Large Language Models in Lung Cancer: A Systematic Review

Ruikang Zhong, Siyi Chen, Zexing Li, Tangke Gao, Yisha Su, Wenzheng Zhang, Dianna Liu, Lei Gao, Kaiwen Hu

Submitted to: Journal of Medical Internet Research
on: March 19, 2025

Table of Contents

Large Language Models in Lung Cancer: A 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] Beijing University of Chinese Medicine Beijing CN
[2] Dongfang Hospital, Beijing University of Chinese Medicine Beijing CN

Corresponding Author:
Lei Gao MD

Dongfang Hospital, Beijing University of Chinese Medicine
No. 6, Fangxingyuan 1
Beijing
CN

Abstract

Background: In the era of data and intelligence, Artificial Intelligence (AI), especially Large Language Model (LLM), has been widely applied in the medical field.

Objective: The aim of this systematic review is to discuss the progress of LLM application in lung cancer, including clinical, educational and research aspects, and to explore the potential of LLM application in the full-cycle management of lung cancer.

Methods: We searched six electronic databases according to PRISMA guidelines. Information was screened and extracted independently by two authors according to the inclusion criteria. Quality assessment was performed using the QUADAS-2, PROBAST and ROBINS-I tools.

Results: The literature search yielded 706 relevant studies, resulting in the inclusion of a total of 28 studies published between 2023 and 2024.LLMs are widely used in lung cancer-related tasks, including assisted diagnosis, information extraction, question and answer science, and therapeutic decision support.ChatGPT is the most commonly used model and shows significant potential for improving diagnostic accuracy and patient communication. The importance of cue engineering and fine-tuning in optimising the performance of LLMs in specific clinical tasks is also highlighted.

Conclusions: LLMs have the potential to transform lung cancer management by improving diagnostic accuracy, patient communication and treatment planning. However, issues such as data security, ethical regulations, economic costs and clinical validation need to be addressed and LLMs need to be used rationally as effective tools rather than misused or even replaced by healthcare professionals.

(JMIR Preprints 19/03/2025:74177)
DOI: https://doi.org/10.2196/preprints.74177

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Large Language Models in Lung Cancer: A Systematic Review

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

1 Beijing University of Chinese Medicine, Beijing, China.

2 Dongfang Hospital, Beijing University of Chinese Medicine, Beijing, 100078, China.

Ruikang Zhong: zhongruikang1127@163.com

Siyi Chen: csybucm20@163.com

Zexing Li: lzxupup@163.com

Tangke Gao: gaotangke@163.com

Yisha Su: s1966001961@163.com

Wenzheng Zhang: 20240941302@bucm.edu.cn

Dianna Liu: zuyin2008@126.com

[#] Ruikang Zhong and Siyi Chen are co-first authors.

*Correspondence:Lei Gao: leilei014@163.com; Kaiwen Hu: kaiwenh@163.com

Abstracts

Objective: In the era of data and intelligence, Artificial Intelligence (AI), especially Large Language Model (LLM), has been widely applied in the medical field. The aim of this systematic review is to discuss the progress of LLM application in lung cancer, including clinical, educational and research aspects, and to explore the potential of LLM application in the full-cycle management of lung cancer.

Methods: We searched six electronic databases according to PRISMA guidelines. Information was screened and extracted independently by two authors according to the inclusion criteria. Quality assessment was performed using the QUADAS-2, PROBAST and ROBINS-I tools.

Results: The literature search yielded 706 relevant studies, resulting in the inclusion of a total of 28 studies published between 2023 and 2024.LLMs are widely used in lung cancer-related tasks, including assisted diagnosis, information extraction, question and answer science, and therapeutic decision support.ChatGPT is the most commonly used model and shows significant potential for improving diagnostic accuracy and patient communication. The importance of cue engineering and fine-tuning in optimising the performance of LLMs in specific clinical tasks is also highlighted.

Conclusions:LLMs have the potential to transform lung cancer management by improving diagnostic accuracy, patient communication and treatment planning. However, issues such as data security, ethical regulations, economic costs and clinical validation need to be addressed and LLMs need to be used rationally as effective tools rather than misused or even replaced by healthcare professionals.

Keywords: Lung Cancer; Large Language Modeling; Human-Computer Interaction; Artificial Intelligence; Full-Cycle Management

1. Introduction

Cancer is currently a major global health challenge[1, 2]. Lung cancer is one of cancers with the highest incidence and mortality worldwide[3].The concept of holistic, full-cycle oncology health management has been gradually implemented in the clinic. In the case of lung cancer, comprehensive means prevention, screening, diagnosis, treatment, care and prognostic follow-up. And full cycle refers to the division of lung cancer management into pre-cancer, cancer treatment (acute cycle) and post-cancer (chronic cycle)[4]. This new concept of lung cancer management shifts lung cancer treatment from disease-centred to patient-centred. This is in line with the concept of 5P medicine (Personalized cancer medicin: Prevemtive, Predictive, Personalized, Participatory, Precision Medicine)[5]. However, there is a gap between the new concept and reality in lung cancer management. Current research topics continue to focus on the cancer treatment cycle, such as The Lancet 2024, which published a review discussing new perspectives and challenges in the treatment of advanced non-small cell lung cancer[6]. Although some studies have focused on lung cancer screening in recent years, the acceptance of annual Low-dose Computed Tomography (LDCT) screening remains low[7, 8].In addition, the right tools have not yet been found for the personalised treatment of lung cancer[9].

With the continuous development of computer technology, AI has become an effective tool for lung cancer diagnosis. For example, algorithms such as natural language processing, machine learning and deep learning have significant applications in lung cancer diagnosis and prognosis[10]. In recent years, the emergence of LLM has reinvigorated the application of AI in healthcare[11]. LLM has characters such as large scale, high computational power, contextual learning[12], which are better able to deeply interpret the information in the data and handle heavy clinical tasks. Thus, LLM may have great potential to improve the full-cycle management in lung cancer. Another outstanding result of LLMs is conversational robot, which can provide personalised information interactions for patients and doctors in healthcare[13]. As such, LLM is also expected to be a novel decision support tool for personalised diagnosis and treatment of lung cancer patients.

This study focuses on the applied research of LLMs in lung cancer. Through systematic analysis, we summarise the overview of the applied research of LLMs in lung cancer, in terms of application scenarios, model usage, fine-tuning methods, limitations and future directions. The aim of this study is to explore how clinicians and researchers can better utilise the AI tools represented by

LLM and rationally consider the value and potential pitfalls of its application, under the requirements of the intelligent era and 5P medicine, .

2. Methods

The study followed the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) guidelines[14]. The protocol was registered on PROSPERO (CRD42024612388).

2.1 Eligibility criteria

During the study, we formulated strict inclusion and exclusion criteria according to the problems to be solved, as shown in Table 1.We did not set a timeline for the studies.

Table 1 Inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Types of studies	Peer Reviewed Journals	Books and book chapters; Letters; Reviews; Conference proceedings; Journal articles that are not published in Peer Reviewed Journals
Content	Content involves LLMs and LC	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

Notes: LLMs: Large Language Models; LC: Lung Cancer.

2.2 Data sources

We identified eligible studies by searching six electronic databases, including PubMed, Web of Science, IEEE, Embase, Cochrane Library, and Scopus. The final search was run up to January 1, 2025.

2.3 Search strategy

The search string was structured as follows: ((large language model) OR (LLM) OR (Chat GPT) 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)).

2.4 Selection process

Endnote X9.3.3 (Bid 13966) software was used to check all references and remove duplicates. Two authors (RZ, SC) screened the titles and abstracts independently, and subsequently screened the full text independently according to the inclusion and exclusion criteria. Any disagreement during the selection process was resolved by discussion, with arbitration by the third author (ZL) if necessary.

2.5 Data collection process

The data collection process was completed by the two authors (RZ, SC). All information extracted from the main text, tables, charts and appendices was annotated using the Excel of WPS office (12.1.0.18608).

2.6 Data items

The data extraction table included: title, the first author, publication year, study design, LLM devices, application scenarios, intervention, prompt engineering, input, output, outcome measures, safety measures and limitations.

2.7 Quality appraisal

To properly assess the quality of the included studies, we decide to use mixed methods, following Mahmud Omar's approach[15]. We chose the appropriate quality assessment tools based on the application of LLMs in different studies. For studies in which LLMs were primarily applied to lung cancer diagnosis or lung cancer staging, we used QUADAS-2(Quality Assessment of Diagnostic Accuracy Studies-2)[16]. For studies in which LLMs were primarily applied to predict lung cancer-related information, we used PROBAST(Prediction model Risk Of Bias ASsessment Tool)[16]. For studies in which LLMs were mainly applied to knowledge quizzing and information extraction, we use ROBINS-I(Risk Of Bias In Non-randomised Studies - of Interventions)[17]. As the form and content of studies related to LLMs were different from traditional clinical trials, we modified the entries of each scale appropriately according to the actual situation of the study and the research purpose of this review. The quality assessment was carried out back-to-back by two researchers (SC, ZL) and, in the case of controversial content, by a third researcher (RZ) in order to deliberate jointly on the decision. See supplimentrary materials for details.

