing visual information [56]. Under carefully designed prompts and instructions, several studies have found that LLMs can perform preliminary analyses of medical images and textual data and, within controlled research settings, offer diagnostic and therapeutic suggestions for LC. Examples include providing initial recommendations for subsequent treatment options in newly diagnosed or suspected patients with NSCLC [57], generating more detailed chemotherapy [58] or radiotherapy [59] plans, and predicting outcome indicators such as overall survival [60] and radiotherapy-induced stress responses [44], thereby assisting treatment and nursing decision-making in research contexts [45]. Furthermore, LLMs pretrained on multilingual corpora have demonstrated potential in transcribing or translating LC radiology reports [61] and surgical records [62] to support multicenter clinical research. It should be noted, however, that most existing evidence is derived from retrospective analyses or small-sample, single-center studies. Robust prospective, multicenter clinical validation remains lacking, and systematic assessments of model interpretability, bias, and safety are still insufficient. Therefore, the reliability and generalizability of these methods in routine clinical practice require further confirmation.

The natural language processing and named entity recognition capabilities of LLMs can not only benefit clinicians and patients in clinical practice but also improve researchers' efficiency. Devi et al [47] used GPT-3.5-turbo to classify patients with NSCLC based on pathological reports to determine their eligibility for clinical trials, assisting researchers with eligibility screening. Kyeryoung et al [25] used GPT-4 to extract safety and efficacy information from clinical trial abstracts and convert it into computable data for comparative analysis across large clinical trial datasets. Liu et al [63] used LLaMA 3.1 (Meta AI) to generate clinical trial annotations, enabling oncologists to stay fully updated with the latest oncology data presented at medical conferences and in journal publications. Similarly, Yuan et al [64] constructed and evaluated 3 machine learning models for predicting LC survival using an LLM-based advanced data analysis approach, making advanced analytics accessible to nontechnical health care professionals.

From the above, it is evident that the current applications of LLMs in LC span multiple stages of care, from early screening and diagnosis to treatment planning, patient follow-up, and research support. However, their maturity, evidence base, and clinical readiness vary substantially. Diagnostic and screening tools are the most developed, yet most rely on retrospective datasets and single-center studies, with limited prospective, multicenter clinical validation. Similarly, treatment planning applications show promise in integrating patient-specific data with clinical guidelines, but they also lack large-scale, prospective evaluations to confirm safety, effectiveness, and adaptability to evolving oncology standards. Patient follow-up and supportive care applications are even less developed, despite their potential to improve adherence, symptom management, and long-term quality of life. These stages are often complex due to diverse patient needs, variable follow-up schedules, and sensitive data management requirements, which may explain their slower technological adoption. Research-support tools, such as automated trial eligibility screening or survival prediction, demonstrate potential for improving efficiency, but their accuracy and reproducibility in real-world practice remain uncertain.

Based on these observations, we identify 3 research priorities. First, rigorous prospective, multicenter clinical validation of both diagnostic or screening and treatment planning applications to ensure generalizability and safety. Second, targeted development of patient follow-up and supportive care applications to address gaps in long-term management and patient engagement. Third, improvement of model interpretability, bias mitigation, and integration strategies to enable safe deployment across diverse health care systems. Addressing these gaps will be essential for the effective integration of LLMs into full-cycle LC management.

Limitations of LLMs and Future Directions in LC

Clinical decision-making for LC in practice is driven by multimodal data, including clinical notes, radiological images, and pathological features. This implies that artificial intelligence tools capable of effectively integrating multimodal data hold significant potential for advancing clinical treatment of LC [65]. However, the research reviewed in this article still primarily focuses on text processing. Although efforts have been made to explore other data modalities, including CT images [34], pathological images [39], and bioinformatics data [46], the accuracy of their outputs has yet to match that of text-based outputs. Furthermore, studies indicate that deep learning models specialized in image processing, such as Convolutional Neural Networks, outperform LLMs in classifying LC cytology images [66]. Therefore, researchers tend to combine LLMs with other deep learning models for multimodal data analysis [38,43]. Nevertheless, the development and advancement of multimodal LLMs remain a key trend. Currently, OpenAI has taken the lead by launching ChatGPT-4o and ChatGPT-4V, spearheading the application boom of multimodal LLMs. In the future, LLMs are expected to overcome single-modality limitations on a large scale and enhance accurate diagnosis and treatment of LC by integrating multimodal reasoning capabilities across medical images, genomic data, biological molecular information, and even audio and video.

