Clinical Acronym and Abbreviation Disambiguation in Electronic Health Records: A Systematic Review

Binod Kumar Mishra[1], Dr. Subrata Chowdhury [2]
[1]Chandigarh University, Mohali, Punjab, India, [2]SVCET College, Chittoor, Andra Pradesh, India,
[1,2]Lincoln University, Malaysia,
[1]bkmishra21@gmail.com,[1]bkmishra21@ieee.org,
[2]subrata895@gmail.com,[2]pdfsv.subrata@lincoln.edu.my

Abstract: Electronic Health Records (EHRs) use universal clinical acronyms and abbreviations which provide the ability to concisely record clinical data, but also create a lot of ambiguity that obstructs the automated analysis of clinical texts. Proper decoding of acronyms is required to have credible applications of clinical natural language processing (NLP) including decision support, information retrieval, and identification of patient cohorts. In the last twenty years, spanning rule-based systems, traditional machine learning, deep learning, and graph-based models are only a few approaches to the wide range of approaches. They are suggested to deal with this challenge. It is a PRISMA-compliant systematic review that summarizes the available literature on clinical acronym and abbreviation disambiguation in EHRs. In the given paper, the data sets, methodology, evaluation measures, and application scenarios are analyzed and a systematic taxonomy of methods is presented. The review provides the emergent trends, research gaps that remain consistent, and future directions of developing the robustness of disambiguation systems, which can be utilized in clinical settings.

Keywords: Clinical NLP, Acronym Disambiguation, Electronic Health Records, Systematic Review, Biomedical Text Mining.

1. Introduction

Electronic Health Records (EHRs) have become an intrinsic part of the contemporary healthcare system, allowing storing and sharing patient data in digital form both within the clinical facilities. Besides organized fields, EHRs include huge amounts of unstructured clinical explanations including discharge summaries, progress notes, operative reports, and radiology interpretations. Free-text documents contain a lot of clinical knowledge, but are difficult to process automatically because of their informal style, domain-specific language, high use of acronym and abbreviations [1], [2].

Healthcare professionals commonly use clinical acronyms and abbreviations to enhance efficiency in documentation and lower the cognitive burden in time-sensitive clinical processes. Nevertheless, most abbreviations are ambiguous by nature, and the meaning of an abbreviation will be different based on the situation, specialty, and the condition of a patient. As an example, “RA” can be treated as rheumatoid arthritis, right atria, or room air, whereas “CP” can be used as a chronicling of chest pain, cerebral palsy or clinical pathway. This ambiguity is a major issue to human interpretation, as well as automated clinical Natural Language Processing (NLP) systems [3], [4].

Precise clinical acronym and abbreviation disambiguation is a pre-requisite to a broad spectrum of downstream clinical NLP applications, which include clinical entity recognition, information extraction,

decision support systems, patient cohort identification, and clinical summarization. Disambiguation errors may spread through such pipelines, resulting in faulty data extraction, biased analytics, and even unsafe clinical advice may be provided [5], [6]. As a result, the acronym disambiguation has been identified as a problematic subtask in the clinical NLP and biomedical text mining studies.

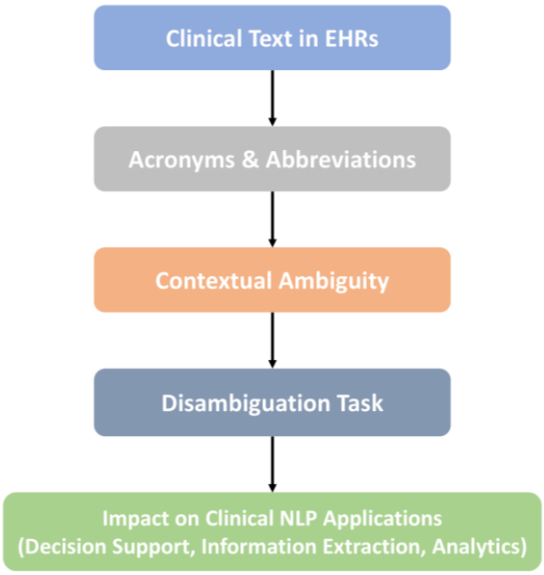

Figure 1. Conceptual overview of clinical acronym and abbreviation disambiguation in electronic health records and its role in supporting downstream clinical natural language processing applications.

The conceptual map of the clinical acronym and abbreviation disambiguation task in the wider Electronic Health Record analytics pipeline is presented in Figure 1. Clinical stories are characterized by ambiguous acronyms, whose meanings are determined by contextual and domain-specific information. Proper disambiguation is the key to making downstream clinical NLP applications involving information extraction, clinical decision support, and patient cohort identification. This summary shows that the acronym de-acronymization plays a key role in converting unstructured clinical text to operational clinical knowledge.

