ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO REAL WORLD DATA AND DIGITAL ONCOLOGY

PERSPECTIVES

AI assistance in tumor multidisciplinary teams

R. J. Geukes Foppen[1]*, M. Morkūnas[2], A. Traverso[3,4] & R. Dienstmann[5,6]

[1]Scientific Director Office IRCCS Regina Elena National Cancer Institute, Rome, Italy; [2]Vilnius Santaros Klinikos Biobank, Hematology, Oncology and Transfusion Medicine Center, Vilnius University Hospital Santaros Klinikos, Vilnius, Lithuania; [3]Digital Institute for Cancer Outcomes Research (DIGICORE), Brussels, Belgium; [4]S-RACE (San-Raffaele Ai CEnter), University Vita-Salute San-Raffaele Hospital, Milan, Italy; [5]Oncology Data Science, Vall d'Hebron Institute of Oncology, Barcelona; [6]OC Medicina de Precisão, Oncoclínicas&Co, São Paulo, Brazil

Available online xxx

Multidisciplinary teams (MDTs), the cornerstone of modern cancer care, are facing significant operational inefficiencies. These challenges include the laborious, manual synthesis of unstructured, multimodal patient data for case preparation, which is time-consuming and prone to information overload. Furthermore, heterogeneous operational processes across MDTs themselves compound these issues. Additionally, clinical decisions are often archived in static documents, preventing the systematic collection of decision rationales essential for continuous learning and research. We propose that artificial intelligence (AI), particularly natural language processing (NLP) and large language models (LLMs), can act as integrated partners to solve these problems. The capacity of AI to seamlessly integrate diverse datasets—including imaging, histopathology, genomics, and clinical data—may be instrumental in enhancing diagnostic accuracy, refining personalized treatment plans within a complex cancer management journey. This integration can be achieved through a tiered approach, utilizing models from small NLP for targeted information extraction to foundational generative NLP for complex evidence synthesis, while addressing key challenges in validation, ethical governance, and regulatory oversight. International initiatives are actively developing validated frameworks to facilitate the widespread and standardized adoption of these AI solutions, while taking into account heterogeneous operational processes. By improving data management, streamlining decision making, and establishing crucial feedback loops, AI integration promises to enhance patient outcomes and optimize resource utilization within cancer care.

Key words: AI, NLP, LLM, multidisciplinary teams, clinical decision support system, oncology

INTRODUCTION

Multidisciplinary teams (MDTs) are a cornerstone of modern cancer care and have become standard practice for 25 years,[1] uniting diverse specialists to discuss complex patient cases and formulate optimal treatment plans. While crucial for navigating oncology's complexities, MDTs face significant inefficiencies due to burdensome case preparation, complex decision making, and a lack of systematic documentation. The process is resource-intensive and time-consuming, with clinicians potentially spending hours preparing for a single patient case by manually synthesizing vast amounts of disparate, unstructured data from various sources like clinical notes, pathology slides, genomic data, and radiology reports (Figure 1, panel EHR). Such manual review leads to information overload and cognitive burden, and without dedicated digital frameworks, the subjective coordination of this complex data strains MDT resources. The time available for in-depth discussion of patient care is thereby negatively impacted. The problem is exacerbated by the fact that valuable input data often comes from unstructured, sometimes paper-based, sources, and decisions are archived in static formats like PDFs. This prevents the systematic collection of the decision rationale, which is crucial for continuous learning and improvement. The lack of structured data, along with heterogeneous operational processes,[2] also diminishes the utility of real-world data for research, and hinders the delivery of truly personalized care.[1]

Artificial intelligence (AI) technologies are transforming cancer care by significantly enhancing data management efficiency,[3] augmenting clinical decision making,[4] and facilitating increasingly personalized patient treatment approaches.[5] Among AI technologies, rapid advancements in natural language processing (NLP) and large language models (LLMs), which are, respectively, computational techniques that enable computers to understand and

**Correspondence to:* Dr Remco Jan Geukes Foppen, Scientific Director Office IRCCS Regina Elena National Cancer Institute, Via Chianesi 53, 00144 Rome, Italy. Tel: +39 06 5266 2728