Existing research on LC predominantly uses general LLMs, such as ChatGPT and LLAMA-2, which are trained on public databases and experience slow knowledge base updates. These models may have gaps in domain-specific LC knowledge, and their outputs are prone to hallucinations and insufficient citations [67]. In recent years, many large models targeting specific tasks within clinical specialties have also emerged. For example, Med-PaLM 2 [68], which excels at lengthy medical question-answering;

BioBERT [69], which specializes in biomedical texts; and ClinicalBERT [70], which focuses on clinical texts. However, studies have found that their performance on cardiac surgery knowledge quizzes [71] and precision LC treatment plans [72] is inferior to that of general LLMs with larger training parameter counts. Retrieval-augmented generation (RAG), a cutting-edge technology in large models, can reference reliable external knowledge (REK) to generate answers or content, enable real-time knowledge updates, and offer strong interpretability and customization capabilities [73]. Combining this technology with general large-scale models has resulted in more satisfactory outcomes. Built on Google's Gemini Pro LLM, MEREDITH uses RAG and chain-of-thought reasoning. MEREDITH was enhanced to incorporate clinical studies on drug response within specific tumor types, trial databases, drug approval status, and oncologic guidelines. The precise treatment recommendations it provides for tumors closely align with expert advice [74]. Tozuka et al [75] summarized the current LC staging guidelines in Japan and supplied these as REK to NotebookLM, a RAG-equipped LLM. NotebookLM achieved 86% diagnostic accuracy in LC staging experiments, outperforming GPT-4o, which recorded 39% accuracy with REK and 25% without. In addition, appropriate prompt engineering can enhance the performance of general-purpose LLMs on specific tasks. Most of the studies included in our review used directives, prompt templates, and fine-tuning. Prompt templates often incorporated role descriptions, case examples, task requirements, LC-specific knowledge, and formatting instructions. Fine-tuning involves retraining a pretrained LLM (eg, ChatGPT and BERT) using labeled data for a specific task or domain to improve its performance on domain-specific tasks [76]. A study by Arzideh et al [77], comparing the extraction of clinical entities from unstructured medical records of patients with LC, found that a fine-tuned BERT model using annotated data achieved a higher F_1-score than an instruction-based LLM. Similarly, Zhu et al [78] developed an open-source, oncology-specific LLM using a stacked alignment and fine-tuning process, which outperformed ChatGPT on medical benchmarks and achieved an area under the receiver operating characteristic curve of 0.95 for LC detection.

The studies included in this paper all used open-source LLMs; however, when deploying open-source LLMs in the cloud, issues related to data security and privacy protection are inevitable. Only 6 studies[20,31,33,38,42,43] have explicitly proposed specific data security measures, including legal constraints, such as the Health Insurance Portability and Accountability Act (HIPAA) [20] or standard protocols [38], data access restrictions [33], and data anonymization [41,43]. With the widespread adoption of LLMs in medical settings and growing awareness of data security, hospitals with significant application demands opt to deploy open-source LLMs locally, enabling models and data to operate entirely within the hospital intranet and thereby avoiding risks associated with cloud transmission [79]. They also mitigate data leakage risks through methods such as data anonymization and deidentification [80], federated learning [81,82], and differential privacy [83], among others. In the future, continued technological advancements and regulatory improvements, strengthened data supervision mechanisms, and a balanced approach between cost and performance will be essential to protect patient privacy.