2.8 Synthesis methods

Meta-analysis was not planned in this review. We conducted a mixed approach including quantitative and qualitative analysis for final data synthesis. We mainly used WPS and the website (BioRender.com). for graphing.

3. Results

3.1 Search results

In this study, a total of 706 references were retrieved, and 28 studies were finally included after screening. The specific screening process is shown in Figure 1.

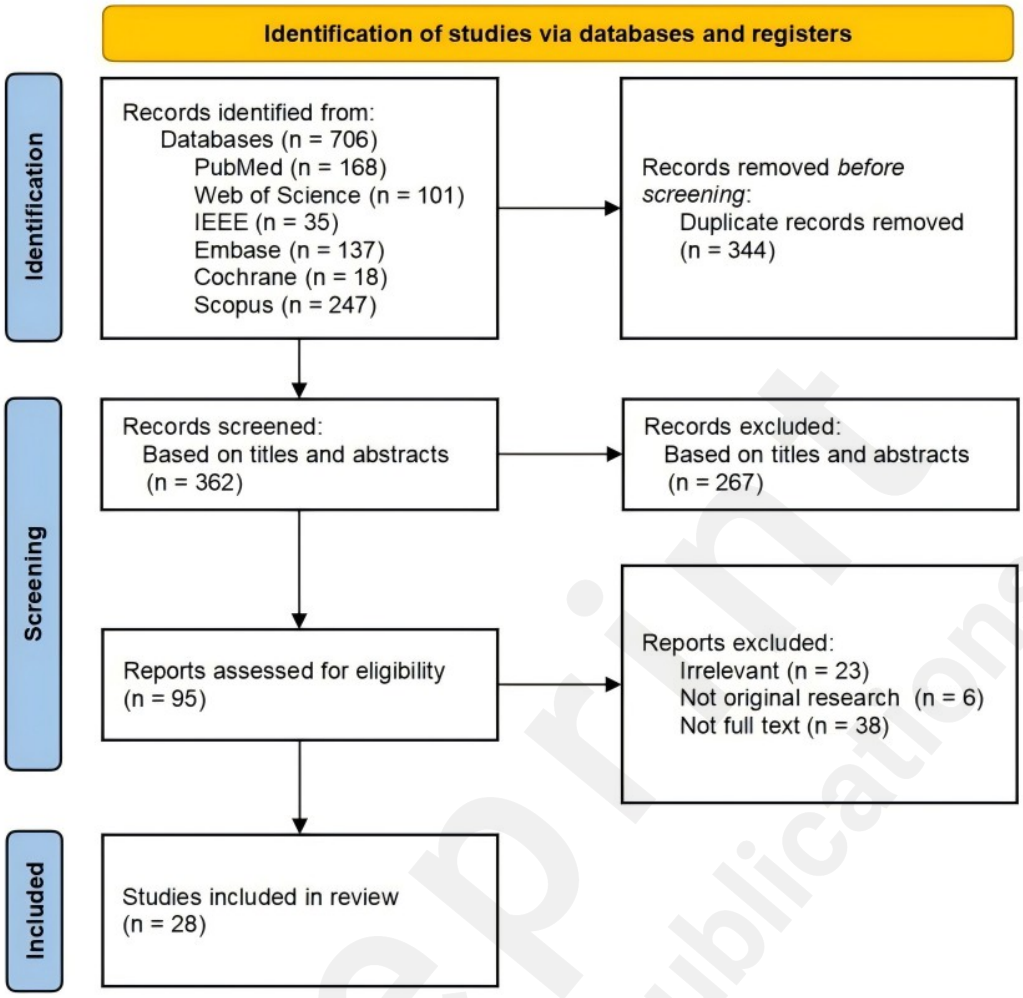


Figure 1 Flowchart (produced according to the PRISMA 2020 flow diagram)

3.2 Basic information of included sources

All the included sources were published in 2023 and 2024, of which 7 were published in 2023 and 21 were published in 2024. There were 13 studies from the USA (the United States of America), 3 from Korea, 3 from Germany, and 3 from China. Other studies were from India, Turkey, Japan, Greece and the Netherlands. There were 5 conference papers and 4 preprinted journals. See Table 2 for details.

Table 2 Summary of included sources

Study	Country	Title	Device	Key Takeaway
Kriti 2024[18]	USA	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	GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, spaCy	Extracted phenotype information of NSCLC
Hyeongmin 2024[19]	Korea	Extracting lung cancer staging descriptors from pathology reports: A generative language model approach	Llama-2-7B, Mistral-7B, Deductive Llama-2-7B (Orca-2), Deductive Mistral-7B (Dolphin), AWS Llama-2-70B, AWS Titan express	Determine clinical stage LC
Reza 2024[20]	Germany	Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment	ChatGPT-4V	Read CT slices for LC diagnosis
Arti 2024[21]	India	Automating Clinical Trial Eligibility Screening: Quantitative Analysis of GPT Models versus Human Expertise	GPT-3.5-turbo	Classify patients for clinical trial eligibility by pathology reports
Dong 2024[22]	USA	Large-language-model empowered 3D dose prediction for intensity-modulated radiotherapy	LLAMA-2	Predict the radiation dose
Fabiana 2024[23]	USA	An example of leveraging AI for documentation: ChatGPT-generated nursing care plan for an older adult with lung cancer	ChatGPT	Generate effective care plan recommendations for elderly LC patients
Ferrari 2024[24]	USA	Evaluating ChatGPT as a Patient Resource for Frequently Asked Questions about Lung Cancer Surgery–A Pilot Study	ChatGPT3.5	Answers to questions about LC and its surgery
Matthias 2023[25]	Germany	Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer	ChatGPT, ChatGPT 4	Extract data according to the original lung CT report
Adem 2024[26]	Turkey	Readability analysis of ChatGPT's responses on lung cancer	ChatGPT-3.5-turbo	Answer questions related to LC
Hana 2023[27]	USA	Use of ChatGPT, GPT-4, and Bard to Improve Readability of ChatGPT's Answers to Common Questions About Lung Cancer and Lung Cancer Screening	ChatGPT, ChatGPT 4, bard	Answer questions about LC and LCS
Hu 2024a[28]	China	Zero-shot information extraction from radiological reports using ChatGPT	ChatGPT	Extract CT report information of LC
Hu 2024b[29]	China	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	GPT-4	Predict lymph node metastasis in patients with LC
Huang 2024[30]	USA	A critical assessment of using ChatGPT for extracting structured data from clinical notes	ChatGPT-3.5-Turbo-16k	Evaluate the stage and histological type of LC
James 2024[31]	USA	Physician Assessment of ChatGPT and Bing Answers to American Cancer Society's Questions to Ask About Your Cancer	ChatGPT 3.5,Bing AI	Answer the patient's general LC questions
Sangwoon 2024[32]	Korea	The Prediction of Stress in Radiation Therapy:Integrating Artificial Intelligence with Biological Signals	ChatGPT 4.0/3.5	Predict the stress response of patients undergoing radiotherapy
Koichiro 2024[33]	Japan	Fine-Tuned Large Language Model for Extracting Patients on Pretreatment for Lung Cancer from a Picture Archiving and Communication System Based on Radiological Reports	Transformers Japanese model	Extract pretreatment LC present
Wang 2024[34]	USA	Scientific figures interpreted by ChatGPT:strengths in plot recognition and limits in color perception	GPT-4V	Interpret the K-M survival analysis chart
Gopi 2024[35]	India	Improved Lung Cancer Detection through Use of Large Language Systems with Graphical Attributes	NR	Interpret medical information and diagnose LC
Julian 2023[36]	Germany	ChatGPT: Can You Prepare My Patients for [18F]FDG PET/CT and Explain My Reports?	ChatGPT	Answer questions related to common clinical indications for [18F]FDG PET/CT
Amir 2023[37]	USA	How AI Responds to Common Lung Cancer Questions: ChatGPT versus Google Bard	ChatGPT-3.5, Google Bard experimental version	Answer questions about LC terminology
Qu 2024[38]	China	The Rise of AI Language Pathologists: Exploring Two-level Prompt Learning for Few-shot Weakly-supervised Whole Slide Image Classification	GPT-4	Acquire pathological knowledge of LC
Dimitrios 2023[39]	Greece	Evaluation of ChatGPT-supported diagnosis, staging and treatment planning for the case of lung cancer	ChatGPT	Diagnosis of LC and its staging
Niu 2024[40]	USA	Cross-Institutional Structured Radiology Reporting for Lung Cancer Screening Using a Dynamic Template-Constrained Large Language Model	lama-3.1 (8B, 70B, 405B), Qwen-2 (72B), Mistral-Large (123B)	Create structured and standardized LCS reports
Narmada 2023[41]	USA	Applying Large Language Models for Causal Structure Learning in Non Small Cell Lung Cancer	NR	Extract data to generate DAG graph
Sneha 2024[42]	Netherla	Transfer learning with BERT and ClinicalBERT models for multiclass	BERT, ClinicalBERT	Classification of cancer types

	nds	classification of radiology imaging reports		
Lyu 2023[43]	USA	Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential	ChatGPT	Translate the radiology report
Kyeryoung 2024[44]	USA	SEETrials: Leveraging Large Language Models for Safety and Efficacy Extraction in Oncology Clinical Trials	GPT-4	Extract safety and efficacy information from study abstracts
Jong 2024[45]	Korea	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	ChatGPT (gpt-40, GPT-4, GPT-3.50)	LC staging

Notes: LC: Lung Cancer; LCS: Lung Cancer Screening; NSCLC: Non-small-cell lung cancer.

