Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2026; Vol 25: Issue 1, https://carcinogenesis.com/

Validation of LLM-Based Data Curation for Oncology Clinical Trials: A Review encompassing comparison with Manual Abstraction

Karnaditya Rana[1], Shashidar Reddy Abbidi[2], Debashree Sinha[3], Shalmali Joshi[4]

[1]Associate Director, Clinical Data Management | Tempus AI, Texas, USA 75071

dr.karan.rana@gmail.com

ORCID: 0000-0002-5716-217X

[2]Senior Clinical Data Manager

Asreddy.cdm@gmail.com

Charlotte, NC, 28262, Independent Researcher

https://orcid.org/0009-0007-7750-8696

[3]Senior Clinical Data Manager

Debashrees013@gmail.com

Independent researcher, Charlotte,NC, 28262

https://orcid.org/0009-0003-7650-6948

[4]Senior Advanced Analytics Analyst, Atlanta, GA, USA

ORCID: 0009-0000-4329-3841

Email: joshishalmalij@gmail.com

ABSTRACT

The quick growth of clinical data in oncology research has made it difficult to extract and manage data which needs to be processed for clinical trials. The medical field considers manual curation to be the highest standard but its use has become restricted because of time and financial requirements and difficulties with different people judging the same material. Large Language Models (LLMs) which belong to artificial intelligence research work as effective automatic systems which can extract data from unstructured clinical documents that include electronic health records and pathology reports and clinical narratives. This review assesses how well LLMs perform data curation in oncology clinical trials by comparing their results to those of manual data curation processes. This research project has been collecting current evidence about accuracy and efficiency and scalability and reliability and cost-effectiveness of the solution. The study examined validation frameworks together with ethical matters and difficulties which arise during system implementation. Existing studies demonstrate that LLM-based approaches achieve high concordance with manual curation while dramatically reducing processing time and resource utilization. The system suffers from multiple problems which include data heterogeneity and difficulties with understanding data and bias issues and challenges from regulatory bodies. Hybrid human-AI models appear to be the most viable approach for near-term clinical integration. The review presents major advancements in research while showing existing research deficiencies and giving recommendations about how to use LLM-based systems in oncology clinical trial operations.

How to Cite: Karnaditya Rana, Shashidar Reddy Abbidi, Debashree Sinha, Shalmali Joshi, (2026) Validation of LLM-Based Data Curation for Oncology Clinical Trials: A Review encompassing comparison with Manual Abstraction, *Journal of Carcinogenesis, Vol.25, No.1*, 422-430

1. INTRODUCTION

Background

Oncology clinical trials are essential for developing new methods to diagnose and treat cancer which will enhance patient survival rates according to Shams and his colleagues from 2023. The research functions as the foundation to establish new treatments which assist medical professionals in their choices while developing worldwide medical treatment protocols. The success of these trials requires organizations to provide access to topnotch clinical data which has been properly collected and organized. The volume and complexity of oncology data has reached an exceptional level throughout the last ten years. The expansion of this field occurs because organizations widely implement Electronic Health Records (EHRs) and scientists make progress in genomic sequencing technologies while the medical community adopts imaging techniques like radiology and digital pathology (Khan et al., 2025). Clinical datasets now contain both organized information and extensive collections of disorganized information which includes doctor comments and patient discharge documents and medical examination results and medical imaging descriptions. The diverse nature of data sources together with their extensive size creates major difficulties which hinder swift data extraction and standard data processing and data application during clinical trials.

Problem statement

The manual data curation process in oncology clinical trials still operates through traditional methods despite the development of advanced data generation technologies according to Ding et al. The process requires trained clinical or research staff to examine patient records and extract necessary data which they transform into structured formats that enable analysis. The manual curation process serves as the industry standard because it provides clinical information yet the method suffers from major drawbacks according to Feng et al. The process requires extensive time because it takes several weeks to several months to complete large dataset analysis which results in postponed trial start and finish dates. The process demands extensive resources because it needs expert staff members and causes higher operational expenses. The process experiences additional difficulties because different reviewers interpret clinical data in diverse ways which results in different datasets that they create. Oncology clinical research experiences reduced efficiency and limited scalability because of these obstacles.