At the same time, it should be acknowledged that LLMs cannot fully replace medical professionals, and it is necessary to clarify the responsibility attribution of LLMs in real clinical scenarios. Ethical frameworks should be established based on the needs of different medical scenarios and acceptable thresholds for patients and applied in a targeted manner [84]. Key applications with low risk of harm to patients' health can be prioritized, such as patient registration codes [85], screening [37,86], and extraction of key information from medical records [87]. Through a "human-on-the-loop" human-machine collaboration model, reinforcement learning techniques are introduced to optimize prompt strategies and model decisions, enhance model transparency and clinician engagement, and strengthen human oversight.

Limitations of This Systematic Review

This review includes studies published up to January 1, 2025. Due to the rapid development of LLMs and fast publication cycles, some recent findings may have been missed. To address this, we expedited manuscript preparation and included several additional studies from the past 7 months (from January to July 2025) in the discussion. To ensure the comprehensiveness and relevance of this review, we included all studies that provided complete data and full-text availability. However, some of the included conference papers and preprints may not have undergone peer review, potentially affecting the reliability of the findings. Nonetheless, our quality assessment indicated that their risk of bias did not differ significantly from that of peer-reviewed journal articles. Although 6 databases were searched, relevant studies outside these sources may have been overlooked. We also limited inclusion to English-language articles, which may affect generalizability, although only 2 non-English articles were excluded. Meanwhile, research conducted across different countries may be affected by population diversity and bias in training datasets. Unlike traditional reviews of clinical interventions, this study applied different quality assessment methods tailored to various application scenarios. Some criteria relied on subjective judgment, and the complexity of the process may have introduced bias. To minimize this, 2 researchers (SC and ZL) assessed studies independently, discrepancies were resolved with a third reviewer, and final decisions were validated by 2 experts (LG and KH).

Conclusions

In summary, this systematic review offers an overview of the applications and research involving LLMs in LC, accompanied by a quality assessment. LLMs can assist physicians in interpreting test reports, delivering diagnostic and treatment recommendations, and supporting education, research, and public outreach efforts. The development of multimodal models, data quality, privacy-preserving mechanisms, and advanced LLM architectures is key to integrating these technologies into the full-cycle management of LC care.

Within an ethical framework and under appropriate human oversight, future efforts should focus on validating LLM applications in real-world clinical settings and the inclusion of underrepresented populations to ensure population diversity, ultimately promoting their development toward greater specialization, accuracy, and patient-centeredness.

Acknowledgments

This work was supported by the Beijing University of Chinese Medicine East Hospital and Beijing Municipal Science & Technology Commission. This research was funded by National Key R&D Program of China (2024YFC3505400), The Science and Technology Plan Project of Beijing (grant Z221100003522029 and grant Z241100007724010), Education Science Research Project, National High Level Chinese Medicine Hospital Clinical Research Funding (DFRCZY-2024GJRC009 and DFRCZY-2024GJRC017).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

KH and LG conceived the study and contributed equally. RZ and SC collected the data and wrote the manuscript. RZ, SC, and ZL extracted information and conducted quality assessment. TG and YS added references and enriched the discussion section. DL and WZ polished the English content of the manuscript. RZ, SC, LG, and KH revised and reviewed the manuscript.
Gao Lei and Hu Kaiwen are co-corresponding authors and contributed equally to this work.
Ruikang Zhong and Siyi Chen contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Specific quality evaluation items and results.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 1]

Checklist 1

PRISMA checklist.
[DOCX File (Microsoft Word File), 273 KB-Checklist 1]

