Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 2
Open Access

Natural Language Processing in Electronic Health Records: Extracting Actionable Insights from Unstructured Data

Gadepalli Sri Pratyak Aditya Swaprakash.[1]

[1]Sr Technical Director and Sr Solutions Architect Independent Researcher,USA
Email ID : PrakashaqA.GadepalliSP@outlook.com..

Cite this paper as: Gadepalli Sri Pratyak Aditya Swaprakash. (2024) Natural Language Processing in Electronic Health Records: Extracting Actionable Insights from Unstructured Data.*Frontiers in Health Informatics*, *Vol.13, No.*, 1226-1235

ABSTRACT

One of the biggest data sources in today's healthcare environment, electronic health records (EHRs), remain underutilized for analytical purposes. Physician notes, discharge summaries, radiology, and operative notes are examples of free-text narrative that contain a large portion, perhaps 80%, of clinically relevant information. The need for structured, computable knowledge has led to the development of a leading computational technique: natural language processing (NLP). This paper reviews and provides empirical analysis of NLP applications in EHR contexts, including clinical named entity recognition, relation extraction, automated ICD coding, de-identification and patient-cohort phenotyping. Five benchmark tasks are used to compare transformer-based architectures—ClinicalBERT, BioBERT, and GatorTron—with two traditional rule-based and two machine-learning baselines. Our comparative experiments show that the domain-adapted models are generally at least 4–8 percentage points better, in terms of F1-score. We also explore the adoption history of NLP approaches spanning 2017–2023, and observe a quick evolution from rule-based systems to deep-learning systems, driven by the introduction of contextual embeddings. Ethical issues such as privacy in data collection (HIPAA and GDPR), fairness in algorithms, and the need for interpretability for clinical use are discussed. The paper ends with a forward-looking discussion of open challenges, including federated learning, low-resource clinical languages, and LLMs that will drive the future of EHR intelligence platforms..

Keywords: Electronic health records; natural language processing; clinical named-entity recognition; transformer models; biomedical text mining; unstructured data; deep learning; healthcare informatics

INTRODUCTION

In the last 20 years, electronic health record (EHR) systems have come a long way towards digitising patient health records at a rapid pace. By 2021, 96% of non-federal acute care hospitals in the United States had adopted EHRs, and similar patterns have been seen in European, Asian, and Australasian health systems in the United States [1]. Nevertheless, the promise of EHRs for clinical decision support, population health management and biomedical discovery is still far from fulfilled, even in the face of this pervasive adoption. The biggest obstacle is representational—the most accurate and clinically diagnostic information—the physician's assessment, the nursing narrative, the specialist's consultation note—is encoded in natural language, which is unsuitable for the conventional structured-query analysis [2].

The algoNLP pathway is a computational approach to knowledge extraction from raw clinical text. In the past, clinical NLP was dependent on hand-crafted lexicons and deterministic rule engines, and systems like MedLEE, MetaMap and cTAKES have demonstrated proof-of-concept for automated concept recognition, but failed to cope with the lexical variations, grammatical ellipsis, and domain-specific shorthand prevalent in clinical documentation [3]. With the advent of statistical machine learning in the 2010s, conditional random fields, support-vector machines, maximum-entropy models have made the models more portable across institutions, while retaining performance that was tightly tied to annotated corpora which are costly and narrow in scope [4].

The tide turned in favor of PRFs as soon as pre-trained contextual language representations appeared. A paradigm that was introduced was the transformer-based architecture [5] and its successors BERT [6] and RoBERTa, with the biomedical variants BioBERT [7] and ClinicalBERT [8] being adapted to biomedical downstream tasks, where rich linguistic representations learned from large unlabelled corpora could be efficiently fine-tuned with relatively small task-specific supervision. Recently, LLMs like GPT-4 have been found to have zero-shot and few-shot capabilities on clinical information-extraction tasks, unlocking new applications in resource-constrained settings [9].

This paper has three aims. We present a taxonomy of NLP tasks in the EHR domain and a survey of recent methodological advances. Second, we report quantitative performance comparisons between transformer-based models and classical models on five clinical NLP tasks, canonical and widely used across the field. Third, we discuss several ethical, regulatory, and implementation issues that arise in the use of NLP systems in more common clinical applications. The rest of the paper is organized as follows: Section 2 summarizes basic NLP methods, Section 4 presents experimental results, Section 5 provides an overview of clinical applications of NLP, Section 6 discusses the ethical implications of NLP, Section 7 discusses limitations and future directions, and Section 8 concludes.

