World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

Transformer-based architectures for domain-aware clinical text automation

FNU Sudhakar Abhijeet *

Northeastern University, Boston.

World Journal of Advanced Engineering Technology and Sciences, 2025, 16(02), 251-258

Publication history: Received on 05 July 2025; revised on 12 August 2025; accepted on 14 August 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.16.2.1280

Abstract

Clinical text processing automation is an essential solution in the development of a more efficient, accurate, and scalable healthcare system. The complexities posed by electronic health records, diagnostic reports, and unstructured clinical narratives that feature heavy terminology, erratic structures, and domain-specific semantics escaped traditional natural language processing methods because they were unable to deal with the complexity of the data. Transformer architecture has become a revolutionary solution by providing self-attention and contextual embedding structures representing long-range dependencies and fine-level word patterns. These models facilitate automated clinical documentation, coding, decision support, and multimodal data integration at higher accuracy and compliance by considering methods that are domain-aware, like biomedical pretraining, ontology integration, federated learning, and privacy-preserving training. In this paper, we will review the history of transformer models in the context of clinical NLP, the domain adaptation methods they use, how to achieve scalability and observability, and what the potential future research opportunities are, such as benchmarking multi-region failover, cost-aware autoscaling of health infrastructure using artificial intelligence.

Keywords: Transformers; Clinical Nlp; Domain Adaptation; Clinical Automation; Healthcare AI

1. Introduction

One of the most difficult and important types of unstructured medical data is clinical text data, which is based on electronic health records (EHRs), diagnostic reports, treatment notes, and patient history. In contrast to general texts, the clinical narratives are very specific as they tend to contain highly specialized vocabulary, abbreviations, the occurrence of irregular sentence structure, and a specialized semantics that makes the process of their processing and automation highly complex in nature. Such complexity has proven problematic for the traditional natural language processing (NLP)pipelines, which have largely used rule-based systems and statistical analyses, which cannot generalize well across different clinical areas [1-3]. The advent of deep learning, more specifically transformer-based models, has transformed the landscape of clinical text automation with models able to learn contextual semantics, long-range dependence, and task particularities with little or no manual control [4-6].

Automated coding of medical records, clinical decision support, and other applications have been redefining the processing, analysis, and incorporation of unstructured healthcare text into digital health infrastructures using these architectures [7-9]. The last few years have seen a rise in transformers becoming the fundamental part of cutting-edge clinical NLP models. Their self-attention mechanisms enable dynamic weights on tokens, which enables them to process complex sentence structures better, as compared to recurrent or convolutional models [10-12]. Like the previously presented examples, the transformer-based models exhibit substantial gains in entity recognition, relation extraction, summarization, and report generation tasks when coupled with the domain-specific adaptations, such as being pretrained on biomedical corpora, the inclusion of ontology-based embeddings, or few-shot learning [13-15]. This has been expedited by the increased availability of clinical datasets and fast development of scalable pretraining and

[*] Corresponding author: FNU Sudhakar Abhijeet.

transfer learning, which help to close the gap between generic language models and domain-savvy LLMs that are domain-optimized to solve problems in healthcare [16-18].

In this article, the landscape of transformer-based architectures in clinical text automation is discussed, including its design, domain adaptation strategies, major uses, and further development. They form a progressive sequence with each section progressing knowledge obtained in the previous section, and commence with overarching principles and move into the details of the used tasks, integration paradigms, and issues. As a starting point, the discussion introduces the fundamentals in terms of how transformer mechanisms have disrupted language modeling, why clinical text presents specific challenges to these systems, and how they adapt to specific domains, are deployed operationally into healthcare processes, and how innovations on this front are changing.