Emergence of AI and LLMs

Researchers study artificial intelligence (AI) to develop natural language processing (NLP) systems which can extract clinical data from medical records according to Derek and Collings 2025. The first generation of NLP systems combined rule-based systems with traditional machine learning methods which required programmers to create specific features but did not understand contextual information. The introduction of Large Language Models (LLMs) which use transformer architecture technology, marks a major advancement in this particular research field (Raiaan et al., 2024). LLMs analyze extensive textual data while they develop their ability to comprehend intricate language structures and the connections between different language elements and specialized knowledge terms. The healthcare models enable extraction of vital clinical data from unstructured medical records which healthcare professionals use to create summaries that assist their decision-making process. In oncology, LLMs show potential to identify tumor characteristics and treatment methods and determine which patients can participate in clinical trials, which makes them valuable for data curation activities (Chen, Parsa, et al., 2025).

Rationale of the study

Current research shows that LLMs can be used for clinical data curation but the supporting evidence remains incomplete. The study lacks comprehensive analysis which assesses the validity and reliability of LLM-based methods and their effectiveness compared to traditional manual curation methods (Laskar et al., 2024). The essential requirement for data accuracy in oncology trials necessitates a complete examination of LLMs ability to achieve clinical research standards. The study requires a comparative analysis which assesses how LLMs perform in different situations while measuring the balance between productive work and accurate results. The research will assess how LLMs validate data curation methods for oncology clinical trials and compare these results with manual abstraction techniques. The study objectives include measuring LLM-based systems accuracy and reliability while comparing their manual methods performance through efficiency and scalability assessment and identifying main obstacles which need to be solved for future clinical research applications.

2. METHODS

Literature Search Strategy

The research team conducted an extensive literature review through electronic databases which included PubMed and Cochrane Library and IEEE Explore to discover all available peer-reviewed articles and newly published preprint research in this fast-developing scientific domain. The search strategy combined controlled vocabulary and free-text terms related to artificial intelligence, oncology, and data curation. The search engine used the following key terms which included "large

language models" OR "LLMs" and "natural language processing" and "clinical data curation" OR "data extraction" and "oncology" OR "cancer clinical trials" and "manual abstraction" OR "human curation" and "validation" OR "accuracy" OR "comparative study".

Inclusion criteria

Studies were included if they met the following criteria:

- Direct comparison between LLM-based or NLP-based automated curation and manual data abstraction
- Focus on oncology-related datasets, including clinical trials, registries, or real-world cancer data
- Reported quantitative performance metrics such as accuracy, precision, recall, F1 score, or concordance rates
- Evaluated at least one operational parameter such as time efficiency, cost, or scalability
- Provided sufficient methodological detail to allow critical appraisal

Only studies that contributed meaningful comparative insights were selected to support further analysis and discussion in this review.

Exclusion criteria

The following studies were excluded

- Studies not involving clinical or oncology-specific data
- AI studies without a manual comparator or validation reference standard
- Purely methodological or theoretical papers without empirical results
- Studies focusing solely on imaging genomics without text-based data curation
- Non-English publications and conference abstracts full data

Data extraction and Synthesis

Relevant data from included studies were systematically extracted and organized into thematic categories to support comparative analysis. Key variables included:

- Accuracy and performance metrics
- Time efficiency
- Cost implications
- Scalability and dataset size
- Error patterns

Only studies with comprehensive and clearly reported outcomes were included to ensure reliability of subsequent discussion. Extracted data were synthesized qualitatively, with emphasis on identifying trends, strengths, and limitations of LLM-based curation relative to manual methods. This structured approach enabled a focused and evidence-based evaluation aligned with the objectives of this review.

Data Abstraction in Oncology Clinical Trials

Role of Data Curation

Oncology clinical trials require data curation because it enables researchers to create analyzable datasets from unprocessed clinical data (Kumar & Joghee, 2025). The system assesses patient records to determine eligibility by applying various detailed inclusion and exclusion criteria. The process depends on EHRs to provide precise information about patients' demographic details and their clinical and pathological medical history (Singhal et al., 2023). Researchers use curated data to monitor patient outcomes through survival rates and treatment responses and adverse events which serve as the foundation for trial endpoints.