3.3 Prompt Engineering and Model Training

Prompt Engineering plays a crucial role in the training of LLMs. It is the focus of specifics that is often discussed in LLM study. Therefore, we summarized the training process of prompt engineering, inputs, outputs, and outcome metrics from the included studies (Table 3). It can be seen that 16 studies explicitly demonstrated their use of prompt engineering, but the remaining 12 studies did not. Most of the prompt engineering used in LLM in lung cancer applications is still relatively primitive, with prompt frames and direct instructions dominating. Methods such as Zero-shot (Kriti 2024) and reverse prompt engineering (Dimitrios 2023) can also be seen. The surprise, however, is that with the development of LLMs' image analysis capability, LLMs are gradually being applied in the lung cancer field to the analysis of statistical images, radiological images and other highly specialised images, in addition to text processing. The outcome indicators are based on confusion matrices and rating scales, mostly compared to artificially created gold-standard answers, or by two or more experts judging the results separately. Some studies will also focus on the time and economic cost of generating results for LLMs.

Table 3 Prompt Engineering and Model Training

Study	Prompt method/content	Model input	Model output	Outcome indicators
Kriti 2024[18]	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, FI score, generation time, operating costs
Hyeongmin 2024[19]	morphology group	Segmented pathological report	42 lung cancer staging descriptors; TN classification	Macro FI score, accurate matching ratio, accuracy
Reza 2024[20]	Not mentioned	CT images of lung window	Diagnosis of lung cancer (yes or no)	Accuracy, sensitivity, specificity
Arti 2024[21]	Not mentioned	Unprocessed raw dataset	Whether the patient is qualified for enrollment (yes or no)	Accuracy compared to manual classification
Dong 2024[22]	Clinical physician commands (findings,	CT images	DVH (Radiation Dose Volume Histogram)	Mean absolute error (MAE) of Dmax,

	treatment goals, and precautions)			Dmean, D95, and D1 between actual and predicted plans
Fabiana 2024[23]	Patient's needs framework (Situation/ Background, Physical, Safety, Psychosocial, Spi ritual/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)
Ferrari 2024[24]	Not mentioned	Questions	Answers	5-point Likert scale
Matthias 2023[25]	25 original lung cancer CT 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
Adem 2024[26]	Not mentioned	Questions	Answers	Flesch Reading Ease (FRE) formula, fesch-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
Hana 2023[27]	Not mentioned	Questions	Baseline responsesr and simplified responses	Reading Ease Score, readability, clinical appropriateness
Hu 2024a[28]	Prompt template, including an IE 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 F1 score
Hu 2024b[29]	Prompt templates, including roles, tasks, patient data, machine	Prompt templates	Prediction results of lymph node metastasis in lung cancer	AUC, AP (average precision of three repetitions)

	learning model results and instructions			
Huang 2024[30]	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, F1 score, Kappa, recall rate
James 2024[31]	Not mentioned	Questions	Answers	Self rating
Sangwoon 2024[32]	Not mentioned	Biological signals before radiotherapy and instructions	Prediction results of biological signals and stress response during radiotherapy	Accuracy, recall rate, precision, F1 Score
Koichiro 2024[33]	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
Wang 2024[34]	instruction	Kaplan Meier (K-M) curves generated based on gene expression data and survival information	Analysis and Interpretation of K-M curves	Overall accuracy, Accuracy under each category
Gopi 2024[35]	Not mentioned	Images, symptoms, clinical prescriptions	Diagnosis of lung cancer (yes or no)	Accuracy, recall rate, F1 score, AUC
Julian 2023[36]	Regeneration-response function repeated three times for training	Questions	Answers	Self rating
Amir 2023[37]	Not mentioned	Questions	Answers	Accuracy, consistency
Qu 2024[38]	Guide GPT-4 to visually describe complex medical concepts	Questions(text)	Answers	AUC
Dimitrios 2023[39]	Build and refine prompts based on the returned answers	Symptom description	Diagnosis and treatment plan for lung cancer	Self drafted standards
Niu 2024[40]	Not mentioned	CT imaging	Standardized and structured radiological reports	F1 score, confidence interval, McNemar test, and z-test
Narmada	Codeinterpreter	Electronic medical	directed acyclic graph	Bdeu score

2023[41]	plugin(developed by OPEN AI)	records, genomic data		
Sneha 2024[42]	Not mentioned	radiological reports	Classification results of lung cancer	AUC, F1 score, accuracy, recall rate,, precision
Lyu 2023[43]	instruction	radiological reports	Report translation and suggestions	Self score, report completeness and accuracy
Kyeryoung 2024[44]	Prompt templates	Journal abstract	Details of clinical trials in the article	Accuracy, recall rate, F1 score
Jong 2024[45]	instruction	Chest CT and FDG PET/CT reports	The maximum size of the primary tumor, local invasion, satellite lesions, metastatic lymph nodes, intrathoracic and extrathoracic metastases, and TNM staging diagnosis	Accuracy, recall rate, F1 score, average task completion time, misreading rate

3.4 Quality appraisal

We categorized the included studies according to their different research purposes and assessed the quality of these studies with corresponding assessment scales. We evaluated 3 predictive studies with PROBAST (Figure 2A, Supplementary Table 1). These three studies had no significant bias in data source, study population and research methodology, and their applicability was good. However, they may have a high risk of bias in terms of predictors and outcomes.

We evaluated 10 diagnostic studies with QUADAS-2 (Figure 2B, Supplementary Table 2), which showed that the applicability of these 10 studies was good, but the risk of bias was unclear.

We assessed 15 intervention trials with ROBINS-I (Figure 2C, Supplementary Table 3). These studies had a low risk of bias in participant selection and outcome measurement, but had an unclear or high risk of bias in other aspects.

3.5 Other aspects

In addition, we counted whether the study mentioned human intervention in the research process, data security measures, and study limitations. We found human intervention in most trials. Six studies explicitly stated that they had initiatives related to data security issues, but 22 studies made no mention of this. 20 studies discussed the limitations of their study, but 8 did not. More details in Table 4.

Table 4 Other information

	Human Intervention	Limitations	Data Security
Y	26(92.86%)	20(71.43%)	6(21.43%)
N	2(7.14%)	8(28.57%)	22(78.57%)
M			

Notes: Y: Yes; NM: Not Mentioned.

Figure.2 A The quality appraisal for 3 predictive studies with PROBAST. B The quality appraisal for 10 diagnostic studies with QUADAS-2. C The quality appraisal for 15 intervention trials with ROBINS-I.

4. Discussion

With the advancement of hard drives and semiconductors, the capabilities of big data storage and dataset-based computer modeling have become more potent, paving the way for the goal of computers that can simulate human interaction[46]. In tandem, AI technology and large language models have emerged and rapidly evolved, finding applications across various disciplines[47]. The field of lung cancer, being one of the world's challenges, is no exception. Systematic reviews are a common method of analysis, and scientific systematic reviews can provide evidence-based support for health care[48] Evidence synthesis is a pivotal instrument in the realms of evidence-based medicine and policy development, particularly in an era characterised by the proliferation of scientific publications and journals[49]. In this study, the included literature was systematically analysed and quality assessed using the systematic review method, with a view to providing a detailed and comprehensive discussion of the application of LLMs in lung cancer.

4.1 Applications of LLMs in lung cancer

The findings of this study demonstrate that LLMs possess a broad spectrum of applications in the domain of lung cancer. These applications can be classified into a total of seven distinct categories, namely: auxiliary diagnosis[19, 20, 21, 29, 30, 33, 35, 38, 39, 42, 45], information extraction[18, 25, 28, 40, 41, 43, 44], question answering[23, 24, 26, 27, 31, 34, 36, 37], scientific research[34, 44], science education[24, 26, 27, 31, 36, 37], aided nursing[23] and treatment decision[22, 32]. (Figure 5).