Initial research initiatives were mainly based on the rule-based systems and ad-hoc abbreviation dictionaries. Although such methods provided interpretability, domain transparency, they were characterized by poor scalability, low coverage, and high maintenance costs, especially due to the changing medical vocabulary and institution-specific documentation practices [7]. To address these drawbacks, conventional machine learning techniques were presented, where acronym disambiguation is presented as a supervised classification problem applied to contextual features, generated based on surrounding words, part-of speech labels, and section headings [8], [9]. With the help of these approaches enhanced the flexibility, but they still trusted on handmade characteristics and annotated data.

With the introduction of deep learning, there was a major change in the field of clinical acronym disambiguation research. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based language models were also shown to be effective neural architectures through learning contextual representations via text alone [10], [11]. The further development of the field was

provided by large-scale pre-trained biomedical language models which utilized domain-specific corpora. But even despite their success, deep-learning methods can be difficult with uncommon abbreviations, and inter-institutional domain changes, and their interpretability- which is especially important in clinical practice [12], [13].

In more modern times, scholars have analyzed graph-based and hybrid methodologies that directly represent the relationship between clinical terms, abbreviations, and biomedical concepts. These methods will improve semantic reasoning of relational structures and external sources of medical knowledge, beyond the textual representations in the form of line formations [14], [15]. Such approaches are promising but also come with issues of computational complexity, scalability, and standardized evaluation.

Considering the fast-moving methodological development, growing clinical significance, and evolving literature, there can be a strong necessity of a structured and recent synthesis of the existing research on clinical acronym and abbreviation disambiguation in EHRs. Available surveys are either too old, or are too small, or are too narrow. Furthermore, the lack of consistency in datasets, evaluation measures, and reporting patterns complicates the comparisons of the outcomes across the studies and the evaluation of the real-life application.

To fill these gaps, the given systematic review contributes as follows:

1. Thorough synthesis of available methods of disambiguation of clinical acronym and abbreviations in EHRs.
2. Systematic taxonomy of techniques between rule-based systems to the current graph-based techniques.
3. Critical interpretation of data-sets, evaluation measures, and experimental procedures.
4. Detection of gaps in the research and address of outstanding issues that impede clinical implementation.
5. Research implications to suggest the future course of robust, generalizable, and clinically trustworthy disambiguation systems.

The purpose of this review is to become a reference source that can be used by researchers and professionals dealing with clinical NLP and healthcare informatics by summarizing the existing knowledge and announcing new trends.

2. Review Methodology

This systematic review was undertaken following the Preferred Reporting Items of Systematic Reviews and meta-analyses (PRISMA) guideline since transparency, reproducibility and methodological rigor were sought in identification, selection, and synthesis of relevant studies [15]. The review protocol was created to be very inclusive to capture research on clinical acronym and abbreviation disambiguation in Electronic Health Records (EHRs).

The research questions included in the review were as follows:

- RQ1: What are the computational methods suggested to be used in the process of disambiguation of clinical acronym and abbreviations in EHRs?
- RQ2: Which datasets and text types of clinical data are widely used?
- RQ3: What are the evaluation metrics that are used in studies?
- RQ4: What are the gaps and future challenges in the literature?

These questions have been developed based on the existing guidelines to systematic review in the research of the biomedical informatics and software engineering [16], [17]. Extensive search in literature was conducted in the following electronic databases:

- IEEE Xplore
- PubMed
- Scopus
- Web of Science
- Google Scholar

The search was conducted using combinations of keywords and Boolean operators, including: ("clinical acronym disambiguation" OR "abbreviation expansion") AND ("electronic health records" OR "clinical notes") AND ("clinical NLP" OR "medical text mining")

Peer reviewed journal articles and conference papers in English were only considered. There was no limit on the year of publication as both the early and recent studies were to be covered. There were 512 records that came up during the initial search of the database. The screening of the studies was left with 468 studies after eliminating the 44 duplicates. Title and abstract screening 332 articles were filtered out because they were irrelevant to the clinical disambiguation of acronyms or were not clinical domains. With Full-text assessment, the rest of the 136 articles were evaluated to be eligible. Out of these, 92 articles were eliminated because of the lack of specific-domain abbreviation disambiguation, inadequate methodology description, or lack of quantitative analysis. Inclusion: 44 articles were chosen as a part of the qualitative synthesis. Figure 2 summarizes this process of selection with the help of a PRISMA flow diagram.