Data curation serves another vital function through its requirement to produce high-quality datasets which regulatory agencies demand for their assessment processes (Feng et al., 2026).Trial results and approval processes face delays because of data that contains either errors or missing information. Clinical data curation serves as the primary method through which oncology research establishes its evidence base. The demand for effective and advanced data curation systems has risen due to the growing use of real-world data according to studies that demonstrate how structured data extraction helps organizations make clinical decisions and conduct extensive oncology investigations.

Types of Oncology Data

Oncology clinical trials require researchers to work with two types of data which include structured data and unstructured data. Structured data include laboratory values, coded diagnosis, and demographic information, which are relatively easy to process (Thatoi et al., 2023). Unstructured formats contain a large amount of clinically important data which includes physician notes and pathology reports and radiology interpretations and discharge summaries.

The detailed information of unstructured clinical text becomes challenging for standardization processes (Seinen et al., 2025). The narrative reports contain essential medical variables which include tumor staging and histological subtype and

treatment response. Studies have shown that data stored in an unstructured way, makes quantitative analysis difficult. The inherent heterogeneity of oncology data requires researchers to work with multiple types of data which include imaging and genomics and longitudinal clinical records (Zhou et al., 2024). The complexity of this process increases the data curation workload while it demonstrates the requirement for sophisticated extraction methods.

Workflow of Manual Curation

The traditional process of manual data curation proceeds through multiple defined steps. The process begins with data abstraction when trained abstractors assess patient records to extract the required variables according to established guidelines (Adamson et al., 2023). The process demands experts to comprehend the clinical context because they need to understand advanced medical terms. The next step involves conducting validation and quality assurance activities which verify the information accuracy and consistency throughout the process. The process requires multiple reviewers to resolve disputes through various adjudication methods. The data reliability in oncology research uses a method that requires independent abstractors to document information while an impartial abstractor verifies their work (Del Campo et al., 2026). The final step requires researchers to transfer their structured data into databases and registries which will be used for subsequent research purposes. The datasets undergo analysis through statistical methods which support both reporting needs and regulatory documentation requirements.

Restrictions of Manual Abstraction

Manual data curation exists as the definitive standard yet it shows multiple limitations. The process requires extensive time because researchers need several months to handle extensive data sets (Ma et al., 2023). A comparative oncology study found that doctors needed about 7 months to finish their manual examination of medical documents (Wiest et al., 2025). The process needs expensive resources because it demands doctors and trained personnel who can make abstracts. Manual processes become unpredictable because different observers understand clinical situations in distinct ways which results in different data collection methods. Interreliability metrics serve as crucial tools to determine how well manual abstraction studies assess their research work. The ability to scale manual curation operations becomes restricted because this method cannot handle extensive data collections that multiple organizations maintain. The increasing volume and complexity of oncology data create a requirement for LLM-based data curation which serves as an automated solution to overcome these existing limitations.

Large Language Models in Clinical Data Curation

Overview of LLM Architecture

Large Language Models (LLMs) function as sophisticated AI systems which operate through transformer architectural frameworks that changed how people use natural language processing in their work. Transformers differ from past approaches which depended on fixed rules or statistical methods because they use attention mechanisms to help models understand how words connect throughout extended texts and build up their comprehension of context. The training of LLMs, which includes BERT and GPT models, uses extensive textual datasets that contain biomedical and clinical materials to improve their ability to understand specialized medical terminology.

Transformer-based models have shown exceptional results when used in healthcare to process different types of data, which includes clinical narratives and HER data. The review of NLA in oncology demonstrates that transformer-based models improve both the speed of data collection and the accuracy of gathering structured information from clinical documents (Hands & Kavuluru, 2025). The introduction of transformer architecture with LLMs has created major progress in clinical NLP tasks, which include data extraction and text understanding (Chen, Alnassar, et al., 2025). The models use pre-training on extensive datasets which allows users to adapt the models through fine-tuning or prompting methods for their specific oncology data abstraction needs.

Capabilities Relevant to Healthcare

LLMs possess several capabilities that are particularly relevant to clinical data curation. Natural language understanding serves as the principal function because it allows the model to comprehend intricate clinical documentation which includes abbreviations and synonyms together with context-dependent meanings. This requirement becomes essential for oncology because different medical facilities and healthcare workers maintain their unique documentation approaches.