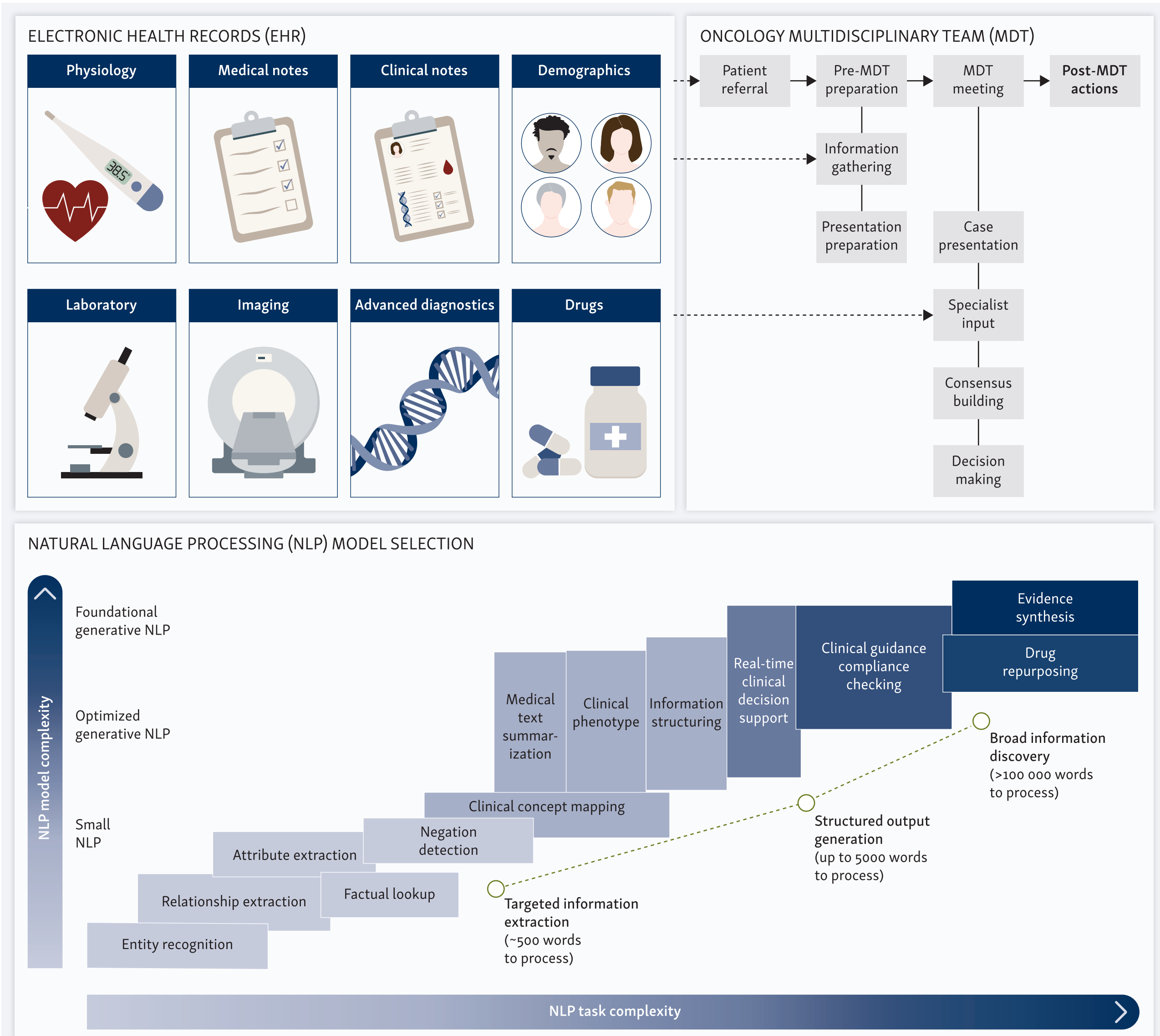

Figure 1. Overview of data flow and NLP integration in oncology multidisciplinary team (MDT) meetings. This figure illustrates the journey of patient data from electronic health records (EHRs) through an oncology MDT process, highlighting critical junctures for NLP integration.
EHR panel (left): depicts the diverse sources of patient data within an EHR system. Patient data typically flow from various EHR sources, into MDT stages via direct manual access by MDT members within the EHR system. Increasingly, the data flow in MDTs often encompasses multimodal data (e.g. clinical notes, imaging reports, laboratory results).Oncology MDT panel (middle): details the sequential workflow of an oncology MDT, emphasizing points where NLP can enhance efficiency and decision making. Critical points of NLP integration are highlighted, aiming to enhance information flow and decision support. Solid lines depict the established, internal workflow steps and direct data flow exclusively within the MDT process. Dotted lines represent the points where EHR data are accessed and typically processed manually. NLP model selection panel (right): categorizes NLP models based on their complexity and task capabilities, offering a tiered approach for tool selection in MDT applications. The horizontal axis, 'NLP task complexity', signifies the increasing complexity of tasks from left to right. The vertical axis, 'NLP model complexity', indicates the increasing complexity of the models from bottom to top. Three tiers are suggested (small NLP, optimized generative NLP, and foundational generative NLP models) to guide the selection of appropriate tools for applications in the MDT. The dotted line in lower panel illustrates the increasing scale of text processing capacity with higher NLP model complexity. A fuller value continuum of the three tiers is displayed in Table 1. LLMs leverage various NLP techniques as their foundation, and represent the state-of-the-art in foundational generative NLP. While foundational generative NLP might not require the sheer scale or emergent capabilities of modern LLMs, its implementation within MDTs can lead to more manageable practical challenges, including data privacy concerns, regulatory hurdles specific to AI in clinical practice, user adoption, and the ongoing need for maintenance and updates of AI models.
AI, artificial intelligence; LLM, large language models; NLP, natural language processing.

process human text, and sophisticated AI models trained on vast amounts of text data to generate human-like text responses, are fundamentally shifting their role in MDTs from mere assistive tools (Figure 1, panel oncology MDT) to integrated partners in decision making.[6] In this specific context, NLPs and LLMs will potentially refine cancer care by efficiently structuring unstructured clinical text and synthesizing vital information.[7] This allows them to accurately identify and categorize crucial details, such as cancer staging, exposure to prior therapies, and laboratory and molecular tests results, directly from clinical notes or PDF reports. Nevertheless, this evolution necessitates a