2. Evolution of Transformers and the Clinical NLP Paradigm Shift

Transformers became a decisive change in NLP when the previous types of NLP models were replaced with recurrent and convolutional models. The computational essence of transformers is their multi-head self-attention mechanism that does away with the sequential bottlenecks of recurrent architectures and allows parallelization, a computational process which speeds up training and inference multiple times, as on Figure 1 [19-21]. Transformers show great ability to model contextual relationships in the long and frequently irregular clinical papers, where the most important diagnostic codewords may lie very far apart in a sentence. This ability is especially instrumental in medicine stories when the background of the events is dispersed among the sporadic observations, laboratory reports, and histories of care [22, 23]. Such architectures not only have demonstrated positive gains in referencing tasks in the healthcare field but have also been used to perform cross-task transfer learning, where large language models can be such activities fine-tuned to various clinical operations due to little labeled data [24, 25]. Transformers in healthcare rely on domain-aware adaptation, though, to achieve high efficacy. Clinical language varies significantly with regard to the generic language in terms of both structure and semantics, e.g., abbreviations, like HTN (hypertension) or SOB (shortness of breath), denote very specialized meanings, nested entities, such as drug-dosage relations, need a detailed contextual interpretation [26-28]. When not adapted to a domain, such nuance is usually misinterpreted or missed by generic models, causing them to be less useful in practice, in a clinical setting [29-30].

As an extension of the above transformative abilities, the second section discusses the applicability of domain-sensitive schemes, especially biomedical pretraining, ontology-gated embeddings, and hybrid designs, to boost transformer models into the clinical domain, connecting the underlying principles to more specific implementations.

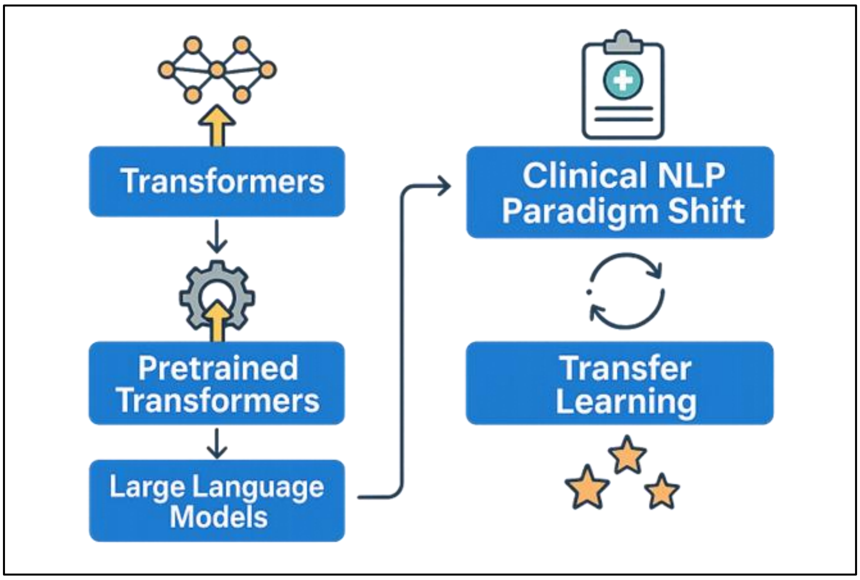

Figure 1 Evolution of Transformers and the Clinical NLP Paradigm Shift showcasing the progression from basic transformer models to large language models, alongside the adoption of transfer learning and few-shot learning in clinical natural language processing

3. Domain-Aware Adaptation of Transformer Architectures

The reason that transformer models succeed in automating clinics is the ability to specialize the model to the complexities of healthcare text. Models trained on large biomedical and clinical datasets, including EHR notes, PubMed abstracts, and patient records that had been de-identified, can absorb medical jargon, medical syntax, and word co-occurrences [4,11,14]. This pretraining in the domain, with the help of masked language modeling tasks, leads to a large downstream improvement in the entity recognition, relation extraction, and classification tasks in comparison with models trained on general corpora only [17,19]. Moreover, the introduction of ontology-based inferences, e.g., embeddings of well-defined terminologies (e.g., SNOMED CT or ICD), will offer such models a structured semantic comprehension and minimize the chances of error during the interpretation of rare or ambiguous word occurrences [13, 20].

In addition to the concept of pretraining, hybrid methods are frequently used in domain adaptation, in which transformers are adapted to an external clinical knowledge base. Such integrations aid models to better clear context-sensitive confusions, like differentiating between homonyms (between a chemical component called leads and a technical component on a medical device called leads) or comprehending encapsulated concepts (as in the case of medications and dosage) [9,15]. Moreover, the transformer model can be generalized in uncommon clinical situations (i.e., rare diagnoses or uncommon drug interactions in rare cases) because the learned representations can be used to solve related tasks via few-shot or zero-shot learning [18, 22]. Although these approaches provide an effective boost of domain knowledge, they make the training computationally challenging and introduce regulatory issues when pretraining on sensitive patient data. This has raised concerns that require privacy-preserving fitting techniques like differential privacy and federated learning to be fit to satisfy regulations in healthcare [6,16]. These intricacies usher us to the next part on how we operationalize such adapted transformers to automate some of the key clinical text functions within health care systems.