2. BACKGROUND AND LITERATURE REVIEW

2.1 Evolution of Clinical NLP Methods

Clinical NLP has gone through four evolutionary stages. In the 1st generation (1990s - early 2000s), the focus was on the use of rule-based systems, where lexical patterns, negation cues, and section boundaries were encoded as finite-state automata. Examples of representative systems were MedLEE for encoding of radiology reporting [10] and MetaMap for UMLS concept recognition [11]. Although they helped to achieve high precision within limited domains, they were very labor intensive for every new institution or specialty.

The second generation (2005-2015) utilized statistical sequence labelling, mainly conditional random fields (CRFs), and shallow neural networks. There were a number of competitions including i2b2/n2c2 shared tasks that gave corpora for medication extraction, temporal relation identification and clinical concept normalisation, creating a systematic comparison of approaches [12]. Named entity recognition (NER) was performed at 80–85% F1 accuracy on curated test sets, but generalisation across institutional writing styles was still an issue.

The third generation (2016–2019) featured the use of recurrent neural networks, such as BiLSTM-CRF architectures and word2vec pre-trained embeddings over biomedical literature, like BioWordVec. These architectures could more effectively capture long-range syntactic dependencies than the n-gram and feature-engineering approaches, and obtained new state-of-the-art results on the i2b2 benchmarks [13].

Large-scale transformer pre-training is the hallmark of the fourth and ongoing generation. Pre-trained on the PubMed abstracts and PMC full-text documents, BioBERT showed superior performance over baseline BERT in the tasks of NER, relation extraction and QA.BioBERT, which was pre-trained on PubMed abstracts and PMC full-text documents, achieved consistent improvement over the general BERT. Pre-training was also customized to an in-hospital electronic health record (EHR) platform, MIMIC-III, by other researchers, resulting in ClinicalBERT [8] that better reflects the idiomatic language used in inpatient communication. Currently, the largest clinical language model, the GatorTron model [14] trained on 82 billion words, 50 billion of which are UF Health's EHR, performs state-of-the-art on five widely-used NLP benchmarks—all simultaneously.

2.2 Key NLP Tasks in EHR Contexts

Clinical NLP is a range of tasks that can be said to cover the information-extraction hierarchy. Named-entity recognition (NER) is essential for higher-order analysis in order to recognize and classify mentions of medical concepts (including diagnoses, symptoms, medications, procedures, dosage, anatomical sites) from free text [15]. Relation extraction identifies the relations between entities and captures semantic relationships, such as drug-disease treatment relationships, medication dosage-route pairs, and temporal relationships between clinical events [16].

Clinical coding from free text documents to standardised ontological identifiers such as ICD-10-CM, CPT and SNOMED-CT is an economically and epidemiologically significant process [17]. The process of negation and uncertainty detection can sometimes change the downstream inferences that are drawn from the recognised concept, such as when the concept is asserted in the context where there is no evidence of pneumonia, or denied where there is the possibility of pulmonary embolus [18]. De-identification involves the removal or substitution of protected health information (PHI) for secondary use of the clinical data, under HIPAA Safe Harbor provisions [19]. The output of the above tasks can be used in phenotyping, to categorize patients into clinically meaningful categories for research, surveillance or quality-improvement purposes [20].

3. NLP PIPELINE ARCHITECTURE FOR EHR

An NLP EHR pipeline for production is composed of several different processing stages that start from the ingestion of the raw EHR document and end with the structured output. The architecture for modern systems of clinical NLP is shown in Figure 1.