Table 1. Tiered language processing value continuum: from extraction to synthesis

	Small NLP	Optimized generative NLP	Foundational generative NLP
Performance metrics	Standard metrics	Standard and qualitative metrics	Primarily qualitative metrics
Scalability and efficiency	High efficiency for specific tasks. Low scalability and cost	Moderate efficiency and scalability. Moderate cost	Low efficiency, high scalability. Very high cost
Accuracy and reliability	High accuracy on specific tasks. Brittle with new data	Good accuracy within fine-tuned domain. More robust than small NLP	High variability. Risk of fluent but incorrect information
Usability in health care	Ideal for specific, repetitive tasks. Needs domain experts	Good for summarization, data structuring, decision support	Broad use for complex tasks (problem-solving). Requires validation
Flexibility and computational cost	Low flexibility (one task/model). Very low cost	Moderate flexibility (adaptable). Moderate cost	Very high flexibility (multitask). Extremely high cost
Risk of hallucination	Very low. Extractive and non-generative	Moderate risk. Can generate plausible but incorrect information	High risk. Needs robust validation
Ease of capture of temporality	Requires manual engineering. Challenging	Better with fine-tuning on relevant data	Excellent. Requires careful prompting

Small NLP: represents foundational NLP techniques designed for specific, less complex tasks. Entity recognition, relationship extraction, negation detection, attribute extraction, factual lookup, clinical concept mapping: examples of targeted information extraction tasks, typically processing up to ∼500 words. Optimized generative NLP: refers to more advanced generative models capable of handling moderately complex tasks. Medical text summarization, information structuring, clinical guideline compliance, clinical decision support: tasks involving processing up to 5000 words. Foundational generative NLP: encompasses state-of-the-art models, including LLMs, suited for highly complex tasks. Evidence synthesis, repurposing: represent broad information discovery tasks, capable of processing over 100 000 words. Evaluation Metrics:
Standard metrics: F1-score, precision, recall. Qualitative metrics: subjective, human-based assessments of AI output. They often evaluate qualities like coherence, relevance, and factual accuracy, which are difficult to capture with a simple numerical score.
AI, artificial intelligence; LLMs, large language models; NLP, natural language processing.