To better illustrate how domain-aware enhancements impact model performance across clinical tasks, the following table compares representative transformer configurations and their average performance improvements over general-domain baselines.

Table 1 Comparative Impact of Domain-Specific Enhancements on Transformer Models for Clinical NLP Tasks

Domain Adaptation Strategy	Example Clinical Task	Average Accuracy Gain vs. General Baseline (%)	Typical Data Scale Used for Pretraining (Tokens)
Biomedical Pretraining (EHR + PubMed)	Clinical Entity Recognition	+18%	2–5 billion
Ontology-Integrated Embeddings (SNOMED)	Relation Extraction (Drug–Disease Links)	+22%	500 million – 1 billion
Federated Pretraining (Cross-Hospital)	Rare Disease Classification	+15%	1–3 billion
Few-Shot Transfer Learning	Rare Event Summarization	+10%	<100 million (labeled)

The observed performance gains underscore the importance of tailoring transformers to clinical text through domain-specific pretraining and knowledge integration. These adapted models form the foundation for scalable automation frameworks discussed in the subsequent sections, where computational efficiency and clinical reliability are equally prioritized.

4. Automating Clinical Documentation and Coding with Transformers

Codifying clinical data is among the most time-consuming activities in contemporary healthcare practice, and clinicians spend considerable parts of their day typing in patient histories, proceedings, and diagnostic results, as revealed in Figure 2. Transformer systems mitigate this work by automating organizational tasks, summarizing processes, and coding. These architectures have the potential to reduce entry-level EHRs that are multi-page long into abstract summaries with adequate contextual accuracy to be used as decision support systems or patient discharge matrimony [7, 10, 14] when fine-tuned to clinical summarization. In an analogous manner, transformers can be trained to perform automatic coding tasks, converting free-text notes to standardized billing categories, like information content with ICD or CPT codes, which saves administrative redundancy and eliminates coding mistakes to a considerable extent [12,21]. In contrast to the previous rule-based or statistical applications, transformer-powered automation is capable of adjusting to the changes in writing style, regional terminology, and other changing clinical guidelines dynamically. These models incorporate attention, which allows them to pick, or focus on, clinically relevant phrases like temporal markers (postoperative day three), or critical values (elevated troponin), enhancing both accuracy and interpretability [8,19]. Additionally, they can be introduced into the hospital information system, and due to their resiliency, are able to handle millions of documents at a time with the same performance level, which will enable large analytics and other research projects.

This transformative impact on documentation lays the foundation for more advanced applications, such as real-time decision support and automated clinical report generation, which require even deeper contextual understanding of topics addressed in the following section.

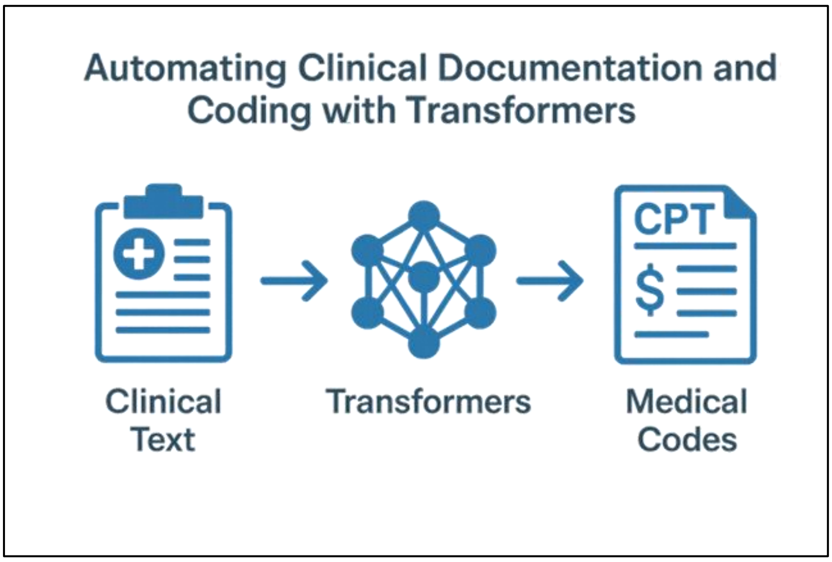
Automating Clinical Documentation and Coding with Transformers
CPT
$
Clinical Text
Transformers
Medical Codes