References

1. Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. CA Cancer J Clin. 2025;75(1):10-45. [doi: 10.3322/caac.21871] [Medline: 39817679]
2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
3. Thiruvengadam R, Singh CD, Kondapavuluri BK, Gurusamy S, Venkidasamy B, Thiruvengadam M. Biomarkers in lung cancer treatment. Clin Chim Acta. May 15, 2025;572:120267. [doi: 10.1016/j.cca.2025.120267] [Medline: 40154724]
4. Peeters S, Lau K, Stefanidis K, et al. New diagnostic and nonsurgical local treatment modalities for early stage lung cancer. Lung Cancer (Auckl). Oct 2024;196:107952. [doi: 10.1016/j.lungcan.2024.107952] [Medline: 39236577]
5. The National Lung Screening Trial Research Team, Adams AM, Aberle DR. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. Aug 4, 2011;365(5):395-409. [doi: 10.1056/NEJMoa1102873] [Medline: 31995683]
6. Jonas DE, Reuland DS, Reddy SM, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US preventive services task force. JAMA. Mar 9, 2021;325(10):971-987. [doi: 10.1001/jama.2021.0377] [Medline: 33687468]
7. Kim HJ, Lee MK. Effectiveness of nursing interventions based on lung cancer trajectory: a systematic review and meta-analysis. Int Nurs Rev. Sep 2025;72(3):e13074. [doi: 10.1111/inr.13074] [Medline: 39604008]
8. Xiang R, Li Q. Development status and thinking of the "integrated diagnosis and treatment, full-course management" model of lung cancer- based on the experience of the lung cancer MDT team of Sichuan Cancer Hospital. Zhongguo Fei Ai Za Zhi. Apr 20, 2020;23(4):211-215. [doi: 10.3779/j.issn.1009-3419.2020.101.12] [Medline: 32316710]
9. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. Nov 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
10. Zhao W, Li J, Zhou K, et al. A survey of large language models. arXiv. Preprint posted online on Mar 31, 2023. [doi: 10.48550/arXiv.2303.18223]

11. Turner JH. Triangle of trust in cancer care? The physician, the patient, and artificial intelligence chatbot. Cancer Biother Radiopharm. Nov 2023;38(9):581-584. [doi: 10.1089/cbr.2023.0112] [Medline: 37707991]
12. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. Jan 2024;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
13. Athaluri SA, Manthena SV, Kesapragada V, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. Apr 2023;15(4):e37432. [doi: 10.7759/cureus.37432] [Medline: 37182055]
14. Garg A, Gupta S, Vats S, Handa P, Goel N. Prospect of large language models and natural language processing for lung cancer diagnosis: a systematic review. Expert Systems. Nov 2024;41(11):e13697. URL: https://onlinelibrary.wiley.com/toc/14680394/41/11 [Accessed 2025-09-17] [doi: 10.1111/exsy.13697]
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: a systematic review. J Affect Disord. Feb 15, 2025;371:234-244. [doi: 10.1016/j.jad.2024.11.052] [Medline: 39581383]
17. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18, 2011;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
18. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. Jan 1, 2019;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
19. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. Oct 12, 2016;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
20. Bhattarai K, Oh IY, Sierra JM, et al. Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning-based methods. JAMIA Open. Oct 2024;7(3):ooae060. [doi: 10.1093/jamiaopen/ooae060] [Medline: 38962662]
21. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. Sep 2023;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
22. Hu D, Liu B, Zhu X, Lu XD, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. Int J Med Inform. Mar 2024;183:105321. [doi: 10.1016/j.ijmedinf.2023.105321] [Medline: 38157785]
23. Naik N, Khandelwal A, Joshi M, et al. Applying large language models for causal structure learning in non small cell lung cancer. Presented at: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI); Jun 3, 2024; Orlando, FL, USA. [doi: 10.1109/ICHI61247.2024.00110]
24. Niu C, Kaviani P, Lyu Q, Kalra MK, Whitlow CT, Wang G. Cross-institutional structured radiology reporting for lung cancer screening using a dynamic template-constrained large language model. arXiv. Preprint posted online on Sep 26, 2024. [doi: 10.48550/arXiv.2409.18319]
25. Lee K, Paek H, Huang LC, et al. SEETrials: leveraging large language models for safety and efficacy extraction in oncology clinical trials. Inform Med Unlocked. 2024;50:101589. [doi: 10.1016/j.imu.2024.101589] [Medline: 39493413]
26. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. May 18, 2023;6(1):9. [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
27. Ferrari-Light D, Merritt RE, D'Souza D, et al. Evaluating ChatGPT as a patient resource for frequently asked questions about lung cancer surgery-a pilot study. J Thorac Cardiovasc Surg. Apr 2025;169(4):1174-1180. [doi: 10.1016/j.jtcvs.2024.09.030] [Medline: 39326732]
28. Gencer A. Readability analysis of ChatGPT's responses on lung cancer. Sci Rep. Jul 26, 2024;14(1):17234. [doi: 10.1038/s41598-024-67293-2] [Medline: 39060365]
29. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. AJR Am J Roentgenol. Nov 2023;221(5):701-704. [doi: 10.2214/AJR.23.29622] [Medline: 37341179]
30. Janopaul-Naylor JR, Koo A, Qian DC, McCall NS, Liu Y, Patel SA. Physician assessment of ChatGPT and Bing answers to American Cancer Society's questions to ask about your cancer. Am J Clin Oncol. Jan 1, 2024;47(1):17-21. [doi: 10.1097/COC.0000000000001050] [Medline: 37823708]
31. Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: can you prepare my patients for [^{18}F]FDG PET/CT and explain my reports? J Nucl Med. Dec 1, 2023;64(12):1876-1879. [doi: 10.2967/jnumed.123.266114] [Medline: 37709536]

32. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology. Jun 2023;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
33. Cho H, Yoo S, Kim B, et al. Extracting lung cancer staging descriptors from pathology reports: a generative language model approach. J Biomed Inform. Sep 2024;157:104720. [doi: 10.1016/j.jbi.2024.104720] [Medline: 39233209]
34. Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. Jpn J Radiol. Oct 2024;42(10):1168-1177. [doi: 10.1007/s11604-024-01606-3] [Medline: 38867035]
35. Hu D, Liu B, Zhu X, Wu N. The power of combining data and knowledge: GPT-4o is an effective interpreter of machine learning models in predicting lymph node metastasis of lung cancer. arXiv. Preprint posted online on Jul 25, 2024. [doi: 10.48550/arXiv.2407.17900]
36. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med. May 1, 2024;7(1):106. [doi: 10.1038/s41746-024-01079-8] [Medline: 38693429]
37. Yasaka K, Kanzawa J, Kanemaru N, Koshino S, Abe O. Fine-tuned large language model for extracting patients on pretreatment for lung cancer from a picture archiving and communication system based on radiological reports. J Imaging Inform Med. Feb 2025;38(1):327-334. [doi: 10.1007/s10278-024-01186-8] [Medline: 38955964]
38. Vallabhaneni GV, Rahul YS, Kumari KS. Improved lung cancer detection through use of large language systems with graphical attributes. Presented at: 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT); Feb 9-10, 2024; Greater Noida, India. May 2024.[doi: 10.1109/IC2PCT60090.2024.10486290]
39. Qu LH, Luo XY, Fu KX, Wang MN, Song ZJ. The rise of AI language pathologists: exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. arXiv. Preprint posted online on May 29, 2023. [doi: 10.48550/arXiv.2305.17891]
40. Panagoulias DP, Palamidas FA, Virvou M, Tsihrintzis GA. Evaluation of chatgpt-supported diagnosis, staging and treatment planning for the case of lung cancer. Presented at: 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA); Dec 4-7, 2023; Giza, Egypt. Nov 2024.[doi: 10.1109/AICCSA59173.2023.10479348]
41. Mithun S, Sherkhane UB, Jha AK, et al. Transfer learning with BERT and ClinicalBERT models for multiclass classification of radiology imaging reports. Preprint posted online on Jul 22, 2024. [doi: 10.21203/rs.3.rs-4443132/v1]
42. Lee JE, Park KS, Kim YH, Song HC, Park B, Jeong YJ. Lung cancer staging using chest CT and FDG PET/CT free-text reports: comparison among three ChatGPT large language models and six human readers of varying experience. AJR Am J Roentgenol. Dec 2024;223(6):e2431696. [doi: 10.2214/AJR.24.31696] [Medline: 39230409]
43. Dong Z, Chen Y, Gay H, et al. Large-language-model empowered 3D dose prediction for intensity-modulated radiotherapy. Med Phys. Jan 2025;52(1):619-632. [doi: 10.1002/mp.17416] [Medline: 39316523]
44. Jeong S, Pyo H, Park W, Han Y. The prediction of stress in radiation therapy: integrating artificial intelligence with biological signals. Cancers (Basel). May 22, 2024;16(11):1964. [doi: 10.3390/cancers16111964] [Medline: 38893087]
45. Dos Santos FC, Johnson LG, Madandola OO, et al. An example of leveraging AI for documentation: ChatGPT-generated nursing care plan for an older adult with lung cancer. J Am Med Inform Assoc. Sep 1, 2024;31(9):2089-2096. [doi: 10.1093/jamia/ocae116] [Medline: 38758655]
46. Wang J, Ye Q, Liu L, Guo NL, Hu GQ. Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. NPJ Precis Oncol. Apr 5, 2024;8(1):84. [doi: 10.1038/s41698-024-00576-z] [Medline: 38580746]
47. Devi A, Uttrani S, Singla A, et al. Automating clinical trial eligibility screening: quantitative analysis of GPT models versus human expertise. Presented at: 17th International Conference on PErvasive Technologies Related to Assistive Environments; Jun 26, 2024; Crete Greece. [doi: 10.1145/3652037.3663922]
48. Ashofteh Barabadi M, Zhu X, Chan WY, Simpson AL, Do RKG. Targeted generative data augmentation for automatic metastases detection from free-text radiology reports. Front Artif Intell. 2025;8:1513674. [doi: 10.3389/frai.2025.1513674] [Medline: 39981192]
49. Mao Y, Xu N, Wu Y, et al. Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images. Cell Rep Med. Mar 18, 2025;6(3):101988. [doi: 10.1016/j.xcrm.2025.101988] [Medline: 40043704]
50. Chansky K, Detterbeck FC, Nicholson AG, et al. The IASLC Lung Cancer Staging Project: external validation of the revision of the TNM stage groupings in the eighth edition of the TNM classification of lung cancer. J Thorac Oncol. Jul 2017;12(7):1109-1121. [doi: 10.1016/j.jtho.2017.04.011] [Medline: 28461257]
51. Bhamani A, Creamer A, Verghese P, et al. Low-dose CT for lung cancer screening in a high-risk population (SUMMIT): a prospective, longitudinal cohort study. Lancet Oncol. May 2025;26(5):609-619. [doi: 10.1016/S1470-2045(25)00082-8]
52. Ding H, Xia W, Zhou Y, et al. Evaluation and practical application of prompt-driven ChatGPTs for EMR generation. NPJ Digit Med. Feb 2, 2025;8(1):77. [doi: 10.1038/s41746-025-01472-x] [Medline: 39894840]