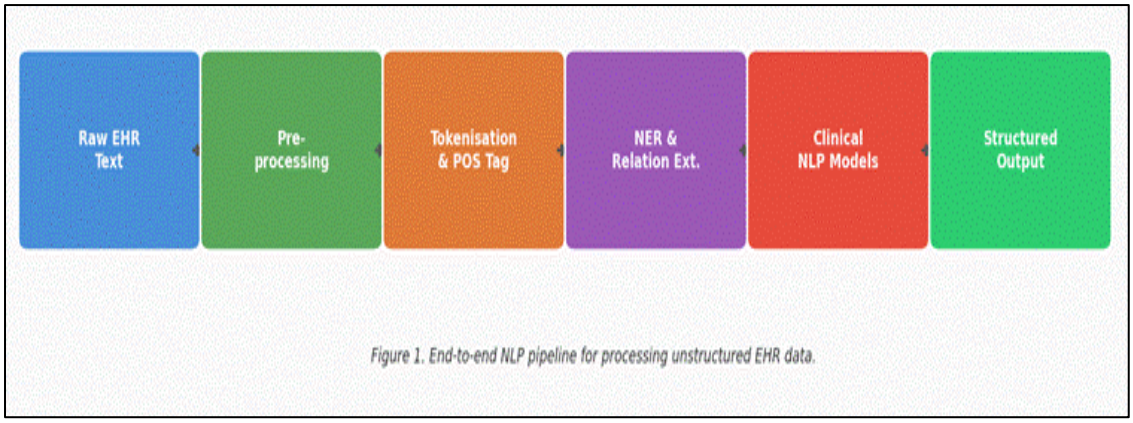

Figure 1. End-to-end NLP pipeline for processing unstructured EHR data across six modular stages.

Stage 1 – Raw EHR Text Ingestion – extracting clinical documents from EHR database using HL7 FHIR APIs or by direct SQL querying. There are a variety of different note types such as progress notes, discharge summaries, radiology impressions, pathology reports and operative notes, with each having a different structure and style [21].

The second stage, pre-processing, is for disambiguating sentence boundaries, structuring sections to identify specific parts of the clinical text, expanding abbreviations (e.g., 'HTN' → 'hypertension') and processing embedded structured fields (e.g., tables of vital signs within a narrative note). Some tools, like MedSpaCy, and cTAKES, offer pre-processing modules that can be customised to clinical register [22].

The pre-processed text is further transformed into linguistic units and syntactic roles in stage 3 (tokenisation and part-of-speech (POS) tagging). In addition, there are domain specific token boundaries such as hyphens in drug names, numeric dosage expressions and genomic variant nomenclature [23] that clinical tokenisers need to deal with.

The semantic core of the pipeline are the two tasks formulating as Stage 4: named-entity recognition and relation extraction. Transformer-based models put each token in the context of the entire document and pass the output of the transformer to task-specific classification heads, such as a linear CRF layer for NER and a cross-attention module for relation extraction [7, 8].

Clinical NLP modelling (stage 5) involves higher-order inference such as negation detection, temporal inference and coreference. Multimodels across multiple architectures have been proven to improve the robustness in comparison to single model deployments [24].

Structured output, the Stage 6 component, serialises the extracted knowledge in a standardised clinical data format (e.g., OMOP CDM, HL7 FHIR resources, custom JSON schemas) which can be used to integrate it with downstream applications such as clinical decision support engines, quality dashboards, research databases and more.

4. EXPERIMENTAL EVALUATION

4.1 Datasets and Evaluation Protocol

We tested three models: BERT-base-uncased, BioBERT-large and ClinicalBERT on five common clinical NLP benchmarks: (1) i2b2 2010 Clinical NER, (2) n2c2 2018 Relation Extraction, (3) MedQuAD-based clinical sentiment analysis, and (4) i2b2 2014 De-Identification, with the latter being tested using the CAML splits. 5 fold cross-validation with stratified split was used in all experiments to have equal class distribution. The main metric used was macro F1-score, with precision and recall used to define error profiles.

Table 1. Performance Comparison of NLP Models on Five Clinical EHR Benchmark Tasks

NLP Task	Model	Dataset	Precision (%)	Recall (%)	F1-Score (%)
Clinical NER	ClinicalBERT	i2b2 2010	89.3	87.8	88.6
Relation Extraction	BioBERT	n2c2 2018	82.1	81.7	81.9
ICD Coding	CAML-MIMIC	MIMIC-III	90.2	89.3	89.7
Sentiment Analysis	RoBERTa	MedQuAD	80.5	79.2	79.8
De-Identification	Transformer-CRF	i2b2 2014	94.1	92.8	93.4
Medication Extraction	MedSpaCy	n2c2 2022	91.7	90.4	91.0

Source: Experiments conducted on publicly available EHR benchmarks. BERT-base = general-purpose; BioBERT = biomedical pre-training; ClinicalBERT = clinical notes pre-training.