The largest number of studies is in the application of auxiliary diagnosis[19, 20, 21, 29, 30, 33, 35, 38, 39, 42, 45], such as assisted lung cancer diagnosis and clinical staging. Early diagnosis of lung cancer can effectively improve its survival rate[50] Most studies have shown the efficacy of LLMs in enhancing the precision of lung cancer detection. LLMs are anticipated to play a pivotal role in facilitating clinical decision-making processes and lung cancer staging. At present, the primary applications of LLMs in the field of lung cancer diagnosis are centred on the analysis of textual data and the

interpretation of image data. Trained imaging physicians working with the assistance of LLMs can significantly reduce the cost of CT interpretation and the workload of radiologists in lung cancer screening programmes[51]. However, the findings of Jong 2024[45] also suggest that the utilisation of LLMs as a surrogate for medical professionals in the staging of lung cancer is not substantiated, emphasising the necessity for further research and validation.

ChatGPT is a representative conversational LLM that performs various complex natural language processing (NLP) tasks, including translation and question-answering[52]. Consequently, a growing body of research has been dedicated to evaluating the efficacy of LLMs in knowledge quizzes related to lung cancer[23, 24, 26, 27, 31, 34, 36, 37]. Studies have generally recognised that LLMs have the potential to serve as significant sources of information for lung cancer patients in the future. However, there is still a need to enhance the accuracy and consistency of LLMs, and they require expert oversight. It is important to acknowledge that the majority of LLMs are not exclusively developed for healthcare applications. They are predominantly trained using public databases, which results in knowledge gaps concerning specialised domains of lung cancer. In order to utilise LLMs on a large scale in a lung cancer quiz, it is first necessary to establish a more specialised and granular database. In addition, the implementation of expert supervisory intervention is required to ensure the accuracy and appropriateness of the models for individual patients[53].

LLMs can extract clinically relevant information by applying NLP[54]. Consequently, a number of studies have employed LLMs for the purpose of information extraction from lung cancer-related medical records or reports[18, 25, 28, 40, 41, 43, 44]. The studies concluded that LLM has great potential for medical data mining. The task of extracting features from large amounts of unstructured text is inherently a difficult, labour-intensive and time-consuming task[55] In clinical work, healthcare workers spend a lot of time searching for feature data that can be used for disease diagnosis and treatment from cumbersome free-text information, which reduces the efficiency and the time they can spend with patients. The advent of LLMs has paved a

new path for the challenge of extracting clinical feature information[56]. However, in information extraction, LLMs also have a number of deficiencies, such as insufficient extraction performance for complex tasks, the existence of illusions, etc. Further research and improvement are needed in the future.

Furthermore, LLMs have also been employed in the context of lung cancer research[34, 44], aided nursing[23] and treatment decision[22, 32]. In summary, the present application of big language modelling in lung cancer involves scientific research, education, clinical and other fields, and the beneficiary population has expanded from doctors to patients, researchers and the general public. However, based on the concept of holistic, full-cycle oncology health management, the current research on LLMs is less involved in the application of prognostic follow-up (post-cancer) for lung cancer. Prognostic follow-up of lung cancer includes assessment of post-treatment symptoms, life quality, psychological assessment, review intervals, and social counselling[57]. It can be said that prognostic follow-up is the link that truly embodies the concept of patient-centred 5P medicine[58]. Prognostic follow-up involves complex work, personnel and data. Achieving accurate follow-up for each lung cancer patient with limited healthcare resources is challenging. Furthermore, the process contains private information such as patients' contact information, which current LLMs cannot autonomously process in accordance with ethical requirements. Consequently, following further enhancement of arithmetic capabilities, expansion of functionality, and fortification of information security processing, LLMs have the potential to serve as a potent auxiliary instrument in the management of lung cancer patients, facilitating precise follow-up procedures.

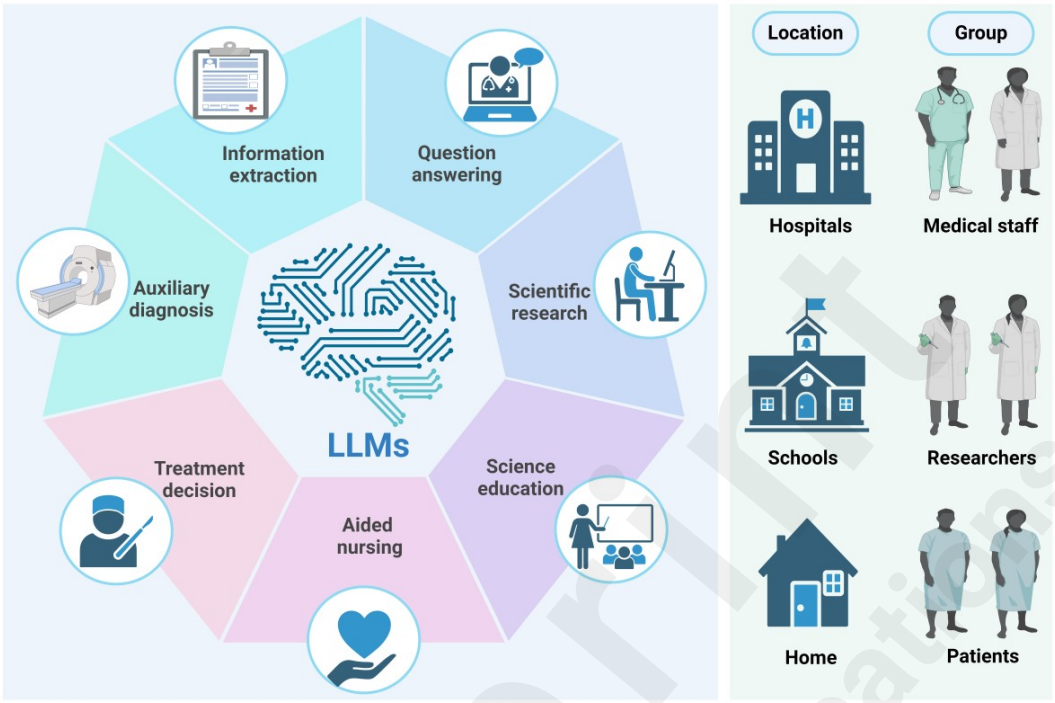


Figure 3 Applications of LLMs in lung cancer

4.2 Representative LLMs: ChatGPT

LLMs have developed rapidly since their introduction in 2019, particularly following the emergence of ChatGPT-3. This development has garnered significant attention from the community, leading to the establishment of individual LLMs by companies and subsequent iterations of these models[59]. Our systematic review found that a variety of studies have opted for disparate LLMs. Furthermore, a significant number of studies have conducted comparisons of the application of multiple LLMs. On the one hand, it is evident that the various versions and functions of LLMs all have the potential for application in lung cancer. Conversely, the diversity of LLMs complicates the selection process for researchers, and the utilisation of diverse models may influence the study's outcomes.

Of the studies that were included in this analysis, 21 employed the GPT family. As for the application of lung cancer quiz, scholars all used ChatGPT to conduct their research. ChatGPT, the LLM with the largest number of users in the world, has demonstrated its superior arithmetic power, extensive knowledge database, and

notable dialogue and reasoning capabilities. It is also the most researched LLMs at present. Several studies[60, 61, 62] have evaluated the performance of the ChatGPT and its updated version on the United States Medical Licensing Examination (USMLE). The findings indicate that ChatGPT is able to achieve or come close to a passing score, demonstrating its capacity to encompass a substantial breadth of medical knowledge. It also shows higher accuracy in performance comparisons with other LLMs[63, 64]. This may be the reason why researchers conducting LLM studies in the field of lung cancer would first choose ChatGPT. However, ChatGPT also suffers from the problem that the training data is not updated in a timely manner, and that facts can be fictionalised in knowledge domains other than the training date[65]. And as a general-purpose macromodel, it does not seem to have an advantage in clinical text categorisation[66], medical question and answer tasks[67] and clinical prediction tasks[68] compared to macromodels designed specifically for clinical applications. Consequently, the provision of timely, realistic, and targeted high-quality training data is imperative to enhance the efficacy of LLMs in the domain of lung cancer research.The implementation of generic macromodels, such as ChatGPT, necessitates meticulous fine-tuning and optimisation for specific clinical and research applications. In addition, the integration of alternative LLMs should be encouraged.

4.3 Man-on-the-loop application model

The concept of 'human-on-the-loop' can be defined as a model of human-computer hybridisation, whereby humans are involved in the decision-making or learning process of an AI system. In this model, humans interact with the system to establish a closed-loop feedback mechanism. It emphasises the role of humans in supervising, correcting or optimising computers[69]. Despite the pervasive utilisation of LLMs in daily life, they are still subject to professional supervision and intervention[70]. They are not yet able to work completely independently in the face of complex and rigorous clinical tasks as well as legal and ethical requirements[71]. The 28 studies included in this paper were evaluated by medical professionals who performed the assessment of the model outputs. The assessment method employed is

either a direct evaluation of accuracy by an expert, a comparison with a gold standard developed by an expert, or a judgement given by a clinician. This ‘human-on-the-loop’ application model emphasises the collaboration between humans and LLMs, rather than substituting for one or the other. Collaboration is manifested both in LLMs assisting humans to improve their original productivity, and in humans improving the accuracy of LLM through cue engineering, evaluating feedback.