Figure 2 Automating clinical documentation and coding using transformer models, illustrating the flow from raw clinical text to structured medical codes

5. Real-Time Clinical Decision Support and Report Generation

Leaving documentation aside, transformer-based solutions are also finding their way into clinical decision support (CDS) frameworks, where they distill patient-specific information with a set of evidence-based guidelines to make guidance that can be acted upon. Transformer multimodal models can provide potential diagnoses, suggestions on what to do, or warnings of adverse drug interactions in nearly real time through the power of contextual embeddings derived using multimodal sources of information such as text, structured lab data, and imaging reports [13,17,22]. Their capacities to represent long-range dependencies enable them to accommodate longitudinal patient histories, a decisive feature when evaluating chronic or multi-morbidity patients. In radiology and pathology, the transformers run automated report generation in which imaging results are summarised and the results are matched with historical patient data. They lower the latency of reporting, minimize terminological inconsistencies, and minimize variation introduced by humans, leading to a better continuity of care and diagnostic correctness [16,18]. Automation of tasks as complex and high-stakes as these requires extremely high validation, including the use of clinician-in-the-loop review processes that increase clinical safety and compliance within a regulatory framework. Decision support and report automation are finding more traction, and the role of strong observability, interpretability, and compliance auditing is emerging as a priority, an area that will be discussed in the following section, where monitoring and enforcement of policies are discussed relative to clinical-grade deployments.

6. Observability and Compliance in Transformer-Driven Clinical Systems

With transformer-based clinical text automation shifting into production, observability and compliance have become fundamentally critical because they guarantee the reliability of operations and conformance to regulations. In the healthcare industry, every given workflow is so critical that models used during clinical documentation or coding, or decision support procedures, should not only work correctly but also be transparent and auditable. Observability starts with real-time monitoring of the system metrics, i.e., model drift, latency, throughput, and accuracy degradation. Coupled sets of Prometheus, Grafana, and ELK pipelines collect operating and clinical performance metrics and enable system administrators to identify abnormalities threatening the care delivery of patients [19,20]. In addition to the operational metrics, the clinical systems should provide policy enforcement as well as a security baseline. Another configurable criterion is dedicated to automated compliance scanners that constantly verify containerized deployment to the healthcare regulatory baselines, including the ones in line with the HIPAA or ISO 27001 framework [21]. It is important to detect policy drift, in which the absence of configurations should be treated as a potential vulnerability or misconfiguration case. An example is of using kube-bench in the runtime audits to identify deviations of a baseline in the Session Management Function (SMF) container runtime, marking the unauthorized capabilities bestowed to pods, and responding in order to remediate and rollback [22]. These kinds of proactive enforcement safeguard against configuration-based attacks that may result in corruption of patient information.

Such a mechanism of observability also underlies explainability efforts. Decision-trace logs and attention-weight visualizations allow seeing inside the model to understand its way of working and incorporate this information to be comfortable with automated results and fulfill the reporting requirements required by regulators. observability and compliance could be synthesized to support operational trust and meet the two imperatives of performance and accountability by means of transformer-driven systems. This focus on system resilience leads instinctively to the topic of scalability and multi-institutional integration that follows, wherein the complexity of the infrastructure to support it is magnified due to its adoption.