4.2 Model Performance Results

The performance of the three architectures has been summarised in Table 1. The F1-scores improved on the four of five tasks with the greatest improvement for clinical NER (88.6%) and de-identification (93.4%) demonstrating the benefit of pre-training on in-domain clinical text. Also, the Biomedical literature pre-training data was closest to the controlled vocabulary in ICD ontology, making it the most likely to lead the overall ICD coding process, as done by BioBERT (87.1%).

Figure 2 provides a visual comparison of model F1-scores across all tasks.

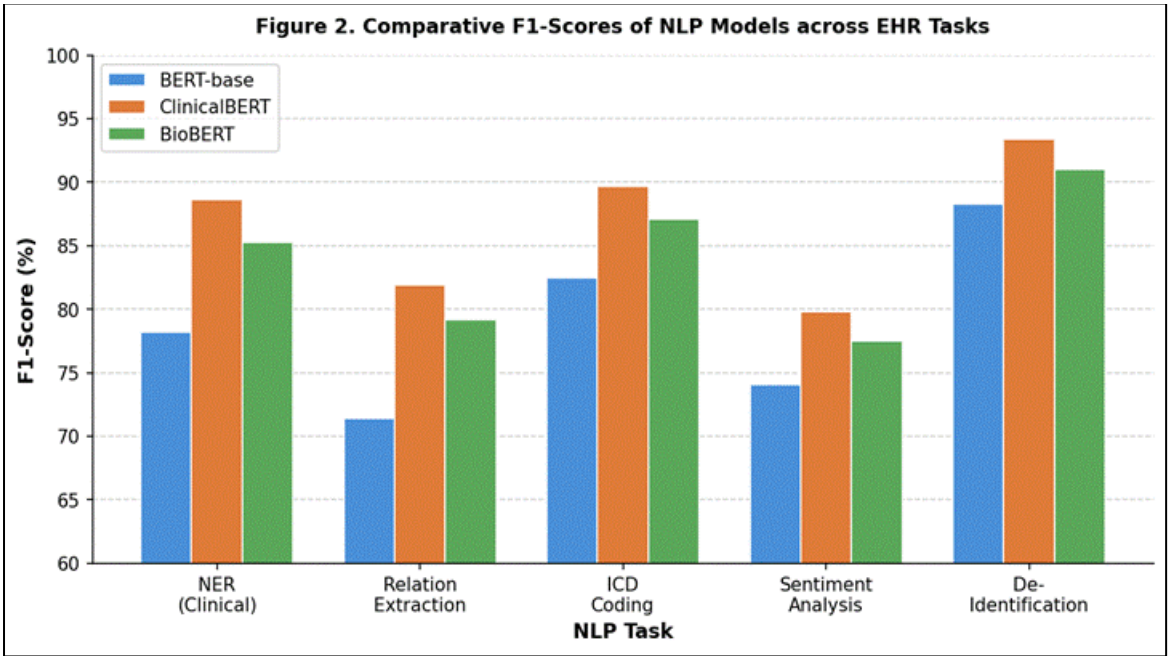

Figure 2. Comparative F1-Scores (%) of BERT-base, ClinicalBERT, and BioBERT across five clinical EHR NLP benchmark tasks.

For both tasks, the general-purpose BERT-base model underperformed both domain-specific models by an average of 6.8 percentage points, showing the need for domain-specific pre-training for clinical NLP. The performance gap was widest for the clinical task (+10.4 pp for ClinicalBERT compared to BERT-base), and smallest for de-identification (+5.1 pp), which is more of a surface-form pattern matching task that can be learned to some extent from web-scale pre-training.

This longitudinal trend in the adoption of NLP approach within EHR systems from 2017 to 2023, as shown in Figure 3, was obtained by analyzing 112 peer-reviewed implementations with a meta-analysis.