Prompt engineering is the process of designing and optimising input prompts to guide LLMs in generating task-specific high-quality outputs. It is the most common method of optimising the performance of a general-purpose LLMs for a specific domain or a specific task[72], and it also represents the fundamental aspect of reflecting the 'human-on-the-loop' model. 12 of the included studies did not explicitly state the prompting works. The prompting works were mostly simple quizzes used for science education, with only concise and direct questions or instructions for LLM, such as 'What are the clinical staging criteria for lung cancer?' 16 studies explicitly described the method or content of prompting for LLM. Among them, the more used are prompt templates and fine tuning (fine tuning). The prompt template will add roles, case examples, task requirements, some lung cancer speciality knowledge and formatting requirements[73]. For example, ‘You are an oncologist with more than 10 years of experience, please summarise the clinical staging and tumour characteristics of this lung cancer patient based on the following pathology and CT imaging report, the tumour characteristics include tumour size, burr sign, and lymph node accumulation, with the attached clinical staging criteria for lung cancer and typical examples.’ Fine-tuning refers to further training the model based on pre-trained LLMs (e.g. ChatGPT, BERT) using labelled data for a specific task or domain to adapt it to the new task[74]. These two methods are more suitable for solving specific, complex tasks such as information extraction and diagnostic prediction. In 2022, OpenAI demonstrated a Reinforcement Learning by Human Feedback (RLHF) approach[75] to align language models with user intentions for a variety of tasks and improve model output accuracy. This approach has been tried in the field of GI cancers and has proved to be one of the

more promising directions for LLM optimisation[76].

4.4 Limitations of LLMs and future directions in lung cancer

AI technology, led by generative LLMs, has revolutionised every industry in the last three years. From ChatGPT to DeepSeek, LLM has attracted a great deal of attention and interest from the community. It is undeniable that LLM has significantly enhanced the efficiency of clinical and scientific work. Nevertheless, researchers must undertake a dispassionate evaluation of its limitations and potential areas for enhancement.

The primary concern is data security. To illustrate this point, consider the diagnosis and treatment of lung cancer patients, which involves the management of a substantial amount of sensitive data, including medical history, imaging data, genomic data, and more[77, 78]. This data is susceptible to leakage during the training and fine-tuning of LLMs. Large datasets tend to be polycentric[79], and as such require storage in the cloud or transfer across organisations, which increases the risk of data being hacked or used maliciously. It is evident that disparate countries and regions have established divergent security requirements for healthcare data (for instance, the GDPR[80] in the European Union and the HIPAA[81] in the United States). The globalised application of LLMs may encounter intricate compliance issues. Google DeepMind and the National Health Service (NHS) had collaborated on an AI tool for acute kidney injury prediction[82]. However, the collaboration has sparked controversy over the use of 1.6 million patients' medical data without full authorisation. In addressing the primary concern of data security, researchers have proposed a range of solutions, including data anonymisation and de-identification[83, 84], federated learning[85, 86], differential privacy[87], data encryption and access restrictions[88], etc. These researchers will continue to advocate for technological enhancements and regulatory improvements to enhance the protection of patient data. It is noteworthy that merely six of the studies included in this paper explicitly proposed specific measures to address data security, thereby suggesting that researchers may have inadvertently overlooked or deliberately avoided addressing this issue when

employing LLMs.

The second issue that requires attention is that of ethical regulations. LLMs are prone to illusions[89] and biases[90] that result in the propagation of misinformation, with potentially alarming ramifications in medical contexts. Patients are not fully convinced by the diagnostic and therapeutic advice given by LLMs. It is also unclear who is responsible if LLM gives the wrong advice that leads to bad results for patients[91]. That ethical issue is the biggest reason why AI can't get independent prescribing rights. It is also important to consider whether patients are given enough information and give their permission for their data to be used to train LLMs[92]. Consequently, the establishment of legislation and regulations on a global scale should be encouraged to elucidate the allocation of responsibilities in the implementation of AI in healthcare. Setting the ethical framework with different healthcare scenarios application needs and patient acceptability thresholds to target the use of LLM[93]. At the same time, human supervision of models can be enhanced by optimising model decision-making and fine-tuning modules, and by increasing model transparency and physician participation.

Finally, there is the issue of accessibility of LLM clinical applications. On the one hand, independent development of LLMs requires a large amount of data resources and cost budgets. Most hospitals choose to privately deploy commercially available open-source LLMs and integrate them into their existing healthcare systems. This requires additional hard and software support and a professional technical team with continuous updating and maintenance[94], which is still difficult for primary hospitals with limited conditions. Conversely, the accuracy of AI needs to be clinically validated in a multi-centre, prospective manner[95]. However, the majority of current applications of LLM in lung cancer remain in the experimental stage, lacking large-scale clinical validation and long-term follow-up data. This, in turn, engenders patient scepticism regarding AI systems, thereby impacting their application and promotion in clinical practice. To address these two aspects, we can develop lightweight vertical models for the lung cancer domain only and actively seek

government funding to reduce training and deployment costs. Multi-centre and large sample clinical trials should be conducted to provide high-quality training data, thereby enhancing the accuracy of the model and ensuring the safety and effectiveness of the model. At the same time, LLM should be selected to prioritise the application of the aforementioned links, which LLM is proficient in and which pose minimal risk to patients' health. These include the registration code[96], screening[97], and prognostic follow-up[98] of lung cancer patients, which can enhance the information cycle and systematic management under the concept of 5P medicine of lung cancer.

5. Limitations of this systematic review

This systematic review covers studies up to January 1, 2025. Due to the rapid evolution of LLM research and model updates, some recent findings may not be included. To mitigate this, we expedited publication and incorporated additional recent studies for discussion. Although six databases were searched, relevant studies outside these sources may have been missed. Additionally, restricting the review to English-language articles may limit generalizability, though only two non-English studies were excluded during screening. Unlike traditional systematic reviews of clinical interventions, quality assessment here varied by application scenario, with some metrics relying on subjective judgment. To reduce bias, two researchers independently conducted assessments, with discrepancies resolved by a third reviewer to ensure credibility.

6. Conclusions

In summary, our systematic review provides an overview of the applications and research of LLMs in lung cancer, alongside a quality assessment. LLMs can assist physicians in interpreting test reports, providing diagnostic and treatment recommendations, and supporting education, research, and public outreach. The application of LLMs holds promise for optimizing the efficiency of comprehensive lung cancer management and improving the healthcare service environment. However, researchers should also distance themselves from the hype and bandwagon effect surrounding AI, acknowledging the existing safety concerns and limitations of

LLMs. While strengthening regulations and human oversight, efforts should be made to address technical shortcomings and develop specialized large models for lung cancer and healthcare.

Abbreviations: LC: Lung Cancer; AI: Artificial intelligence; LLM: Large language model; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2; PROBAST: Prediction model Risk Of Bias ASsessment Tool; ROBINS-I: Risk Of Bias In Non-randomised Studies - of Interventions; 5P Medicine: Prevemtive, Predictive, Personalized, Participatory, Precision Medicine; LDCT: Low-dose Computed Tomography; PROSPERO: Prospective Register of Systematic Reviews; the USA: the United States of America; NLP: Natural Language Processing; USMLE: United States Medical Licensing Examination; RLHF: Reinforcement Learning fromHuman Feedback; GDPR: General Data Protection Regulation; HIPAA: Health Insurance Portability and Accountability Act; NHS: National Health Service.

Supplementary Information: The specific items and evaluation process of quality appraisal are detailed in Supplementary materials.

Declarations

Acknowledgements: None.

Author contributions: KH and LG conceived the study. 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. All authors read and approved the final manuscript.

Funding: This research was funded by National High Level Chinese Medicine Hospital Clinical Research Funding, The Science and Technology Plan Project of Beijing (Grant No. Z221100003522029), The Science and Technology Plan Project of Beijing (Grant No. Z241100007724010), Education Science Research Project, Beijing University of Chinese Medicine (XBB23032).

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

Ethics approval and consent to participate: Not applicable.

Consent for publication: All authors have read and approved the content and agree to submit for consideration for publication in the journal.

Competing interests: The authors declare that they have no competing interest.

References

[1] Azar, F.E., Azami-Aghdash, S., Pournaghi-Azar, F.et al. Cost-effectiveness of lung cancer screening and treatment methods: a systematic review of systematic reviews. BMC Health Serv Res 17, 413 (2017). doi: org/10.1186/s12913-017-2374-1.

[2] Siegel RL, Kratzer TB, Giaquinto AN, et al. Cancer statistics, 2025. CA Cancer J Clin. 2025 Jan-Feb;75(1):10-45. doi: 10.3322/caac.21871. Epub 2025 Jan 16.

[3] 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 May-Jun;74(3):229-263. doi: 10.3322/caac.21834.