proactive approach to evaluation, ethical governance, and regulatory oversight to ensure patient safety and maintain trust.[8]

THE OPERATIONAL PROCESS FOR AI IMPLEMENTATION IN MDTS

Utilizing models of different complexity tiers, namely small NLP, optimized generative NLP, and powerful foundational generative NLP, an NLP-assisted workflow can be designed to enhance the MDT process (Table 1). The choice of the NLP model and its corresponding validation approach are critical and vary based on the task's complexity and the nature of the desired output.[9-13]

The NLP model selection can be envisioned as a dynamic library of NLP methods, selectable to construct tailored workflows (Figure 1, panel NLP model selection). Following the principles laid out here and given the variability of MDT operational processes, these workflows can in principle be extended to any MDT beyond oncology. Raw clinical data from electronic health records (EHRs), including patient status reports, laboratory test results, and detailed pathology reports are routed to specialized NLP modules based on the specific information extraction and synthesis tasks required.

The first option for workflow construction, targeted information extraction, utilizes models belonging to the small NLP tier. As demonstrated in Figure 1 (panel NLP model selection), these models excel at various information extraction tasks, including named entity recognition (NER),[14-16] question answering (Q/A),[17-19] relation extraction,[20-22] and classification.[23,24] For instance, small NLP models have demonstrated strong performance across diverse tasks and data scales. A convolutional neural network-based small NLP model achieved high accuracy across five information extraction tasks from a large dataset of over 23 000 pathology reports.[25] Furthermore, sophisticated training techniques have been shown to improve the performance of small NLP models when assigning ICD-O-3 codes to pathology reports.[26] The potential of such models for underrepresented languages was recently by training a Spark NLP model on just 100 manually annotated Romanian colonoscopy reports and still achieving excellent performance for NER.[27] These models are typically trained on standard servers (i.e. cost-effective systems equipped with powerful consumer-grade components, having a shorter life span) with graphical processing units (GPUs) and, once optimized, can be deployed for real-time inference on more accessible central processing unit (CPU)-based systems or even edge devices. Their validation is straightforward, relying on traditional performance metrics such as accuracy and F1-score, which are calculated automatically on a held-out, human-annotated dataset. The output is a single, verifiable data point.[9,28] Such projects can now be implemented without a huge investment into infrastructure or computational power, as the necessary hardware may already be available within the hospital and the expertise can often be found within its IT staff with recent university training.

Another option for workflow construction, structured output generation, employs optimized generative NLP models to carry out more complex contextual understanding and data structuring. These models analyze combined inputs, such as structured laboratory data and unstructured patient reports (from various linguistic contexts), to identify and format clinically relevant information. A prime example is the identification and structuring of potential adverse events, like neutropenia or thrombocytopenia, into a standardized format, such as JSON, ready for integration into the MDT's review tools. This process heavily relies on sophisticated prompt engineering, where carefully crafted instructions guide the model to extract and format information with fewer data annotations required, as annotations are primarily needed for validation rather than training. These optimized generative NLP models are designed to run locally on capable servers (i.e. high-speed, reliable, and scalable systems equipped with high-end consumer or enterprise-grade components, designed for continuous use), potentially with GPUs (for parallel processing resulting in faster response times). Crucially, while optimized generative NLP can also handle tasks traditionally carried out by small NLP models, such as NER, this flexibility comes at a significantly higher computational cost, representing a key trade-off for broader applicability. However, for tasks requiring a broader contextual understanding or more flexible output formats, they offer significant advantages without the need for specific training data for each new task.[29-31] Studies have demonstrated practical validation approaches for using optimized generative NLP in NER tasks with sophisticated prompt engineering, often combining automated metrics (achievable with annotated validation data)[32] with automated checks for the required output structure (e.g. validating against a JSON schema)[11] and qualitative human review for nuances.[33,34] Given the complexity and potential clinical impact, the validation of structured outputs from these models involves rigorous manual review by clinicians to verify the relevance, accuracy, and adherence to clinical guidelines (e.g. toxicity grading).[35,36]