7. Scalability and Multi-Institutional Deployment

The installation of transformer-based architectures in hospital systems, research institutes, and telehealth systems requires infrastructure able to dynamically scale to variations in the workload. The nature of the workload of clinical NLP is naturally spiky, and peak demand may coincide with a reporting deadline or an admission period, or a batch process of the previous EHR data. Scalability plans focus on the elastic build patterns via Kubernetes-driven sets in which auto-scaling policies tame and share usage of compute assets and memory [23, 24]. Horizontal scaling allows scaling to vast corpora to allow parallel processing, and vertical scaling to ensure transformer layers, and in particular large pretrained variants, are sufficiently resourced to prevent inference. Cross-institutional deployments present an unorthodox set of issues, such as security of data exchange, latency-sensitive inferencing, and heterogeneity in EHR systems. Privacy risks are addressed through federated learning frameworks, in which training is collaborative across decentralized data sources, without the centralization of sensitive patient data, and such federated settings enhance the generalizability of models [25,26]. Domain-specific transformers combined with privacy-preserving protocols, such protocols like secure multi-party computation and differential privacy, are also a common feature of these types of architectures to ensure that regulatory compliance is not lost through scaling.

The scalability challenge is also applicable to the edge deployments, where the lightweight versions of the transformer are deployed nearer the data source, e.g., in the telemedicine equipment or a diagnostic kiosk. The localized models can handle real-time text inputs, such as patient triage questionnaires, and delegate their complex tasks of summary or code generation to centrally located clusters. Such distributed deployments necessitate orchestration structures that can adjust the workloads among edge and core dynamically depending on the condition of the network and availability of compute [27]. As scalability enables a large user base, it increases the risks in the operation, including model drift and inefficiencies in costs. Such pitfalls emphasize the need to provide smart optimization mechanisms, and in the second part of the paper, the integration of adaptive and cost-aware automation schema and the role emerging in scaling heuristics through AI is explored.

8. Adaptive Optimization and Cost-Aware Automation

Transformer-based clinical NLP systems have significant computational and financial overhead even in their current scaled operations, especially because models with hundreds of millions of parameters may be used for many other tasks, such as summarization or entity recognition. These challenges are overcome with the help of cost-aware automation strategies, which dynamically adjust the provisioned infrastructure to the requirements, prioritize, and optimize the way resources are utilized without compromising performance. Elastic clusters can dynamically right-size using auto-scaling policies using contextually based workload measurements, including request latency or request document length, or request criticality [28].

In recent technology, the use of reinforcement learning and meta-learning algorithms to make autoscaling decisions is relatively new. These systems dynamically learn about the past workloads to make predictions of configuring resources optimally in terms of latency points relative to the costs of operations. As an example, transformers that are utilized in common documentation processes can enforce Spot Instances that compromise moderate availability rates, whereas models engaged in decision support that is accomplished in real-time run on scheduled, high-performance nodes, with a view to holding response times [29]. There is also the capability to compress large models through adaptive pruning and quantization methods to optimize latency and energy requirements, especially at the edge. These approaches will guarantee that even architecturally costly solutions can be used where price is a primary concern, and smaller healthcare organizations. Notably, they have to be proven viable in the regulatory setting because the efficiency advantage cannot negatively affect the accuracy and replicability of the models, which is a factor that relates directly to the following section about the challenges and research frontiers.

In parallel with algorithmic improvements, resource allocation strategies critically affect the sustainability of clinical NLP deployments. The table below summarizes efficiency-oriented configurations that balance performance with operational costs in large-scale healthcare infrastructures.

Table 2 Cost-Aware Resource Strategies for Transformer-Based Clinical NLP Deployments

Infrastructure Strategy	Typical Use Case	Cost Reduction Achieved (%)	Trade-offs and Considerations
Spot Instance Pools with Preemptible Nodes	Routine Documentation Automation	50–70%	Requires fault-tolerant workloads and checkpointing.
Reinforcement Learning–Driven Autoscaling	Real-Time Decision Support Systems	30–45%	Complexity in tuning; must avoid latency spikes.
Quantized Transformer Inference (8-bit)	Edge Deployments (e.g., Telehealth Triage)	40–60%	Slight drop in precision; mitigated by retraining.
Tiered Storage for Intermediate Model States	Large Batch EHR Coding and Summarization	25–35%	Retrieval latency may increase for cold tiers.