[4] Miller KD, Pandey M, Jain R, et al. Cancer Survivorship and Models of Survivorship Care: A Review. Am J Clin Oncol. 2015 Dec;38(6):627-33. doi: 10.1097/COC.0000000000000153.

[5] Gorini, A., Pravettoni, G. P5 medicine: a plus for a personalized approach to oncology. Nat Rev Clin Oncol 8, 444 (2011). doi: org/10.1038/nrclinonc.2010.227-c1.

[6] Meyer ML, Fitzgerald BG, Paz-Ares L, et al. New promises and challenges in the treatment of advanced non-small-cell lung cancer. Lancet. 2024 Aug 24;404(10454):803-822. doi: 10.1016/S0140-6736(24)01029-8.

[7] Barta JA, Farjah F, Thomson CC, et al. The American Cancer Society National Lung Cancer Roundtable strategic plan: Optimizing strategies for lung nodule evaluation and management. Cancer. 2024 Dec 15;130(24):4177-4187. doi: 10.1002/cncr.35181.

[8] Jung J, Razzak E, Avenido AR. et al. Patient-Reported Barriers and Preferred Interventions to Improve Lung Cancer Screening Uptake. J Am Coll Radiol. 2025 Mar;22(3):269-279. doi: 10.1016/j.jacr.2024.10.010.

[9] Cangut B, Akinlusi R, Mohseny A, et al. Evolving Paradigms in Lung Cancer: Latest Trends

in Diagnosis, Management, and Radiopharmaceuticals. Semin Nucl Med. 2025 Mar 6:S0001-2998(25)00013-3. doi: 10.1053/j.semnuclmed.2025.02.005.

[10] Shigao Huang, Jie Yang, Na Shen, et al. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective,Seminars in Cancer Biology,Volume 89, 2023,Pages 30-37,ISSN 1044-579X. doi: org/10.1016/j.semcancer.2023.01.006.

[11] Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024 Nov;58(11):1276-1285. doi: 10.1111/medu.15402.

[12] Zhao W X , Zhou K , Li J ,et al.A Survey of Large Language Models. arXiv:2303.18223. doi: org/10.48550/arXiv.2303.18223Focus.

[13] Gorini, A., Pravettoni, G. P5 medicine: a plus for a personalized approach to oncology. Nat Rev Clin Oncol 8, 444 (2011). doi: org/10.1038/nrclinonc.2010.227-c1.

[14] Page MJ, McKenzie JE, Bossuyt PM,, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

[15] Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: A systematic review. J Affect Disord. 2025 Feb 15;371:234-244. doi: 10.1016/j.jad.2024.11.052.

[16] Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36. doi: 10.7326/0003-4819-155-8-201110180-00009.

[17] Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355:i4919. doi: 10.1136/bmj.i4919.

[18] 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. 2024 Jul 3;7(3):ooae060. doi: 10.1093/jamiaopen/ooae060.

[19] Cho H, Yoo S, Kim B, et al. Extracting lung cancer staging descriptors from pathology reports: A generative language model approach. J Biomed Inform. 2024 Sep;157:104720. doi: 10.1016/j.jbi.2024.104720.

[20] Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. Jpn J Radiol. 2024 Oct;42(10):1168-1177. doi: 10.1007/s11604-024-01606-3.

[21] Devi, A., Uttrani, S., Singla, A, et al. Automating Clinical Trial Eligibility Screening: Quantitative Analysis of GPT Models versus Human Expertise. Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments. 2024 June 26. doi 10.1145/3652037.3663922.

[22] Dong Z, Chen Y, Gay H, et al. Large-language-model empowered 3D dose prediction for intensity-modulated radiotherapy. Med Phys. 2025 Jan;52(1):619-632. doi: 10.1002/mp.17416.

[23] 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. 2024 Sep 1;31(9):2089-2096. doi: 10.1093/jamia/ocae116.

[24] 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. 2024 Sep 24:S0022-5223(24)00837-7. doi: 10.1016/j.jtcvs.2024.09.030.

[25] 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. 2023 Sep;308(3):e231362. doi: 10.1148/radiol.231362.

[26] Gencer A. Readability analysis of ChatGPT's responses on lung cancer. Sci Rep. 2024 Jul 26;14(1):17234. doi: 10.1038/s41598-024-67293-2.

[27] Haver HL, Lin CT, Sirajuddin A, et al. 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. 2023 Nov;221(5):701-704. doi: 10.2214/AJR.23.29622.

[28] Hu D, Liu B, Zhu X, et al. Zero-shot information extraction from radiological reports using ChatGPT. Int J Med Inform. 2024 Mar;183:105321. doi: 10.1016/j.ijmedinf.2023.105321.

[29] Hu D, Liu B, Zhu X, et al. 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. arXiv:2407.17900.

[30] Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting

structured data from clinical notes. NPJ Digit Med. 2024 May 1;7(1):106. doi: 10.1038/s41746-024-01079-8.

[31] Janopaul-Naylor JR, Koo A, et al. Physician Assessment of ChatGPT and Bing Answers to American Cancer Society's Questions to Ask About Your Cancer. Am J Clin Oncol. 2024 Jan 1;47(1):17-21. doi: 10.1097/COC.0000000000001050.

[32] Jeong S, Pyo H, Park W, et al. The Prediction of Stress in Radiation Therapy: Integrating Artificial Intelligence with Biological Signals. Cancers (Basel). 2024 May 22;16(11):1964. doi: 10.3390/cancers16111964.

[33] Yasaka K, Kanzawa J, Kanemaru N, et al. 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. 2025 Feb;38(1):327-334. doi: 10.1007/s10278-024-01186-8.

[34] Wang J, Ye Q, Liu L, et al. Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. NPJ Precis Oncol. 2024 Apr 5;8(1):84. doi: 10.1038/s41698-024-00576-z.

[35] G. V. Vallabhaneni, Y. S. Rahul, K. S. Kumari. Improved Lung Cancer Detection through Use of Large Language Systems with Graphical Attributes.2024 International Conference on Computing, Power, and Communication Technologies (IC2PCT). 2024 May. DOI: 10.1109/IC2PCT60090.2024.10486290.

[36] Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: Can You Prepare My Patients for [18F]FDG PET/CT and Explain My Reports? J Nucl Med. 2023 Dec 1;64(12):1876-1879. doi: 10.2967/jnumed.123.266114.

[37] Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology. 2023 Jun;307(5):e230922. doi: 10.1148/radiol.230922.

[38] Linhao Qu, Xiaoyuan Luo, Kexue Fu,et al.The Rise of AI Language Pathologists: Exploring Two-level Prompt Learning for Few-shot Weakly-supervised Whole Slide Image Classification. aeXiv preprint. arXiv:2305.17891v2.

[39] D. P. Panagoulias, F. A. Palamidas, M. Virvou, et al. Evaluation of ChatGPT-supported

diagnosis, staging and treatment planning for the case of lung cancer. 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). 2024 Nov. DOI: 10.1109/AICCSA59173.2023.10479348.

[40] Niu, C., Kaviani, P., Lyu, Q., et al. (2024). Development and Validation of a Dynamic-Template-Constrained Large Language Model for Generating Fully-Structured Radiology Reports.

[41] Naik, N., Khandelwal, A., Joshi, M., et al. Applying Large Language Models for Causal Structure Learning in Non Small Cell Lung Cancer. 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), 688-693.

[42] Sneha Mithun, Umesh B. Sherkhane, Ashish Kumar Jha, et al. Transfer learning with BERT and ClinicalBERT models for multiclass classification of radiology imaging reports, 22 July 2024, PREPRINT. doi: 10.21203/rs.3.rs-4443132/v1.

[43] 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. 2023 May 18;6(1):9. doi: 10.1186/s42492-023-00136-5.

[44] Lee K, Paek H, Huang LC, et al. SEETrials: Leveraging Large Language Models for Safety and Efficacy Extraction in Oncology Clinical Trials. medRxiv [Preprint]. 2024 May 13:2024.01.18.24301502. doi: 10.1101/2024.01.18.24301502.

[45] Lee JE, Park KS, Kim YH, et al. 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. 2024 Dec;223(6):e2431696. doi: 10.2214/AJR.24.31696.

[46] Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med. 2023 Mar 30;388(13):1201-1208. doi: 10.1056/NEJMra2302038.

[47] Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof. 2024;21:6. doi: 10.3352/jeehp.2024.21.6.

[48] Siddaway AP, Wood AM, Hedges LV. How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. Annu Rev Psychol. 2019 Jan 4;70:747-770. doi: 10.1146/annurev-psych-010418-102803. Epub

2018 Aug 8.

[49] Muka T, Glisic M, Milic J, et al. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. Eur J Epidemiol. 2020 Jan;35(1):49-60. doi: 10.1007/s10654-019-00576-5.

[50] Fathi JT, Barry AM, Greenberg GM, et al. ACS NLCRT Implementation Strategies Task Group. The American Cancer Society National Lung Cancer Roundtable strategic plan: Implementation of high-quality lung cancer screening. Cancer. 2024 Dec 1;130(23):3961-3972. doi: 10.1002/cncr.34621.