For complex tasks, the system can be integrated with foundational generative NLP models. These powerful models are often cloud based due to their high computational demands. A promising evolution is agentic AI, where the NLP/LLM acts as a central 'brain'. This brain autonomously selects and coordinates specialized tools to carry out multi-step tasks. These AI agents are designed to function with a high degree of autonomy, capable of analyzing a clinical case, identifying the necessary steps for a full evaluation, and then sequentially using the appropriate tools—such as segmenting a medical image or extracting structured data from a report—to gather the required information.[27,37] As demonstrated by Ferber et al.,[38] this integration of an LLM with a toolkit of specialized functions drastically improves the ability to generate precise solutions for complex, realistic medical cases, moving beyond simple data structuring to a more comprehensive and autonomous analysis. As with many AI applications, it might be too early to deploy such autonomous systems in clinical practice, given ongoing concerns about their reliability, transparency, and accountability. Ensuring the safety and validity of these 'black box' systems, alongside establishing clear regulatory frameworks, remains a critical challenge before they can be responsibly integrated into patient care pathways.[39]

While similar in principle to prompting optimized generative NLP, leveraging foundational generative NLP models often involves more intricate and extensive prompt engineering to harness their vast general knowledge, advanced reasoning capabilities, and huge context windows (often exceeding 100 000 words). Furthermore, these models, accessed via application programming interface, offer tools for enhanced validation. For instance, one can prompt the model to verify its generated interpretations of molecular test results against up-to-date online knowledge bases or to cross-reference potential off-label treatment options with current research literature.[40-42] They excel at synthesizing information across diverse and voluminous datasets, such as comprehensive case summaries or insights derived from an entire patient history. For such high-level, generative outputs, validation methods include qualitative human evaluation by multiple independent clinical raters, assessing aspects like coherence, relevance, factual accuracy, and completeness, as automated metrics alone are insufficient to capture clinical utility (Table 1). The necessity for different validation approaches underscores that the reliability of NLP in clinical settings is directly tied to the complexity and criticality of the task it carries out.[30,43,44]

SETTING THE PACE—FROM HETEROGENEITY TO AI-DRIVEN OPERATIONAL EXCELLENCE IN MDTS: ENHANCING PATIENT OUTCOMES

The integration of AI within MDTs presents a transformative opportunity to significantly enhance cancer care. By streamlining operational processes, boosting decision-making efficiency, and establishing crucial feedback loops for continuous learning, NLP can profoundly support MDTs both operationally and scientifically, ensuring that human clinical expertise remains central. This approach promises not only improved patient outcomes, but also a more efficient utilization of resources across health care systems.

However, the inherent heterogeneity in MDT organization across different countries and hospitals poses a significant challenge to the widespread and standardized implementation of AI solutions. This organizational variability, coupled with the prevalent reliance on unstructured data sources and static documentation, critically impedes the systematic collection of decision rationale and hinders the essential feedback loops for continuous learning and improvement.

To address these challenges and foster the necessary trust for regulatory acceptance of AI in clinical decision making, several international initiatives are emerging. In Europe, the Digital Institute for Cancer Outcomes Research (DIGICORE) and its DigiONE network (DIGItal Infrastructure for ONcology) are at the forefront of leveraging AI technologies to improve precision oncology in Europe.[45] While a primary focus of DIGICORE, and indeed a broader movement within health care, is on structuring real-world data for secondary use (e.g. for research, outcomes analysis, and benchmarking across health care systems, often utilizing frameworks like Minimal Essential Description of Cancer and Observational Medical Outcomes Partnership), the power of their approach extends further. By addressing the complexities of unstructured clinical notes through NLP technologies, the very same DIGICORE NLP technologies employed to extract and harmonize critical insights for research protocols and health quality indicators can be readily adapted to structure data for primary use, directly enhancing vital clinical processes such as MDTs.

DIGICORE's roadmap includes defining and executing, together with health care professionals, a survey to map current MDT practices and establish a crucial operational baseline. The concept of a 'placebo' survey is to establish an objective, detailed baseline of the current MDT operational process before the introduction of any new AI-driven technology. The term 'placebo' is used metaphorically; the survey is a diagnostic tool, rather than an active intervention, designed to map existing workflows, identify communication barriers, and pinpoint operational bottlenecks without the bias that might arise from simultaneously introducing a new system. This foundational understanding of current practices is crucial, as it provides a clear benchmark against which the true impact and effectiveness of subsequently implemented AI solutions can be accurately measured. In parallel, the United States-based xCures initiative[46] operates the xDECIDE platform, an AI-augmented clinical decision support system. This system uses a 'human—AI team' approach, combining NLP and machine learning with expert review from oncologists and molecular pharmacologists in a virtual tumor board setting. By structuring complex patient data and integrating real-world evidence, these transatlantic efforts are building the foundational, validated frameworks vital for unlocking AI's full potential in MDTs. These international initiatives are essential for paving the way for scalable advancements in global cancer care and building the confidence required for widespread adoption.