9. Challenges in Clinical Deployment of Transformer Architectures

Possessing a powerful transformative capability, when applied in healthcare instances, the deployment of demanding transformer-based systems is accompanied by a certain range of technical, ethical, and operational barriers. Supervised training, benchmarking may be constrained due to the limited amount of annotated clinical data, due to privacy. Although these limitations can be alleviated by using unsupervised pretraining on de-identified corpora, fine-tuning to more specialized tasks tends to be hurt due to an insufficient number of domain-relevant examples [10,12]. Partial solutions can be regarded as transfer learning and few-shot adaptation, which are not as effective in dealing with subdomains of clinical issues, specifying rare diseases and low-resource languages [20, 21].

Interpretability and trust are the other issues. The model decisions that clinicians need to be able to understand are not only what clinicians need to be able to trust is correct, but also need to understand the reasoning behind what they are doing. Although attention maps and counterfactual reasoning tools are effective approaches in enhancing transparency, they are lacking in communicating all the decision logic to the non-technical stakeholders [17, 22]. In practice, changing large transformers can pose a computational burden to the IT infrastructures in a hospital, especially in parts with low budgets or fewer amenities. Sustainability issues are also associated with high energy consumption, and work is being done on models with efficient architecture, including sparse transformers and knowledge-distilled versions [25, 27]. Also, the regulatory framework is not up to date with technological progress, which brings ambiguity to the approval process, auditability requirements, and liability risk to the AI-based clinical decisions [24].

The variety of challenges presented herein emphasizes the necessity of further research and innovative area development that preconditions further section development of emergent trends and research opportunities in transformer-based clinical text automation.

10. Future Directions and Research Opportunities

The future path of transformer-based clinical NLP is yet to be sped up once the transformers themselves and the models used are better adapted, and the deployment frameworks within this domain improve. The work done in the future is likely to focus on further multimodal integration, with transformers combining text, imaging, genomic, and wearable sensor data to produce complete representations of a patient. This type of cross-sphere integration might drive next-generation decision support with the capability of multidimensional, individualized suggestions on unheard-of scales. The other important area in research is AI-based cost-sensitive autoscaling, in which pay-as-you-go-style autoscaling is determined by predictive models of workload and criticality. With the input of historical demand patterns, patient acuity scores, and energy price signals, these systems would be able to improve the cost structure, yet continue to meet clinical-grade levels of performance. There will also be the study of resiliencies, especially multi-regional failover benchmarking. As more healthcare provision is dependent on cloud-native capacity, more systematic research is required to estimate failover delays, consistency assurances, and the effects of service degradation when regions fail. These benchmarks will guide the ideal practices of disaster recovery planning in a mission-critical clinical setting. Also, there will be a need to adopt innovations in the area of privacy-preserving training, including new advancements in federated learning or homomorphic encryption, to be able to scale models to the conditions of various healthcare systems so that their sensitive patient data will be protected. The combination of these lines of research holds the potential of increasing the scope and confidence of automation powered by transformers, bolstering its reputation as a pillar of digital care change.

11. Conclusion

Transformer-based architectures have emerged as a transformative force in automating the processing of complex, unstructured clinical text. By leveraging self-attention mechanisms, contextual embeddings, and domain-aware adaptations such as biomedical pretraining, ontology integration, and privacy-preserving learning, these models address many of the limitations of traditional NLP methods in healthcare. Their integration into documentation, coding, decision support, and report generation workflows demonstrates tangible improvements in efficiency, accuracy, and scalability, while also enabling advanced capabilities such as multimodal analysis and real-time clinical guidance. However, the deployment of such systems in healthcare brings critical challenges, including the need for interpretability, regulatory compliance, cost-efficient scaling, and resilience across multi-institutional environments. Addressing these requires robust observability frameworks, adaptive optimization strategies, and continued innovation in privacy-preserving and energy-efficient model design. Looking forward, the convergence of transformers with multimodal patient data, intelligent autoscaling, and failover benchmarking presents an opportunity to elevate both the scope and trustworthiness of AI-driven healthcare systems. By balancing computational performance with ethical, legal, and operational imperatives, transformer-powered clinical NLP can mature into a cornerstone of digital healthcare transformation, enabling smarter, faster, and more personalized care delivery on a global scale.

References

[1] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

[2] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23.

[3] Turchin A, Masharsky S, Zitnik M. Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked. 2023 Jan 1;36:101139.

[4] Deznabi, I., Iyyer, M., & Fiterau, M. (2021, August). Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 4026-4031).