The first flexible feature of LLMs enables them to extract entities, which lets them recognize and categorize clinical entities that include diagnoses and medication and tumor characteristics and procedures. The LLM-based systems extract cancer-related variables from unstructured text which decreases the need for manual abstraction. The scoping review points out that LLMs can extract vital patient data from EHR systems through automatic data extraction which includes important patient features (Li et al., 2024).

The LLMs exhibit strong contextual reasoning abilities, which enable them to grasp how different entities connect with each other in various contexts. The system uses its ability to recognize present and previous diagnosis information while it

detects negations and establishes temporal connections. The research demonstrated that transformer-based models reached high performance levels when extracting clinical information through contextual tasks which involved medication and adverse event extraction and achieved F1 scores above 0.93. The capabilities of LLMs make them highly effective for extracting intricate oncology information from narrative text sources.

Applications in Oncology

The field of oncology has seen increasing usage of LLMs as they perform multiple tasks related to data collection and organization. The main function of this system involves tumor staging extraction which enables models to extract staging details from both pathology reports and clinical documentation. The automated pipelines which use LLM technology have demonstrated their ability to extract staging information through their work with oncology records which produces structured data (Lee et al., 2025). The second major function of this system involves treatment identification which enables LLMs to extract details about chemotherapy regimens and surgical procedures and radiation therapy from clinical documentation. The system enables researchers to monitor treatment trends which occur during clinical investigations and through actual patient data collection.

LLMs serve the purpose of detecting adverse events through their capability to recognize complications and treatment-related toxicities that occur in clinical narratives. Transformer-based models have shown strong performance in extraction medication-related events and adverse drug reactions from unstructured data (Chen, Parsa, et al., 2025). LLMs function as crucial tools which enable researchers to determine clinical trial eligibility through their ability to compare patient information with detailed eligibility requirements. LLMs work through patient record analysis and trial protocol examination to help identify appropriate candidates for oncology studies which results in better recruitment outcomes.

Workflow of LLM-based Abstraction

The LLM-based data curation workflow follows a structured pipeline which begins with data input and proceeds through three processing stages before reaching its final output. The model receives unstructured clinical data which includes EHR notes and pathology reports and imaging narratives as input. The LL system examines the text based on its existing knowledge and contextual information to find important entities and their corresponding connections between them. The system uses named entity recognition and relation extraction as NLP techniques during this processing stage. The structured output contains extracted information which transforms into standardized variables and database entries. This system enables seamless integration with clinical trial systems and statistical analysis software. The output undergoes assessment against established standards through performance evaluation which includes accuracy measurement and F1 score testing. The validation process confirms clinical evidence through testing while it reveals which areas need development work. The framework transforms unstructured oncology data into usable formats through an efficient process which eliminates the need for manual curation. Recent studies show that LLM-based workflows effectively automate repetitive data extraction tasks while achieving precise and consistent results.

Validation frameworks for LLM-Based Curation

Importance of Validation

Validation functions as the essential condition which must be fulfilled before hospitals will use LLMs for their oncology data curation work. The clinical reliability of data needs to be established because it has direct implications for clinical trial results and treatment options and submissions to regulatory agencies. The process of data extraction becomes problematic when it produces either wrong information or inconsistent results because this situation creates the potential for dangerous outcomes which will damage both patient safety and research legitimacy. A recent validation study emphasized manual chart review, although reliable, time-consuming and prone to human error, highlighting the need for validated automated alternatives (Dagli et al., 2024). The process of validation holds equal importance to the process of regulatory compliance. The clinical trial data must fulfill the strict regulations which regulatory authorities established because these authorities need convincing proof that all LLM-extracted data maintain accuracy and reproducibility and show no bias. A scoping review of LLM applications in oncology noted that further studies are needed to evaluate the performance and assess real-world applicability (Chen, Alnassar, et al., 2025). The complete validation process needs to be established first before we can use the system in healthcare.