ACKNOWLEDGEMENTS

The work of the author A.T. is supported by the S-RACE (San Raffaele Ai CEnter).

FUNDING

This work was supported by the ERDF Interregional Innovation Investment Instrument funded by the European Union [grant number 101115081]. The work of the author A.T was supported by the Supported by the Italian Ministry of Research, Complementary Actions to the NRRP "D34health - Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care" Grant (# PNC0000001). The funding sources had no role for writing or decision to submit this manuscript.

DISCLOSURE

The authors have declared no conflicts of interest.

REFERENCES

1. Law NLW, Hong LW, Tan SSN, Foo CJ, Lee D, Voon PJ. Barriers and challenges of multidisciplinary teams in oncology management: a scoping review protocol. *BMJ Open*. 2024;14(2):e079559.
2. Appetecchia M, Amantea C, Canfora M, et al. Disease management teams in oncology: state of the art and the experience of a scientific institute of hospitalization and treatment (IRCCS) in Rome, Italy. *Brain Behav Immun Health*. 2024;43:100919.
3. Weissler EH, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*. 2021;22:537.
4. Ah-Thiane L, Heudel PE, Campone M, et al. Large language models as decision-making tools in oncology: comparing artificial intelligence suggestions and expert recommendations. *JCO Clin Cancer Inform*. 2025;9:e2400230.
5. Azenkot T, Rivera DR, Stewart MD, Patel SP. Artificial intelligence and machine learning innovations to improve design and representativeness in oncology clinical trials. *Am Soc Clin Oncol Educ Book*. 2025;45(3):e473590.
6. Hao Y, Qiu Z, Holmes J, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *NPJ Digit Med*. 2025;8(1):450.
7. Maddox TM, Embí P, Gerhart J, Goldsack J, Parikh RB, Sarich TC. Generative AI in medicine — evaluating progress and challenges. *N Engl J Med*. 2025;392:2479-2483.
8. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. *Commun Eng*. 2025;4:128.
9. Bosma JS, Dercksen K, Builtjes L, et al. The DRAGON benchmark for clinical NLP. *NPJ Digit Med*. 2025;8(1):289.
10. Garg M, Raza S, Rayana S, Liu X, Sohn S. The rise of small language models in healthcare: a comprehensive survey. *ArXiv*. 2025. abs/2504.17119.
11. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating large language models for health-related text classification tasks with public social media data. *J Am Med Inform Assoc*. 2024;31(10):2181-2189.
12. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Med Inform*. 2024;12:e55318.
13. Šuvalov H, Lepson M, Kukk V, et al. Using synthetic health care data to leverage large language models for named entity recognition: development and validation study. *J Med Internet Res*. 2025;27:e66279.
14. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: a transferable clinical natural language processing model for electronic health records. *Artif Intell Med*. 2021;118:102086.
15. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
16. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *J Am Med Inform Assoc*. 2022;29:1208-1216.
17. Mitchell JR, Szepietowski P, Howard R, et al. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT Network): development study. *J Med Internet Res*. 2022;24(3):e27210.
18. Moon S, He H, Jia H, Liu H, Fan JW. Extractive clinical question-answering with multianswer and multifocus questions: data set development and evaluation study. *JMIR AI*. 2023;2:e41818.
19. Ying H, Yu S. A Question answering based pipeline for comprehensive Chinese EHR information extraction. *arXiv*. 2024. arXiv:2402.11177.
20. da Silva DP, da Rosa Fröhlich W, de Mello BH, Vieira R, Rigo SJ. Exploring named entity recognition and relation extraction for ontology and medical records integration. *Inform Med Unlocked*. 2023;43:101381.
21. Knez T, Žitnik S. Multimodal learning for temporal relation extraction in clinical texts. *J Am Med Inform Assoc*. 2024;31(6):1380-1387.
22. Wei Q, Ji Z, Si Y, et al. Relation extraction from clinical narratives using pre-trained language models. *AMIA Annu Symp Proc*. 2020;2019:1236-1245.
23. Lilli L, Santoro M, Masiello V, et al. MISTIC: a novel approach for metastasis classification in Italian electronic health records using transformers. *BMC Med Inform Decis Mak*. 2025;25(1):160.
24. Wang Y, Wang Y, Peng Z, Zhang F, Zhou L, Yang F. Medical text classification based on the discriminative pre-training model and prompt-tuning. *Digit Health*. 2023;9:20552076231193213.