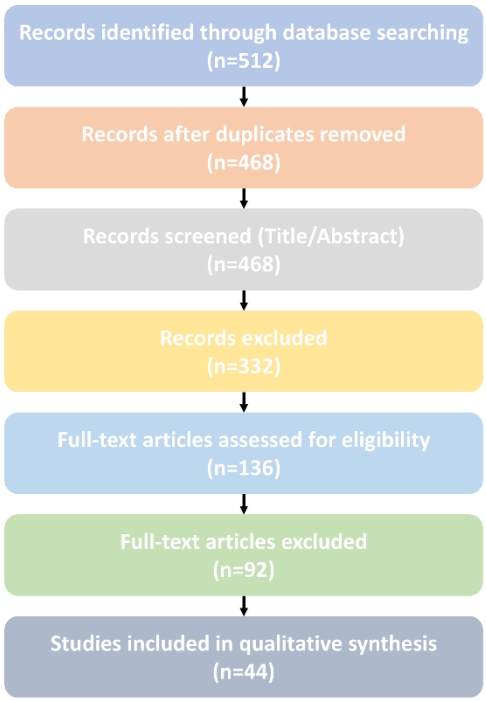

Figure 2. PRISMA flow diagram for the systematic review of clinical acronym and abbreviation disambiguation

Table 1 presents the PICOS framework along with the corresponding inclusion and exclusion criteria adopted in this systematic review to ensure the selection of methodologically rigorous and clinically relevant studies.

Table 1. PICOS Framework with Inclusion and Exclusion Criteria

Component	Inclusion Criteria	Exclusion Criteria
Population (P)	Clinical text from Electronic Health Records (EHRs), including discharge summaries, progress notes, and other clinical narratives containing acronyms or abbreviations	Non-clinical text, general-domain corpora, or non-EHR datasets
Intervention (I)	Computational methods for acronym and abbreviation disambiguation, including machine learning, deep learning, NLP-based, and hybrid approaches	Non-computational, purely manual, or non-NLP-based approaches
Comparison (C)	Baseline methods, rule-based systems, or alternative computational models used for comparative evaluation	Studies without any comparative or baseline analysis
Outcomes (O)	Quantitative performance evaluation using metrics such as accuracy, precision, recall, F1-score, or top-*k* accuracy	Studies lacking quantitative results or evaluation metrics
Study Design (S)	Peer-reviewed journal articles and conference proceedings reporting experimental validation	Non-peer-reviewed articles, editorials, surveys without experiments, theses, or gray literature

These criteria are consistent with recommended practices for systematic reviews in healthcare informatics [17], [18].

Data Extraction and Synthesis

For each included study, the following information was extracted:

- Publication year and venue
- Clinical dataset used (e.g., i2b2, MIMIC)
- Type of clinical text
- Methodological approach
- Evaluation metrics
- Reported performance and limitations

Because of the heterogeneity in the datasets and evaluation protocols, a qualitative narrative synthesis was performed, rather than a meta-analysis, as it is a best practice in clinical NLP reviews [19]. Qualitative assessment of methodological quality of included studies was conducted based on clarity of problem formulation, transparency of dataset, rigor of evaluation and reproducibility. The interpretation of studies

which had little or no experimental detail was done with a heavy dose of caution as advised in previous systematic review methods [18], [20].

3. Taxonomy of Approaches

The study of the clinical acronym and abbreviation disambiguation in Electronic Health Records has progressed significantly within the last twenty years. Existing methods can be ranked into four major categories based on methodological features, learning paradigms, and the representations strategies: (i) rule-based and dictionary-driven ones, (ii) classical machine learning methods, (iii) deep learning-based methods, and (iv) graph-based and hybrid methods. The given taxonomy is a systematic device that explains methodological trends and comparative advantages related to the studies.

Dictionary-Based and Rule-Based: The oldest type of approach towards clinical acronym and abbreviation disambiguation is rule-based and dictionary-driven methods. The techniques would generally be based on lexicons of abbreviation expansion, manually curated or not, and rules of thumb, specific to a local context, document section titles, or document structure [21], [22]. By way of illustration, there are systems that solve abbreviations by searching through contextual key words in a fixed window, or by taking advantage of the structured parts of clinical notes, e.g., Assessment or Medication sections.

Despite the high level of interpretability and transparency provided by rule-based methods, which is crucial in clinical environments, these methods have a low level of efficacy because of various factors. They demand a lot of manual labor to maintain and update dictionaries, have trouble with hidden or institution-specific abbreviations, and are not strong enough to address linguistic variation in clinical narratives [23]. Therefore, such techniques cannot be easily scaled to large and heterogeneous EHR corpora and are not common as standalone applications in recent research.