53. Singh R, Hamouda M, Chamberlin JH, et al. ChatGPT vs. Gemini: comparative accuracy and efficiency in Lung-RADS score assignment from radiology reports. Clin Imaging. May 2025;121:110455. [doi: 10.1016/j.clinimag.2025.110455] [Medline: 40090067]
54. Gong EJ, Bang CS, Lee JJ, et al. Large language models in gastroenterology: systematic review. J Med Internet Res. Dec 20, 2024;26:e66648. [doi: 10.2196/66648] [Medline: 39705703]
55. Huang L, Yu WJ, Ma WT, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv. Preprint posted online on Nov 19, 2024. [doi: 10.48550/arXiv.2311.05232]
56. Schulze Buschoff LM, Akata E, Bethge M, Schulz E. Visual cognition in multimodal large language models. Nat Mach Intell. 2025;7(1):96-106. [doi: 10.1038/s42256-024-00963-y]
57. Zabaleta J, Aguinagalde B, Lopez I, et al. Utility of artificial intelligence for decision making in thoracic multidisciplinary tumor boards. J Clin Med. Jan 10, 2025;14(2):399. [doi: 10.3390/jcm14020399] [Medline: 39860405]
58. Brown EDL, Shah HA, Donnelly BM, Ward M, Vojnic M, D'Amico RS. Precision oncology in non-small cell lung cancer: a comparative study of contextualized ChatGPT models. Cureus. Mar 2025;17(3):e81097. [doi: 10.7759/cureus.81097] [Medline: 40271313]
59. Wang Q, Wang Z, Li M, et al. A feasibility study of automating radiotherapy planning with large language model agents. Phys Med Biol. Mar 21, 2025;70(7). [doi: 10.1088/1361-6560/adbff1] [Medline: 40073507]
60. Paolo D, Greco C, Cortellini A, et al. Hierarchical embedding attention for overall survival prediction in lung cancer from unstructured EHRs. BMC Med Inform Decis Mak. Apr 18, 2025;25(1):169. [doi: 10.1186/s12911-025-02998-6] [Medline: 40251623]
61. Yamagishi Y, Nakamura Y, Hanaoka S, Abe O. Large language model approach for zero-shot information extraction and clustering of Japanese radiology reports: algorithm development and validation. JMIR Cancer. Jan 23, 2025;11:e57275. [doi: 10.2196/57275] [Medline: 39864093]
62. Yang X, Xiao Y, Liu D, et al. Cross language transformation of free text into structured lobectomy surgical records from a multi center study. Sci Rep. May 2, 2025;15(1):15417. [doi: 10.1038/s41598-025-97500-7] [Medline: 40316625]
63. Liu RJ, Forsythe A, Rege JM, Kaufman P. BIO25-024: real-time clinical trial data library in non-small cell lung (NSCLC), prostate (PC), and breast cancer (BC) to support informed treatment decisions: now a reality with a fine-tuned large language model (LLM). J Natl Compr Canc Netw. Mar 28, 2025;23(3.5):BIO25-024. [doi: 10.6004/jnccn.2024.7156] [Medline: 40157350]
64. Yuan Y, Zhang G, Gu Y, et al. Artificial intelligence-assisted machine learning models for predicting lung cancer survival. Asia Pac J Oncol Nurs. Dec 2025;12:100680. [doi: 10.1016/j.apjon.2025.100680] [Medline: 40201531]
65. Xiang J, Wang X, Zhang X, et al. A vision–language foundation model for precision oncology. Nature New Biol. Feb 20, 2025;638(8051):769-778. [doi: 10.1038/s41586-024-08378-w]
66. Teramoto A, Michiba A, Kiriyama Y, Tsukamoto T, Imaizumi K, Fujita H. Automated description generation of cytologic findings for lung cytological images using a pretrained vision model and dual text decoders: preliminary study. Cytopathology. May 2025;36(3):240-249. [doi: 10.1111/cyt.13474] [Medline: 39918342]
67. Lehr SA, Caliskan A, Liyanage S, Banaji MR. ChatGPT as research scientist: probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. Proc Natl Acad Sci U S A. Aug 27, 2024;121(35):e2404328121. [doi: 10.1073/pnas.2404328121] [Medline: 39163339]
68. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
69. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
70. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Nov 29, 2020. [doi: 10.48550/arXiv.1904.05342]
71. Khalpey Z, Kumar U, King N, Abraham A, Khalpey AH. Large language models take on cardiothoracic surgery: a comparative analysis of the performance of four models on American Board of Thoracic Surgery Exam questions in 2023. Cureus. Jul 2024;16(7):e65083. [doi: 10.7759/cureus.65083] [Medline: 39171020]
72. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open. Nov 1, 2023;6(11):e2343689. [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
73. Ammann PJL, Golde J, Akbik A. Question decomposition for retrieval-augmented generation. arXiv. Preprint posted online on Jul 1, 2025. [doi: 10.48550/arXiv.2507.00355]
74. Lammert J, Dreyer T, Mathes S, et al. Expert-guided large language models for clinical decision support in precision oncology. JCO Precis Oncol. Oct 2024;8:e2400478. [doi: 10.1200/PO-24-00478] [Medline: 39475661]