25. Alawad M, Gao S, Qiu JX, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *J Am Med Inform Assoc*. 2020;27(1):89-98.
26. Rios A, Durbin EB, Hands I, Kavuluru R. *Assigning ICD-O-3 codes to pathology reports using neural multi-task training with hierarchical regularization*., 32. ACM BCB; 2021.
27. Ioanovici A-C, Feier A-M, Măruşteri M-Ş, Trâmbiţaş-Miron A-D, Dobru D-E. Enhancing and not replacing clinical expertise: improving named-entity recognition in colonoscopy reports through mixed real—synthetic training sources. *J Pers Med*. 2025;15(8):334.
28. Wyles CC, Fu S, Odum SL, et al. External validation of natural language processing algorithms to extract common data elements in THA operative notes. *J Arthroplasty*. 2023;38(10):2081-2084.
29. Ghali M-K, Farrag A, Sakai H, Baz HE, Jin Y, Lam S. GAMedX: generative AI-based medical entity data extractor using large language models. *ArXiv*. 2024. abs/2405.20585.
30. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
31. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. 2024;31(9):1812-1820.
32. Dagli MM, Ghenbot Y, Ahmad HS, et al. Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review. *Sci Rep*. 2024;14(1):26783.
33. Dencker EE, Bonde A, Troelsen A, Sillesen M. Validation of natural language processing for surgical complication surveillance: detecting eleven postoperative complications from electronic health records. *medRxiv*. 2025.
34. Aali A, Van Veen D, Arefeen YI, et al. A dataset and benchmark for hospital course summarization with adapted large language models. *J Am Med Inform Assoc*. 2025;32(3):470-479.
35. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med*. 2025;8(1):274.
36. Gaber F, Shaik M, Allega F, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *NPJ Digit Med*. 2025;8(1):263.
37. Ferber D, El Nahhas OSM, Wölflein G, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat Cancer*. 2025;6(8):1337-1349.
38. Ferber D, El Nahhas OSM, Wölflein G, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat Cancer*. 2025;6(8):1337-1349.
39. Geukes Foppen RJ, Gioia V, Gupta S, Johnson CL, Giantsidis J, Papademetris M. Methodology for safe and secure AI in diabetes management. *J Diabetes Sci Technol*. 2025;19(3):620-627.
40. Berman E, Sundberg Malek H, Bitzer M, Malek N, Eickhoff C. Retrieval augmented therapy suggestion for molecular tumor boards: algorithmic development and validation study. *J Med Internet Res*. 2025;27:e64364.
41. Hamamoto R, Koyama T, Kouno N, et al. Introducing AI to the molecular tumor board: one direction toward the establishment of precision medicine using large-scale cancer clinical and biological information. *Exp Hematol Oncol*. 2022;11(1):82.
42. Lutz S, D'Angelo A, Hammerl S, et al. Unveiling the digital evolution of molecular tumor boards. *Target Oncol*. 2025;20(1):27-43.
43. Fu YV, Ramachandran GK, Park N, et al. BioMistral-NLU: towards more generalizable medical language understanding through instruction tuning. *AMIA Jt Summits Transl Sci Proc*. 2025;2025:149-158.
44. Lu W, Luu RK, Buehler MJ. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities. *NPJ Comput Mater*. 2025;11:84.
45. Mahon P, Chatzitheofilou I, Dekker A, et al. A federated learning system for precision oncology in Europe: DigiONE. *Nat Med*. 2024;30(2):334-337.
46. Shapiro MA, Stuhlmiller TJ, Wasserman A, et al. AI-augmented clinical decision support in a patient-centric precision oncology registry. *AI Precis Oncol*. 2024;1:58-68.