Validation Metrics

LLM-based abstraction validation used established quantitative performance metrics as its testing standard. Accuracy serves as the primary measurement used by researchers to assess how well their study results match a reference standard which usually consists of manually curated data. A recent oncology-focused validation study demonstrated that LLM-based extraction reached an accuracy level of over 98 to 99 percent when they compared their results with human extraction (May et al., 2025). The two fundamental metrics require evaluation through precision which measures the percentage of true positive results and recall which measures the percentage of actual positive results that were correctly identified. The F1 score combines these metrics to create a model performance assessment tool that evaluates results with equal weight. F1 scores above 0.85 to 0.90 in clinical NLP applications serve as reliable indicators for accurate extraction results.

Concordance functions as another vital metric which assesses how well LLM outputs match human-created annotations. In oncology concordance functions as an essential requirement because tumor staging and treatment response assessment need standardized transformation across diverse treatment programs.

Levels of Validation

The process of validating LLM-based data curation requires multiple evaluation levels to achieve complete assessment.

- The process of variable-level validation tests the precision of each data component which includes tumor stage and diagnosis and treatment type. This level helps identify specific variables where the model performs well or poorly.
- The record-level validation process checks whether essential data points from one patient record have been accurately retrieved. This provides insight into the model's ability to maintain contextual consistency across a clinical document.
- The dataset-level validation process assesses the entire dataset quality which includes efficiency and internal consistency and usability for subsequent analysis.

The recent oncology review found that LLM studies evaluate their performance through multiple datasets and different variables which shows their ability to apply their results across various cancer types and clinical situations (Chen, Alnassar, et al., 2025).

Internal VS External Validation

Validation strategies can be broadly categorized into:

- Internal validation This method tests the LLM using data from the same institution or dataset that developers used to create their model. The method allows researchers to evaluate their results but produces an erroneous assessment of their work because the system uses familiar data.
- External validation The model undergoes evaluation through testing on datasets that researchers gathered from different institutions and various population groups. This process establishes the extent to which test results can apply to actual life situations.

In oncology research, researchers now use multiple institutional datasets to conduct validation studies. A scoping review found that a majority of LLM studies (75%) utilized multi-center datasets which show that external validation plays a vital role in establishing research credibility through various clinical environments (Chen Alnassar et al. 2025).

Emerging Validation Frameworks

The researchers develop assessment methods which assess LLM-based clinical data extraction through standardized testing procedures. The framework employs three testing elements together with expert human abstraction assessment and internal consistency evaluation and cross-dataset assessment for its multi-dimensional validation process. The research team developed these frameworks to provide solutions which LLMs need to overcome their operational challenges of hallucination and bias and contextual errors. The newest validation approaches assess model performance through benchmark datasets which establish standardized reference points for different research evaluations. Researchers use such datasets to create standard testing methods which enable them to compare different models while maintaining test results from previous experiments.

New research studies demonstrate that LLMs require ongoing verification which needs LLMs to undergo constant evaluation throughout their operational lifespan. The need for ongoing validation particularly applies to oncology because both clinical procedures and documentation methods undergo rapid changes. The development of standardized validation protocols functions as an essential requirement which establishes trust in LLM-based systems and supports their usage in clinical trial operations.

Comparative analysis: LLM vs Manual Abstraction

Accuracy and Concordance

The accuracy of LLM-based curation tests which compare LLM-based curation methods with manual abstraction methods shows high accuracy but their results differ according to the difficulty of the tested data. A landmark comparative study in breast oncology reported an overall LLM accuracy of 90.8%, which closely matched the results of physician review. The study found strong agreement across key clinical variables (Kang et al., 2025). The validation study found that LLM-based extraction achieved >98 – 99% accuracy across multiple variables, which demonstrated that LLMs extract structured data with accuracy that matches human performance (May et al., 2025).

LLMs demonstrate their best performance when they handle structured variables which medical documents record at exact times. The theory performance decreases when it needs to operate with complex variables which require multiple data elements to be combined for analysis. LLM outputs showed a higher rate of missing data for staging variables, with 12.2% missing compared to 3.1% missing for manual abstraction (Kang et al., 2025). The analysis of concordance shows that LLMs provide consistent results but human experts still outperform LLMs in interpreting complex clinical situations.

Efficiency and Time Comparison

The main benefit of using LLM-based curation systems results in better operational efficiency. The oncology study demonstrated that manual curation needed 1025 physician hours for 7 months of work while LLM-based processing required 12 days and 96 hours of physician oversight which led to a 91 to 95 percent reduction in time and effort (Kang et al., 2025). The research shows LLMs transform clinical trials through faster processes that replace human data extraction tasks.