75. Tozuka R, Johno H, Amakawa A, et al. Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging. Jpn J Radiol. Apr 2025;43(4):706-712. [doi: 10.1007/s11604-024-01705-1] [Medline: 39585559]
76. Giuffrè M, Kresevic S, Pugliese N, You K, Shung DL. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. Liver Int. Sep 2024;44(9):2114-2124. [doi: 10.1111/liv.15974] [Medline: 38819632]
77. Arzideh K, Schäfer H, Allende-Cid H, et al. From BERT to generative AI - comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. Comput Biol Med. Sep 2025;195:110665. [doi: 10.1016/j.compbiomed.2025.110665] [Medline: 40554973]
78. Zhu M, Lin H, Jiang J, et al. Large language model trained on clinical oncology data predicts cancer progression. NPJ Digit Med. Jul 2, 2025;8(1):397. [doi: 10.1038/s41746-025-01780-2] [Medline: 40604229]
79. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: techniques and applications. Comput Biol Med. May 2023;158:106848. [doi: 10.1016/j.compbiomed.2023.106848] [Medline: 37044052]
80. Clunie D, Taylor A, Bisson T, et al. Summary of the National Cancer Institute 2023 virtual workshop on medical image de-identification—part 2: pathology whole slide image de-identification, de-facing, the role of AI in image de-identification, and the NCI MIDI datasets and pipeline. J Digit Imaging Inform med. Feb 2025;38(1):16-30. [doi: 10.1007/s10278-024-01183-x]
81. Zhou S, Li GY. Federated learning via inexact ADMM. IEEE Trans Pattern Anal Mach Intell. Aug 2023;45(8):9699-9708. [doi: 10.1109/TPAMI.2023.3243080] [Medline: 37022837]
82. Li S, Liu P, Nascimento GG, et al. Federated and distributed learning applications for electronic health records and structured medical data: a scoping review. J Am Med Inform Assoc. Nov 17, 2023;30(12):2041-2049. [doi: 10.1093/jamia/ocad170] [Medline: 37639629]
83. Shiri I, Salimi Y, Maghsudi M, et al. Differential privacy preserved federated transfer learning for multi-institutional ^{68}Ga-PET image artefact detection and disentanglement. Eur J Nucl Med Mol Imaging. Dec 2023;51(1):40-53. [doi: 10.1007/s00259-023-06418-7] [Medline: 37682303]
84. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med. Jul 8, 2024;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
85. Wang CK, Ke CR, Huang MS, et al. Using large language models for efficient cancer registry coding in the real hospital setting: a feasibility study. Pac Symp Biocomput. 2025;30:121-137. [doi: 10.1142/9789819807024_0010] [Medline: 39670366]
86. Moore CL, Socrates V, Hesami M, et al. Using natural language processing to identify emergency department patients with incidental lung nodules requiring follow-up. Acad Emerg Med. Mar 2025;32(3):274-283. [doi: 10.1111/acem.15080] [Medline: 39821298]
87. Geevarghese R, Solomon SB, Alexander ES, et al. Utility of a large language model for extraction of clinical findings from healthcare data following lung ablation: a feasibility study. J Vasc Interv Radiol. Apr 2025;36(4):704-708. [doi: 10.1016/j.jvir.2024.11.029] [Medline: 39662619]

Abbreviations:

CT: computed tomography
GI: gastrointestinal
HIPAA : Health Insurance Portability and Accountability Act
LC: lung cancer
LLM: large language model
NSCLC: non-small cell carcinoma
OS: overall survival
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction model Risk Of Bias Assessment Tool
PROSPERO: Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
RADS: Reporting and Data System
RAG: retrieval-augmented generation
REK: reliable external knowledge
ROBINS-I: Risk Of Bias In Non-randomized Studies - of Interventions

Edited by Javad Sarvestan; peer-reviewed by Chaochen Wu, Yuvanesh Vedaraju, Yuyun Yueniwati; submitted 19.03.2025; final revised version received 13.08.2025; accepted 14.08.2025; published 30.09.2025

Please cite as:
Zhong R, Chen S, Li Z, Gao T, Su Y, Zhang W, Liu D, Gao L, Hu K
Large Language Models in Lung Cancer: Systematic Review
J Med Internet Res 2025;27:e74177
URL: https://www.jmir.org/2025/1/e74177
doi: 10.2196/74177