[5] Fang, L., Chen, Q., Wei, C. H., Lu, Z., & Wang, K. (2023). Bioformer: an efficient transformer language model for biomedical text mining. *ArXiv*, arXiv-2302.

[6] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-46.

[7] Sabeenian RS, Vinodhini CM. A systematic review on machine learning/deep learning model-based detection of sleep apnea using bio-signals. Recent Pat Eng. 2025;19(4):E230724232158.

[8] Liu L, Perez-Concha O, Nguyen A, Bennett V, Jorm L. Automated ICD coding using extreme multi-label long text transformer-based models. Artif Intell Med. 2023;144:102662.

[9] Li D, Zhang Y, Zhang Y, Lu H, Yu B, Peng S, et al. A hybrid model based on deep convolutional neural network for medical named entity recognition. In: 2022 IEEE 24th International Conference on High Performance Computing & Communications; 8th International Conference on Data Science & Systems; 20th International Conference on Smart City; 8th International Conference on Dependability in Sensor, Cloud & Big Data Systems & Applications (HPCC/DSS/SmartCity/DependSys). 2022 Dec. p. 2353-7.

[10] Guo B, Liu H, Niu L. Integration of natural and deep artificial cognitive models in medical images: BERT-based NER and relation extraction for electronic medical records. Front Neurosci. 2023;17:1266771.

[11] Sarkar AR, Chuang YS, Mohammed N, Jiang X. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. Sci Rep. 2024;14(1):29669.

[12] Mai G, Cundy C, Choi K, Hu Y, Lao N, Ermon S. Towards a foundation model for geospatial artificial intelligence (vision paper). In: Proceedings of the 30th International Conference on Advances in Geographic Information Systems. 2022 Nov. p. 1-4.

[13] Feng J, Shaib C, Rudzicz F. Explainable clinical decision support from text. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020 Nov. p. 1478-89.

[14] Keszthelyi D, Gaudet-Blavignac C, Bjelogrlic M, Lovis C. Patient information summarization in clinical settings: scoping review. JMIR Med Inform. 2023;11(1):e44639.

[15] Skreta M, Arbabi A, Wang J, Drysdale E, Kelly J, Singh D, et al. Automatically disambiguating medical acronyms with ontology-aware deep learning. Nat Commun. 2021;12(1):5319.

[16] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3(1):119.

[17] Xu Q, Xie W, Liao B, Hu C, Qin L, Yang Z, et al. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: a systematic review. J Healthc Eng. 2023;2023(1):9919269.

[18] Kusal S, Patil S, Choudrie J, Kotecha K, Mishra S, Abraham A. AI-based conversational agents: a scoping review from technologies to future directions. IEEE Access. 2022;10:92337-56.

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.

[20] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964. 2020.

[21] Kargarandehkordi A, Li S, Lin K, Phillips KT, Benzo RM, Washington P. Fusing wearable biosensors with artificial intelligence for mental health monitoring: a systematic review. Biosensors. 2025;15(4):202.

[22] Oladoja T. Comprehensive security for Kubernetes in healthcare: modern challenges and solutions. 2022.

[23] Falkenberg R. 9.2 Power consumption analysis and uplink transmission power. Also of interest. 2022;423.

[24] Casper S, Ezell C, Siegmann C, Kolt N, Curtis TL, Bucknall B, et al. Black-box access is insufficient for rigorous AI audits. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024 Jun. p. 2254-72.

[25] Cho HN, Jun TJ, Kim YH, Kang H, Ahn I, Gwon H, et al. Task-specific transformer-based language models in health care: scoping review. JMIR Med Inform. 2024;12:e49724.

[26] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598.

[27] Liu V, Yin Y. Green AI: exploring carbon footprints, mitigation strategies, and trade-offs in large language model training. Discov Artif Intell. 2024;4(1):49.

[28] Simon BD, Ozyoruk KB, Gelikman DG, Harmon SA, Türkbey B. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. Diagn Interv Radiol. 2025;31(4):303.

[29] Xu S, Manshaii F, Xiao X, Chen J. Artificial intelligence-assisted nanogenerator applications. J Mater Chem A. 2025;13(2):832-54.

[30] Dasari KK. Cross-cloud continuity: a scalable framework for resilient and regulated digital infrastructure. 2023.