Cost Analysis

The analysis of economic impacts between human-based work and LLM-based abstracting methods shows that automated systems deliver considerable financial advantages. The manual curation process demands multiple trained clinicians and EHR access infrastructure and administrative support which results in extensive operational expenditures. LLM-based systems need less human resources than companies use for their regular manual operations.

The oncology study determined that LLM-based processing costs about 260 for 1734 cases which shows a high-cost efficiency (Kang et al., 2025). The reduction of physician duties leads to total cost reductions which occur because operational needs have decreased. Organizations need to invest money in infrastructure expenses during model deployment because they require resources for processing and data handling and ongoing maintenance. Organizations achieve better cost-effectiveness through LLM-based systems because these systems function well with large datasets.

Scalability

LLM-based systems have better scalability than manual abstraction systems which face major restrictions. Manual approaches are inherently constrained by human capacity, making them unsuitable for large, multi-institutional datasets. LLMs can process extensive data sets from numerous sites while needing only few system resources. A multi-center oncology study involving data from five academic hospitals, demonstrated LLM-based systems can effectively handle large datasets while maintaining accuracy and consistency (Kang et al., 2025). The LLM system integrates multiple data types which include clinical notes and pathology reports and registry data to operate effectively in big data environments. LLM systems need to scale their operations because modern oncology research requires handling increasing data volumes which result from precision medicine advancements and real-world evidence creation.

Table 1. Comparison of Manual vs LLM-based Abstraction

Parameter	Manual Abstraction	LLM-Based Abstraction
Accuracy	High (gold standard)	Comparable (90 – 99%)
Time	Weeks to months	Hours to days
Cost	High	Low per case
Scalability	Limited	High
Consistency	Variable	High
Suitable condition	Complex Integration	Large-scale structured extraction

Hybrid Human-AI Models

Concept of Human-in-the-Loop

The hybrid human-AI model which people commonly call human-in-the-loop systems uses automated large language model outputs together with human expertise to improve clinical data curation processes (Srivastava et al., 2025). The initial extraction of structuredvariables from unstructured oncology data through LLMs uses clinical notes and pathology reports as their source material. Human experts review the extracted outputs to validate them and make corrections when required. The iterative process provides ongoing feedback which helps improve both model performance and data quality throughout the entire process. The systems show their highest effectiveness in oncology because that field needs doctors to interpret complex patient data.

Advantages

The hybrid human-AI model provides multiple benefits which surpass the advantages of both complete manual work and total machine automation (Sauer & Burggräf, 2025). First, it leads to improved accuracy. Human oversight works together with LLMs which efficiently extract data from large volumes of information to guarantee accurate understanding of complex clinical data points. Studies evaluating AI-assisted data abstraction have shown that combining automated extraction with expert validation significantly enhances overall data reliability comparedto standalone methods. The second advantage of hybrid systems lies in their ability to decrease operational demands. The team members have experienced a significant reduction in work requirements because they can use LLMs to handle their repetitive tasks which involve processing clinical notes and gathering essential data (Al-Abdulkarim et al., 2025). Humans who are experts in their field use this system which helps them produce better results through better decision-making and case evaluation. Third, hybrid approaches foster enhanced trust and acceptance. The absence of transparent information together with doubts about system

dependability serves as the primary obstacle which prevents healthcare organizations from embracing artificial intelligence technology. Human validation processes create trust among doctors and regulatory authorities because those processes ensure system accountability while allowing clinical supervision.

Real-World Use Cases

Hybrid human-AI models are increasingly being implemented in semi-automated curation systems with oncology research and clinical trials (Stenhouse et al., 2025).Multiple real-world studies have used LLM-based pipelines to extract oncology variables which were then verified for accuracy and completeness by clinicians through manual checking. The systems showed substantial reductions in curation time, while achieving high matching rates with complete manual datasets. Hybrid systems which process extensive patient information from cancer registries and multi-institutional research databases enable efficient creation of real-world evidence. LLMs function as pre-screening tools for patient eligibility in clinical trial recruitment, while human reviewers make final decisions about patient eligibility after LLM testing (Callies et al., 2025). The combination of hybrid human-AI systems with LLM-based data curation for oncology workflows creates a practical solution which achieves operational efficiency and trustworthy clinical results.