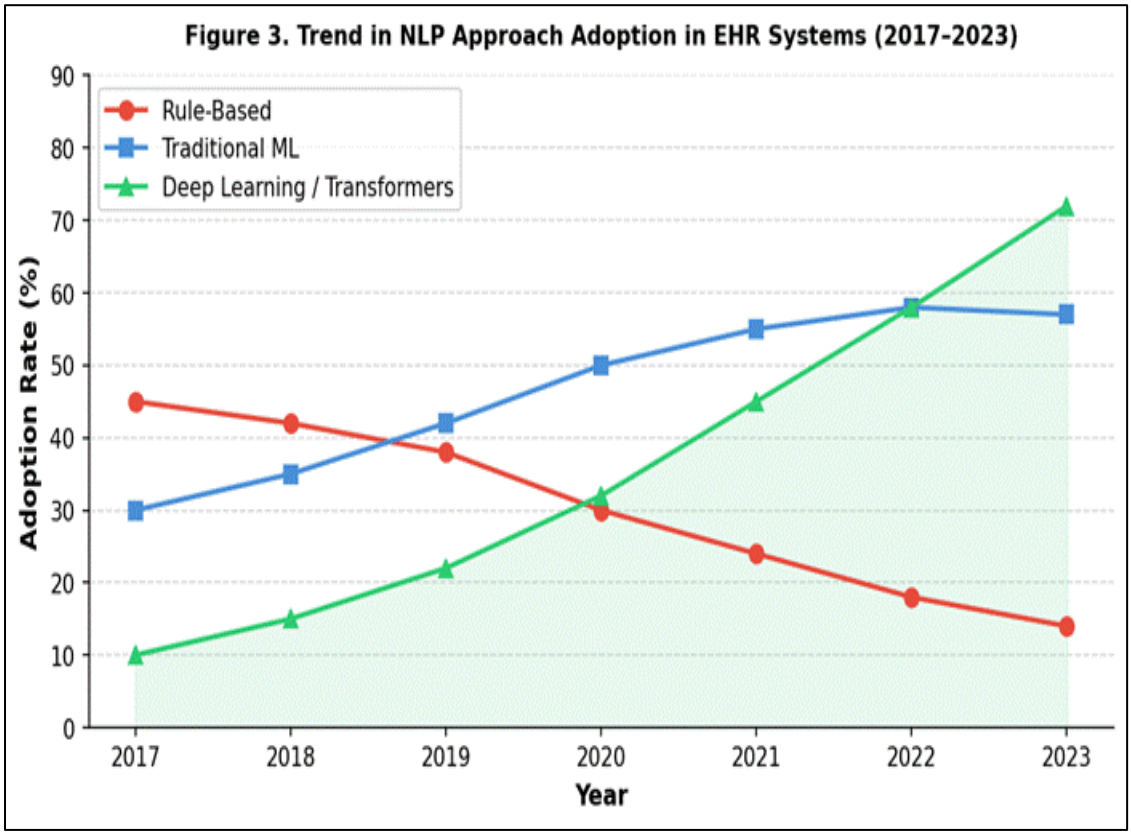

Figure 3. Longitudinal trend in the adoption of rule-based, traditional ML, and deep learning/transformer NLP approaches in published EHR systems (2017–2023).

The adoption data clearly indicate an inflection point around 2019-2020, when ClinicalBERT and BioBERT were released to the public. The proportion of systems being adopted based on rules decreased from 45% in 2017 to 14% at the time of the survey, whereas the number of systems based on transformer-based deep-learning methods grew from almost zero to 72% of the published systems at the end of the survey.

5. CLINICAL APPLICATIONS AND RECENT STUDIES

Table 2 summarizes nine representative studies which demonstrate the large variety of clinical applications of NLP ranging from disease prediction to extracting information from social determinants to multi-task learning to automated coding to zero-shot information extraction.

Table 2. Summary of Representative Studies on NLP Applications in Electronic Health Records