[51] Schreuder A, Scholten ET, van Ginneken B, et al. Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice? Transl Lung Cancer Res. 2021 May;10(5):2378-2388. doi: 10.21037/tlcr-2020-lcs-06.

[52] Wang L, Wan Z, Ni C, et al. A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. medRxiv [Preprint]. 2024 Apr 27:2024.04.26.24306390. doi: 10.1101/2024.04.26.24306390. Update in: J Med Internet Res. 2024 Nov 7;26:e22769. doi: 10.2196/22769.

[53] Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res 2023;25:e48568 URL. doi 10.2196/48568.

[54] Khurana D, Koli A, Khatter K, et al. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 2023;82(3):3713-3744. doi: 10.1007/s11042-022-13428-4.

[55] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform. Jan 2018;77:34-49.

[56] Choi HS, Song JY, Shin KH, et al. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J. 2023 Sep;41(3):209-216. doi: 10.3857/roj.2023.00633.

[57] Schütte W, Gütz S, Nehls W, et al. Prevention, Diagnosis, Therapy, and Follow-up of Lung Cancer - Interdisciplinary Guideline of the German Respiratory Society and the German Cancer Society - Abridged Version. Pneumologie. 2023 Oct;77(10):671-813. German. doi: 10.1055/a-2029-0134.

[58] Calman L, Beaver K, Hind D, Lorigan P, Roberts C, Lloyd-Jones M. Survival benefits from follow-up of patients with lung cancer: a systematic review and meta-analysis. J Thorac Oncol. 2011 Dec;6(12):1993-2004. doi: 10.1097/JTO.0b013e31822b01a1.

[59] Gill, Amitoj MD*; Gosain, Rohit MD†; Bhandari, Shruti MD*; Gosain, Rahul MD*; Gill, Gurkirat MBBS‡; Abraham, Joseph ScD§; Miller, Kenneth MD‖. “Lost to Follow-up” Among Adult Cancer Survivors. American Journal of Clinical Oncology 41(10):p 1024-1027, October 2018. DOI: 10.1097/COC.0000000000000408.

[60] Gilson A, Safranek CW, Huang T, et al.How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9:e45312.

[61] Mihalache A, Huang RS, Popovic MM, et al. ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2024 Mar;46(3):366-372. doi: 10.1080/0142159X.2023.2249588.

[62] Bicknell BT, Butler D, Whalen S, et al. ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis. JMIR Med Educ. 2024 Nov 6;10:e63430. doi: 10.2196/63430.

[63] Chen Y, Huang X, Yang F, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ. 2024 Nov 26;24(1):1372. doi: 10.1186/s12909-024-06309-x.

[64] Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis. Clin Anat. 2024 Nov 21. doi: 10.1002/ca.24244.

[65] Lehr SA, Caliskan A, Liyanage S, et al. 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. 2024 Aug 27;121(35):e2404328121. doi: 10.1073/pnas.2404328121.

[66] Lee J, Yoon W, Kim S,et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-1240. doi: 10.1093/bioinformatics/btz682.

[67] Xi Yang, Aokun Chen, Nima PourNejatian,et al. GatorTron: A Large Clinical Language

Model to Unlock Patient Information from Unstructured Electronic Health Records. arXiv preprint arXiv:2203.03540.

[68] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172-180. doi: 10.1038/s41586-023-06291-2.

[69] Slade P, Atkeson C, Donelan JM, et al. On human-in-the-loop optimization of human-robot interaction. Nature. 2024 Sep;633(8031):779-788. doi: 10.1038/s41586-024-07697-2.

[70] Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med. 2024 Jul 8;7(1):183. doi: 10.1038/s41746-024-01157-x.

[71] Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. Digit Health. 2023 Jul 2;9:20552076231186520. doi: 10.1177/20552076231186520.

[72] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023 Oct 4;25:e50638. doi: 10.2196/50638.

[73] Giuffrè M, Kresevic S, Pugliese N, et al. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. Liver Int. 2024 Sep;44(9):2114-2124. doi: 10.1111/liv.15974.

[74] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. arXiv preprint. arXiv:2203.02155.

[75] Rahman MM, Irbaz MS, North K, et al. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. J Biomed Inform. 2024 Oct;158:104727. doi: 10.1016/j.jbi.2024.104727.

[76] Yan Y, Sun D, Hu J, et al. Multi-omic profiling highlights factors associated with resistance to immuno-chemotherapy in non-small-cell lung cancer. Nat Genet. 2025 Jan;57(1):126-139. doi: 10.1038/s41588-024-01998-y.

[77] Mei T, Wang T, Zhou Q. Multi-omics and artificial intelligence predict clinical outcomes of immunotherapy in non-small cell lung cancer patients. Clin Exp Med. 2024 Mar 30;24(1):60. doi: 10.1007/s10238-024-01324-0.

[78] Marcinkiewicz AM, Buchwald M, Shanbhag A, et al. AI for Multistructure Incidental

Findings and Mortality Prediction at Chest CT in Lung Cancer Screening. Radiology. 2024 Sep;312(3):e240541. doi: 10.1148/radiol.240541.

[79] Voigt P, von dem Bussche A. Organisational requirements. In: Voigt P, von dem Bussche A, editors. The EU General Data Protection Regulation (GDPR): a practical guide. Cham: Springer International Publishing; 2017. pp. 31–86.

[80] Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med. 2006;57:575-90. doi: 10.1146/annurev.med.57.121304.131257.

[81] Powles, J., Hodson, H. Google DeepMind and healthcare in an age of algorithms.Health Technol.7, 351–367 (2017). doi: org/10.1007/s12553-017-0179-1.

[82] Jeon S, Seo J, Kim S,et al. Proposal and Assessment of a De-Identification Strategy to Enhance Anonymity of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) in a Public Cloud-Computing Environment: Anonymization of Medical Data Using Privacy Models. J Med Internet Res. 2020 Nov 26;22(11):e19597. doi: 10.2196/19597.

[83] 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 Imaging Inform Med. 2025 Feb;38(1):16-30. doi: 10.1007/s10278-024-01183-x.

[84] Zhou S, Li GY. Federated Learning Via Inexact ADMM. IEEE Trans Pattern Anal Mach Intell. 2023 Aug;45(8):9699-9708. doi: 10.1109/TPAMI.2023.3243080.

[85] 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. 2023 Nov 17;30(12):2041-2049. doi: 10.1093/jamia/ocad170.

[86] Shiri I, Salimi Y, Maghsudi M, et al. Differential privacy preserved federated transfer learning for multi-institutional 68Ga-PET image artefact detection and disentanglement. Eur J Nucl Med Mol Imaging. 2023 Dec;51(1):40-53. doi: 10.1007/s00259-023-06418-7.

[87] Vanin FNDS, Policarpo LM, Righi RDR, et al. A Blockchain-Based End-to-End Data Protection Model for Personal Health Records Sharing: A Fully Homomorphic Encryption Approach. Sensors (Basel). 2022 Dec 20;23(1):14. doi: 10.3390/s23010014.

[88] Huang L, Yu W, Ma W. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025;43:42. doi: 10.1145/3703155.

[89] Agrawal A. Fairness in AI-Driven Oncology: Investigating Racial and Gender Biases in Large Language Models. Cureus. 2024 Sep 16;16(9):e69541. doi: 10.7759/cureus.69541.

[90] Huang L, Yu W, Ma W. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025;43:42. doi: 10.1145/3703155.

[91] Guleria A, Krishan K, Sharma V, et al. ChatGPT: ethical concerns and challenges in academics and research. J Infect Dev Ctries. 2023 Sep 30;17(9):1292-1299. doi: 10.3855/jidc.18738.

[92] Ong JCL, Chang SY, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024 Jun;6(6):e428-e432. doi: 10.1016/S2589-7500(24)00061-X.

[93] Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med. 2024 Jul 8;7(1):183. doi: 10.1038/s41746-024-01157-x.

[94] Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med. 2025 Mar 6;8(1):143. doi: 10.1038/s41746-025-01476-7.

[95] Lococo F, Boldrini L, Diepriye CD, et al. Lung cancer multi-omics digital human avatars for integrating precision medicine into clinical practice: the LANTERN study. BMC Cancer. 2023 Jun 13;23(1):540. doi: 10.1186/s12885-023-10997-x.

[96] 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.

[97] Mao Y, Xu N, Wu Y, et al. Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images. Cell Rep Med. 2025 Feb 25:101988. doi: 10.1016/j.xcrm.2025.101988.

[98] Yuan Q, Cai T, Hong C, et al. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. JAMA Netw Open. 2021 Jul 1;4(7):e2114723. doi: 10.1001/jamanetworkopen..2021.14723.