3. CONCLUSION

The introduction of LLM-based data curation creates an entirely new approach for conducting oncology clinical trials. Manual curation which has been considered the most accurate method for a long time now shows its limitations because it cannot handle large numbers of items and needs too much time to complete tasks. LLMs provide an effective solution because they deliver precise results while users complete tasks more quickly than before. The process of switching from manual work to automated systems requires careful management. Verification functions as a vital process which establishes product trustworthiness together with operational security. The upcoming period will see hybrid models which integrate both methods achieve widespread adoption. LLM-based data curation has the potential to revolutionize oncology clinical trials through its ability to enhance operational efficiency while increasing research capacity and reducing costs. The current evidence supports their application however three main issues which include accuracy, interpretability, and regulatory requirements need to be resolved. Hybrid human-AI systems represent the most effective approach because they deliver both operational efficiency and reliable system performance. Research and validation efforts need to continue indefinitely.

REFERENCES

[1] Adamson, B., Waskom, M., Blarre, A., Kelly, J., Krismer, K., Nemeth, S., Gippetti, J., Ritten, J., Harrison, K., & Ho, G. (2023). Approach to machine learning for extraction of real-world data variables from electronic health records. *Frontiers in Pharmacology*, *14*, 1180962.

[2] Al-Abdulkarim, M., Bakouri, M., & Alassaf, A. (2025). A Metrics-Driven Approach to Develop a Hybrid Model of Staffing and Workload Balance in the NGHA Hospitals. *Journal of Healthcare Leadership*, 395-416.

[3] Callies, A., Bodinier, Q., Ravaud, P., & Davarpanah, K. (2025). Real-world validation of a multimodal LLM-powered pipeline for high-accuracy clinical trial patient matching. *Communications Medicine*.

[4] Chen, A., Yu, Z., Yang, X., Guo, Y., Bian, J., & Wu, Y. (2023). Contextualized medication information extraction using transformer-based deep learning architectures. *Journal of biomedical informatics*, *142*, 104370.

[5] Chen, D., Alnassar, S. A., Avison, K. E., Huang, R. S., & Raman, S. (2025). Large language model applications for health information extraction in oncology: scoping review. *JMIR cancer*, *11*, e65984.

[6] Chen, D., Parsa, R., Swanson, K., Nunez, J.-J., Critch, A., Bitterman, D. S., Liu, F.-F., & Raman, S. (2025). Large language models in oncology: a review. *BMJ oncology*, *4*(1), e000759.

[7] Dagli, M. M., Ghenbot, Y., Ahmad, H. S., Chauhan, D., Turlip, R., Wang, P., Welch, W. C., Ozturk, A. K., & Yoon, J. W. (2024). Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review. *Scientific reports*, *14*(1), 26783.

[8] Del Campo, A. O. B., Lituiev, D., Varma, G., Manoharan, M., Kumar Ravi, S., Aman, A., Kansagra, A., Greshock, J., Venkatakrishnan, A., & Batavia, A. S. (2026). Automated abstraction of clinical parameters of multiple myeloma from real-world clinical notes using large language models. *BMC Medical Informatics and Decision Making*.

[9] Derek, V., & Collings, P. (2025). Natural Language Processing (NLP) in Healthcare AI: Enhancing Clinical Insight Extraction from Unstructured Patient Data. *Authorea Preprints*.

[10] Ding, L., Bradford, C., Kuo, I.-L., Fan, Y., Ulin, K., Khalifeh, A., Yu, S., Liu, F., Saleeby, J., & Bushe, H. (2022). Radiation oncology: future vision for quality assurance and data management in clinical trials and translational science. *Frontiers in Oncology*, *12*, 931294.

[11] Feng, S., McMahan, C., & Hu, A. (2026). Data Curation. In *The Radiology AI Handbook* (pp. 163-186). Elsevier.

[12] Hands, I., & Kavuluru, R. (2025). A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications. *Artificial Intelligence Review*, *58*(10), 314.