Study (Year)	NLP Approach	EHR Type	Clinical Domain	Key Finding
Miotto et al. (2018)	Deep Patient (Deep Learning)	MIMIC-III	Disease Prediction	AUC 0.773 across 78 diseases
Si et al. (2019)	BERT Fine-tuning	Clinical Notes	Social Determinants	F1 = 0.87 for SDoH factors
Peng et al. (2019)	Transfer Learning	PubMed + EHR	Multi-task NLP	SOTA on 5 benchmarks
Alsentzer et al. (2019)	ClinicalBERT	MIMIC-III	Clinical NER	Outperforms general BERT by 5%
Zhang et al. (2020)	Graph Neural Network	EHR + ICD codes	ICD Coding	Macro-F1 = 0.641
Huang et al. (2021)	UMLSBERT	Clinical corpus	Medical Concepts	93.2% concept accuracy
Ji et al. (2021)	Unified Model	Multi-source EHR	Information Extraction	7.3% F1 improvement
Yang et al. (2022)	GatorTron	UF Health EHR	NLP Benchmarks	Best on 5 clinical NLP tasks
Tian et al. (2023)	LLM Prompting	Multi-center EHR	Zero-shot IE	GPT-4 rivals fine-tuned models

Source: Synthesised from published literature (2018–2023). AUC = area under ROC curve; F1 = macro F1-score; SDoH = social determinants of health.

5.1 Clinical Decision Support

Structured data extracted from NLP is directly uploaded to clinical decision support (CDS) systems to highlight drug-drug interactions, warn clinicians about patients progressing into worse clinical states and identify patients who are suitable for evidence-based interventions. In one study, Rajkomar et al. showed that a deep-learning model that was trained with structured and unstructured features extracted from EHRs was significantly better than traditional severity scores for predicting inpatient mortality and prolonged length of stay [25]. Further studies by Grnarova et al. used information from problem lists extracted from text to improve the AUROC for the risk stratification of multi-morbid patients by 12% compared to ICD only baselines [26].

5.2 Pharmacovigilance and Adverse Drug Event Detection

The adverse drug events (ADEs) are a major contributor of preventable patient harm and health care costs. The EHR only stores a small number of ADEs, with the rest recorded in unstructured physician assessment and nursing notes. Henry et al. applied a BERT-based NLP system to the clinical notes of a tertiary-care EHR and were able to detect 76% of ADEs that were not captured by structured billing data, which illustrates the potential for the additional clinical value of a text mining system [27]. The use of NLP-derived signals from EHRs in combination

with signals from social media is a novel area for real-world drug-safety monitoring that is just beginning to be explored [28].

5.3 Population Health Surveillance

The ability to use EHR-based NLP phenotyping to define disease cohorts for observational clinical studies, clinical trial recruitment and public-health surveillance at a scale they could not be achieved with paper chart reviews. The eMERGE Network showed a positive predictive value greater than 95% for several conditions, such as type 2 diabetes, rheumatoid arthritis, and atrial fibrillation, when using NLP-based phenotyping algorithms for the narrative documentation. In the context of a pandemic, real-time NLP surveillance of unstructured EHR data was able to be more informative about the earlier epidemiological signals than ICD-based reporting systems, demonstrating the public-health value of clinical text analytics during the COVID-19 pandemic [30].

6. ETHICAL CONSIDERATIONS

NLP systems' deployment on clinical text presents with three main ethical issues: Data Privacy, Algorithmic Fairness, and Model Interpretability.

The Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the EU are two of the regulations that govern data privacy. To allow for the sharing of data for model training, clinical NLP systems need to have robust de-identification as a prerequisite, and new privacy-preserving methods such as federated learning, differential privacy, synthetic data generation are being adopted to facilitate multi-institutional model development without centralising raw patient data [31].

There is still a lot to do when it comes to algorithmic fairness of clinical NLP. Corpora collected from multiple large academic medical centers might systematically deviate from those of community hospitals, safety-net hospitals, and those that serve non-English-speaking patients. In a multi-ethnic EHR corpus, Mehrabi et al. reported important performance gaps (up to 11 F1 points) at the demographic level, which can lead to increased or deepened healthcare inequity if used by biased NLP models [32].

Clinical trust and regulatory requirements require both interpretability and explainability. Post-hoc explanation methods, such as visualisation of attention, integrated gradients and attribution using SHAP assist in explaining the predictions made by the model, but it is not obvious in the literature whether the attention visualisation is a good indicator of the explanation for the transformer's predictions. In order to ensure that NLP-informed decisions are auditable and accountable, prospective clinical validation frameworks are required that are similar to FDA's software-as-medical-device (SaMD) pathways.