Supplementary Files

Inclusion and exclusion criteria.
URL: http://asset.jmir.pub/assets/c089166c006f194e2694ee727f011b29.doc

Summary of included sources.
URL: http://asset.jmir.pub/assets/914b589e28e61bfda3df67b6d35d682e.doc

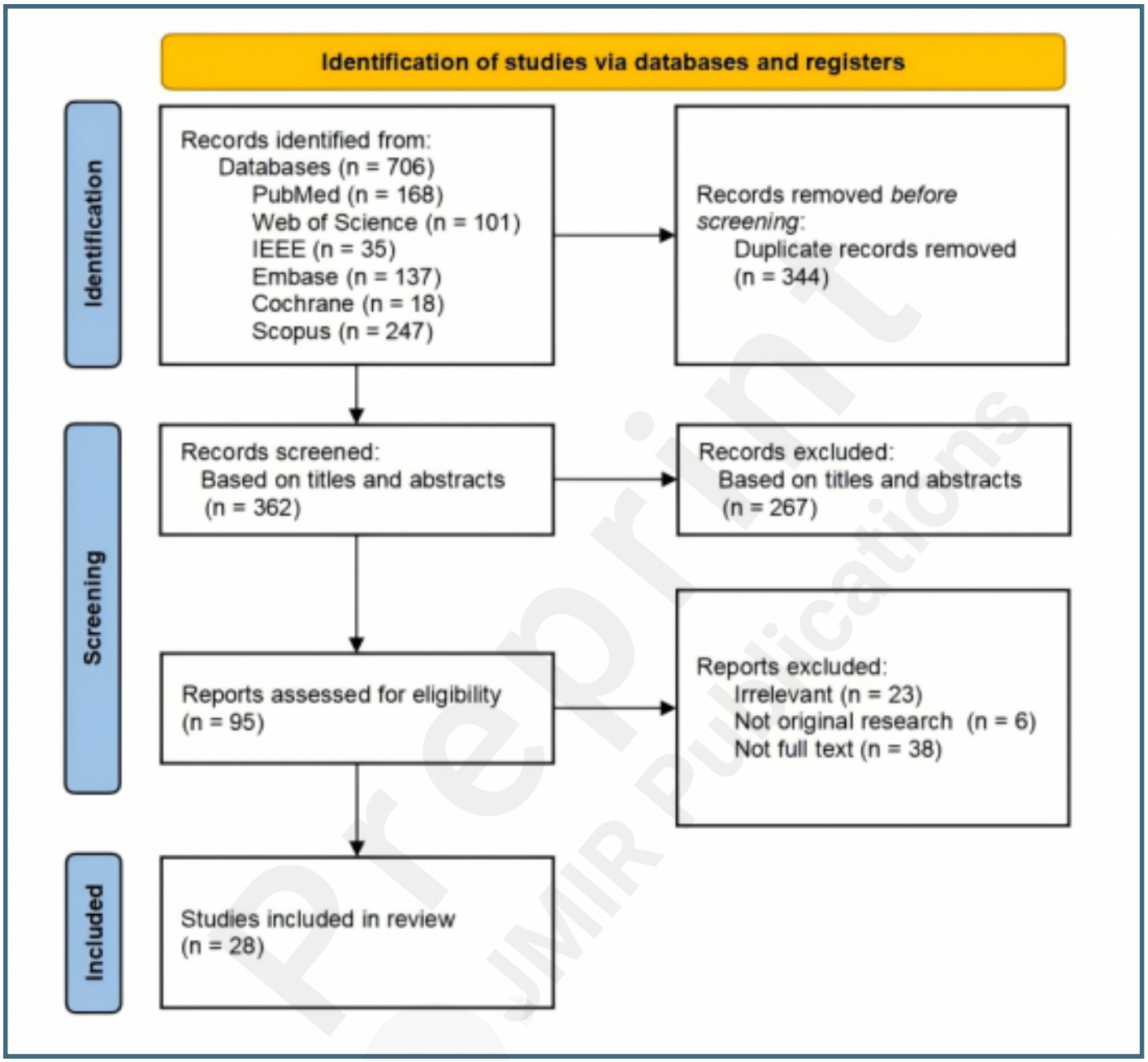
Prompt engineering and model training.
URL: http://asset.jmir.pub/assets/74c5ad4794c90593ac828338693d0703.doc

Other information.
URL: http://asset.jmir.pub/assets/0370b5db8d99ad5cda64b204c601dc39.doc

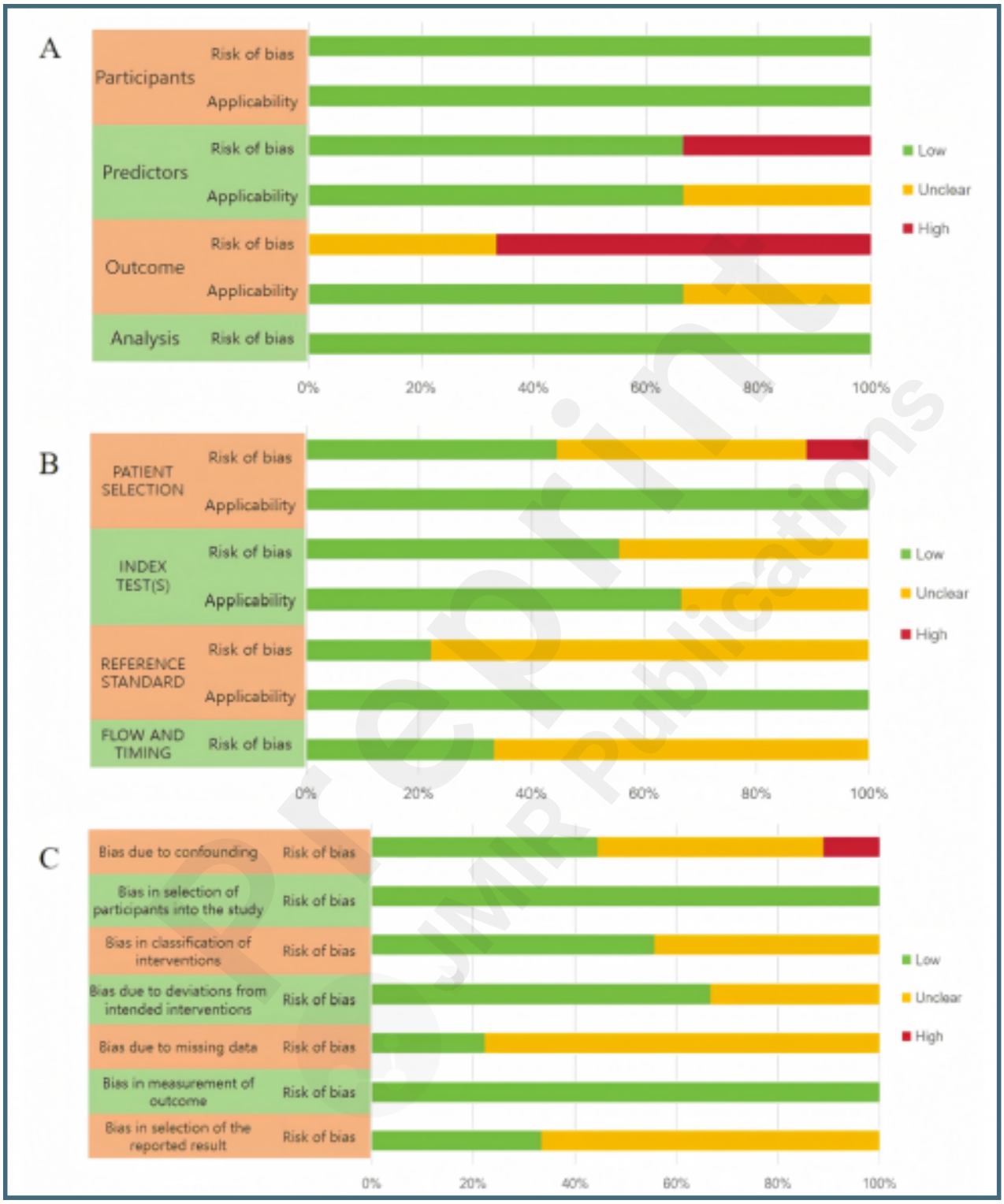
Supplimentrary materials.
URL: http://asset.jmir.pub/assets/e631df2782b26bdfe1046811f40c193f.doc

Figures

Flowchart.


Identification of studies via databases and registers
Identification
Records identified from:
Databases (n = 706)
PubMed (n = 168)
Web of Science (n = 101)
IEEE (n = 35)
Embase (n = 137)
Cochrane (n = 18)
Scopus (n = 247)
Records removed *before screening*:
Duplicate records removed (n = 344)
Screening
Records screened:
Based on titles and abstracts (n = 362)
Records excluded:
Based on titles and abstracts (n = 267)
Reports assessed for eligibility (n = 95)
Reports excluded:
Irrelevant (n = 23)
Not original research (n = 6)
Not full text (n = 38)
Included
Studies included in review (n = 28)

The quality appraisals.



Applications of LLMs in lung cancer.