[13] Kang, Y.-J., Lee, H., Yi, J. P., Kim, H., Yoon, C. I., Baek, J. M., Kim, Y.-s., Jeon, Y. W., Rhu, J., & Lim, S. H. (2025). Large Language Model Versus Manual Review for Clinical Data Curation in Breast Cancer: Retrospective

Comparative Study. *JMIR Medical Informatics*, *13*, e73605.
[14] Khan, S. N., Danishuddin, Khan, M. W. A., Guarnera, L., & Akhtar, S. M. F. (2025). Multi-modal AI in precision medicine: integrating genomics, imaging, and EHR data for clinical insights. *Frontiers in Artificial Intelligence*, *8*, 1743921.
[15] Kumar, R. M. R., & Joghee, S. (2025). A review on integrating breast cancer clinical data: A unified platform perspective. *Current Treatment Options in Oncology*, *26*(1), 1-13.
[16] Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C. W., & Parvez, M. R. (2024). A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,
[17] Lee, D., Vaid, A., Menon, K. M., Freeman, R., Matteson, D. S., Marin, M. L., & Nadkarni, G. N. (2025). Using large language models to automate data extraction from surgical pathology reports: retrospective cohort study. *JMIR Formative Research*, *9*(1), e64544.
[18] Li, L., Zhou, J., Gao, Z., Hua, W., Fan, L., Yu, H., Hagen, L., Zhang, Y., Assimes, T. L., & Hemphill, L. (2024). A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*.
[19] Ma, P., Ding, R., Wang, S., Han, S., & Zhang, D. (2023). Insightpilot: An llm-empowered automated data exploration system. Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations,
[20] May, P., Greß, J., Seidel, C., Sommer, S., Schuler, M. K., Nokodian, S., Schröder, F., & Jung, J. (2025). Enabling Just-in-Time Clinical Oncology Analysis With Large Language Models: Feasibility and Validation Study Using Unstructured Synthetic Data. *JMIR Medical Informatics*, *13*, e78332.
[21] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, *12*, 26839-26874.
[22] Sauer, C. R., & Burggräf, P. (2025). Hybrid intelligence–systematic approach and framework to determine the level of Human-AI collaboration for production management use cases. *Production Engineering*, *19*(3), 525-541.
[23] Seinen, T. M., Kors, J. A., van Mulligen, E. M., & Rijnbeek, P. R. (2025). Using structured codes and free-text notes to measure information complementarity in electronic health records: feasibility and validation study. *Journal of Medical Internet Research*, *27*, e66910.
[24] Shams, M., Abdallah, S., Alsadoun, L., Hamid, Y. H., Gasim, R., Hassan, A., & Hassan, A. (2023). Oncological horizons: The synergy of medical and surgical innovations in cancer treatment. *Cureus*, *15*(11).
[25] Singhal, P., Tan, A., Drivas, T., Johnson, K., Ritchie, M., & Beaulieu-Jones, B. (2023). Opportunities and challenges for biomarker discovery using electronic health record data. *Trends in molecular medicine*, *29*(9), 765-776.
[26] Srivastava, T., Irfan, H., Babiy, V., & Swami, S. (2025). Integration of Generative AI with Human Expertise in HEOR: A Hybrid Intelligence Framework. *Advances in Therapy*, *42*(9), 4103-4130.
[27] Stenhouse, A., Fisher, N., Lepschi, B., Schmidt-Lebuhn, A., Rodriguez, J., Turco, F., Reeson, A., Paris, C., & Thrall, P. H. (2025). A vision of human–AI collaboration for enhanced biological collection curation and research. *Bioscience*, *75*(6), 457-471.
[28] Thatoi, P., Choudhary, R., Shiwlani, A., Qureshi, H. A., & Kumar, S. (2023). Natural language processing (NLP) in the extraction of clinical information from electronic health records (EHRs) for cancer prognosis. *International Journal*, *10*(4), 2676-2694.
[29] Wiest, I. C., Wolf, F., Leßmann, M.-E., van Treeck, M., Ferber, D., Zhu, J., Boehme, H., Bressem, K. K., Ulrich, H., & Ebert, M. P. (2025). A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *npj Precision Oncology*, *9*(1), 313.
[30] Zhou, H., Zhou, F., Zhao, C., Xu, Y., Luo, L., & Chen, H. (2024). Multimodal data integration for precision oncology: Challenges and future directions. *arXiv preprint arXiv:2406.19611*.