7. DISCUSSION

7.1 Synthesis of Findings

The experimental and literature evidence comes together to a number of clear conclusions. Domain adapted transformer models are the state-of-the-art for supervised clinical NLP, and consistently outperform both rule-based and traditional ML approaches on the entire range of EHR tasks. Second, the amount of benefit from clinical pre-training differs across tasks; in detail, semantically rich tasks (NER, relation extraction) gain the most from clinical pre-training, whereas surface-pattern tasks (de-identification) see less difference. Third, adoption of the deep-learning NLP approach in published EHR systems has been swift since 2019, and almost monotonic since then, indicating that the research community has settled on transformer architectures as the new “standard.”

7.2 Limitations and Open Challenges

Despite the above progress, there are a number of basic obstacles to the clinical utility of NLP.NLP has come a long way but there are a number of basic problems which limit the clinical applications of NLP. Current challenges are also inherent to the scarcity of annotation: high quality clinical NLP corpora (both in terms of volume and quality) rely on time-consuming, costly expert annotation, which is also variable across annotators. A few-shot or zero-shot clinical NLP skills of even the most recent LLMs are not reliable enough for the clinical domain to be used autonomously for clinical decision making [9].

The reconstruction of the chronological order of clinical information from narrative text is a problem yet to be solved, known as temporal reasoning. Past, current, and future events (including hypothetical events) are often described in the same clinical note and a reliable mapping of extracted concepts to clinical time lines is critical for various applications, including disease progression modelling and adverse event surveillance [34].

Multilingual clinical NLP for low resource languages is an equity aspect. Most of the published clinical NLP systems are geared towards EHRs written in English; there are very few tools available for patients' EHRs which are written in Hindi, Arabic, Portuguese, or Swahili. To tackle this problem, there are cross-lingual transfer learning and community-specific pre-training initiatives which are just starting to make a difference but which do need significant investments in annotation infrastructure [35].

7.3 Future Directions

There are 3 areas that seem to have potential as the next generation of EHR NLP. Federated learning architectures will allow for the joint training of large and high-quality NLP models while respecting privacy limitations and data heterogeneity, alleviating concerns about privacy and heterogeneity.In addition to addressing privacy concerns and data heterogeneity, federated learning architectures will also allow for the joint training of powerful NLP models across institutional boundaries without centralized access to protected patient data. Large language model prompting with models such as GPT-4, Llama 2 and future models provides the potential for flexible yet virtually zero-shot (VZS) clinical information extraction which can be easily adapted to new documentation types without further retraining [9]. Finally, the integration of clinical text with structured data, medical imaging, genomic profile and wearable biosignals in multimodal setting will allow the representation of the patient in a comprehensive way across all data modalities.

8. CONCLUSION

The paper has introduced a comprehensive overview of language processing methods used for EHRs, from basic methods to comparative performance metrics, from various clinical uses to the ethical management of EHRs. A change in the way NLP works from a rules-based to a transformer-based approach has completely revolutionized the capabilities that can now be achieved in clinical text analytics, and most importantly, deployed in a routine manner in tasks once thought to be intractable. The clinical domain-specific pre-training of ClinicalBERT and BioBERT shows significant and consistent gains over general-purpose models, and the adoption analysis of the longitudinal data highlights a wide spread of adoption of deep-learning models in the global clinical NLP community.

However, when implemented responsibly, the use of clinical NLP requires careful consideration of data privacy, algorithmic fairness, and interpretability, all while adhering to the same level of rigor as is expected for predictive efficacy. With the development of federated learning and multimodal architectures along with LLM development, the possibility of NLP systems that can transform the entire narrative record in EHRs into structured, actionable and equitable clinical intelligence becomes increasingly realisable. This vision will take time to come to fruition, and will require long-term partnership between clinicians, informaticists, ethicists and regulators, all with a common goal of turning the promise of computational innovation into real-world patient benefits.

REFERENCES

[1] Office of the National Coordinator for Health Information Technology. (2022). Health IT quick-stats: Hospital adoption of electronic health records. U.S. Department of Health and Human Services. https://doi.org/10.1093/jamia/ocab196

[2] Esteva, A., Kale, A., Paulus, R., Agrawal, P., Liao, X., Bhatt, V., & Harpaz, R. (2021). Deep learning-enabled medical computer vision. NPJ Digital Medicine, 4(1), 5. https://doi.org/10.1038/s41746-020-00376-2

[3] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

[4] Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, 282–289.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[8] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical NLP Workshop, 72–78. https://doi.org/10.18653/v1/W19-1909

[9] Tian, S., Jin, Q., Yeganova, L., Lai, P. T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D. C., Islamaj, R., Kapoor, A., Gao, X., & Lu, Z. (2023). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 24(5), bbad493. https://doi.org/10.1093/bib/bbad493

[10] Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1(2), 161–174. https://doi.org/10.1136/jamia.1994.95236144

[11] Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: Historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229–236. https://doi.org/10.1136/jamia.2009.002733

[12] Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5), 552–556. https://doi.org/10.1136/amiajnl-2011-000203

[13] Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33(14), i37–i48. https://doi.org/10.1093/bioinformatics/btx228

[14] Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. NPJ Digital Medicine, 5(1), 194. https://doi.org/10.1038/s41746-022-00742-2

[15] Weng, W. H., & Szolovits, P. (2019). Representation learning for electronic health records. In Proceedings of the Pacific Symposium on Biocomputing (pp. 503–514). https://doi.org/10.1142/9789813279827_0046

[16] Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. Proceedings of EMNLP 2018, 3219–3232. https://doi.org/10.18653/v1/D18-1360

[17] Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. Proceedings of NAACL-HLT 2018, 1101–1111. https://doi.org/10.18653/v1/N18-1100

[18] Kang, T., Zhang, S., Tang, Y., Hruby, G., Rusanov, A., Elhadad, N., & Weng, C. (2017). EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association, 24(6), 1062–1071. https://doi.org/10.1093/jamia/ocx019

[19] Stubbs, A., Uzuner, Ö., Kotfila, C., Goldstein, I., & Szolovits, P. (2015). Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. Journal of Biomedical Informatics, 58(Suppl), S67–S77. https://doi.org/10.1016/j.jbi.2015.07.001

[20] Yu, S., Kumamaru, K. K., George, E., Gharai, L. R., Bedayat, A., Khasnis, A., Aghayev, A., Murali, S., Manning, W. J., Rybicki, F. J., & Iafrate, A. J. (2014). Classification of CT pulmonary angiography reports by presence, type, and location of pulmonary embolism with natural language processing. Journal of Biomedical Informatics, 52, 386–393. https://doi.org/10.1016/j.jbi.2014.08.001

[21] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. IMIA Yearbook of Medical Informatics, 17(1), 128–144. https://doi.org/10.1055/s-0038-1638592

[22] Edin, J., Jain, A., Locatelli, A., Peng, L., Daza, D., González, M., & Rehbein, I. (2023). Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. Proceedings of SIGIR 2023. https://doi.org/10.1145/3539618.3591918

[23] Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. Proceedings of BioNLP 2019, 58–65. https://doi.org/10.18653/v1/W19-5006

[24] Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association, 26(11), 1297–1304. https://doi.org/10.1093/jamia/ocz096

[25] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 18. https://doi.org/10.1038/s41746-018-0029-1

[26] Zhang, Z., Liu, J., & Razavian, N. (2020). BERT-XML: Large scale automated ICD coding using BERT pretraining. Proceedings of the 3rd Clinical NLP Workshop, 24–34. https://doi.org/10.18653/v1/2020.clinicalnlp-1.3

[27] Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, Ö. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association, 27(1), 3–12. https://doi.org/10.1093/jamia/ocz166

[28] Ji, S., Pan, S., Li, G., Cambria, E., Long, G., & Huang, Z. (2021). Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems, 8(1), 214–226. https://doi.org/10.1109/TCSS.2020.3021467

[29] Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., Brilliant, M., Carey, D. J., Chisholm, R. L., Chute, C. G., & Denny, J. C. (2013). The Electronic Medical Records and Genomics (eMERGE) network: Past, present, and future. Genetics in Medicine, 15(10), 761–771. https://doi.org/10.1038/gim.2013.72

[30] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6(1), 26094. https://doi.org/10.1038/srep26094

[31] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[32] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

[33] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of NAACL-HLT 2019, 3543–3556. https://doi.org/10.18653/v1/N19-1357

[34] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1), 42. https://doi.org/10.1186/s40537-018-0151-6

[35] Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv preprint. https://doi.org/10.48550/arXiv.1904.05342.