Conventional Machine Learning Methods: To surmount the inflexibility of rule-based systems, classical methods of machine learning treated acronym disambiguation as a supervised classification problem. In such techniques, the ambiguous acronyms are considered the target variables and the surrounding textual context is now modeled in terms of handcrafted features like bag-of-words, n-grams, part-of-speech labels, and section identifiers [24], [25]. Some of the commonly used classifiers are the Naive Bayes, Support Vector Machines, Logistic Regression and Decision Trees.

These procedures proved to be more adaptable and generalized than rule-based systems, especially in case there were adequate annotated data. Nonetheless, feature engineering is an essential aspect of their performance and demands domain knowledge and might not be applicable across datasets or institutions. Moreover, conventional machine learning models are frequently incapable of obtaining long-range contextual associations and semantic subtleties found in clinical text [26]. With the advent of deep learning mechanisms, purely feature-based models became less popular.

Deep Learning-based solutions: Deep learning was an important breakthrough in disambiguating clinical acronym and abbreviations by allowing learning contextual representations in raw text. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are examples of neural architectures that were first used to model local and sequential context around acronyms [27]. These models minimized the use of manual feature engineering and, they gained significant performance over other traditional classifiers.

Transformer-based language models that were trained on large biomedical and clinical datasets have become more dominant in recent years. Most they can capture rich contextual semantics, and have been

broadly used in disambiguation tasks through EHRs using models like BioBERT, ClinicalBERT and others [28], [29]. Although deep learning methods are effective, they have several weaknesses. They are usually large in size, prone to cross-institutional domain shifts, and have little interpretability, which is a significant issue in clinical decision support [30].

Graph-Based and Hybrid Methodologies: Graph-based and hybrid techniques are a newer methodological approach that seeks to overcome the constraints of a completely text-based model. These methods model the connections between acronyms, surrounding words, clinical concepts and in some cases external biomedical knowledge in the form of graphs. Nodes can be words, abbreviations or concepts and the edges can reflect semantic, syntactic or co-occurrence relationships [31], [32].

Graph-based methods are more able to model dependencies that are not easily modeled with linear text representations alone because relational reasoning and collective inference are possible. Hybrid methods typically make use of both contextual embeddings and deep language models trained to make use of both local context and global semantic structure [33]. Though these applications demonstrate positive outcomes, they come with the issues of computational complexity, scalability, and the requirement of standardized graph building protocols. These are therefore yet to be adopted in the actual clinical systems.

Relative Overview of Approaches: Methodological advancement between rule-based systems and learning-based and relational models. Although newer techniques prove to be better in terms of performance and flexibility, there is no single paradigm that can be used to address issues that may arise with the areas of generalization, explainability, and deployment. This fact highlights the necessity to persist with research on hybrid and scalable frameworks that can balance between performance and clinical usability.

4. Data and Measures of Evaluation

Data sets and evaluation procedures used to experiment with acronym and abbreviation disambiguation systems in clinical settings have a great influence on their performance and generalizability. The section examines widely used datasets, their weaknesses, and evaluation metrics applied in different studies.

Clinical Acronym and Abbreviation Disambiguation Datasets: The disambiguation of clinical acronym and abbreviations have been tested on a range of datasets. Most datasets are based on actual Electronic Health Records, which represent real-life language application but also create issues of privacy, cost of annotation and domain specificity.

- **i2b2 Clinical Notes Dataset:** One of the most popular benchmark datasets within the clinical NLP research is the i2b2 (Informatics for Integrating Biology and the Bedside) dataset. It has annotated examples of clinical narratives, discharge summaries, progress notes, labelled abbreviations, and their expansions [34]. The i2b2 dataset has greatly been utilized in supervised learning methods because its annotations are of high quality and its evaluation protocols are standard. But its cross-domain generalization is limited by its small size and institutional particularism.
- **CASI Dataset:** The Clinical Abbreviation Sense Inventory (CASI) dataset is created to be used in disambiguation tasks of acronym and abbreviations. CASI consists of several senses to usual clinical abbreviations, which are annotated in different clinical scenarios [35]. This data has been commonly utilized to test traditional machine learning and neural models. Although CASI is

narrowly designed, it represents a smaller scope of clinical specialties, and it is not representative of the variance of language found in large-scale EHR systems.

- **MIMIC-III and MIMIC-IV Data:** MIMIC-III and MIMIC-IV, the Medical Information Mart of Intensive Care (MIMIC) datasets, are extensive publicly available critical care databases that consist of de-identified clinical records, laboratory outcomes, and discharge reports [36], [37]. These datasets are becoming useful in acronym disambiguation research because of their volume and variety. Nevertheless, there are no explicit annotations of acronym sense in MIMIC datasets, and the researcher must use weak supervision or distant supervision techniques, which could lead to labeling noise.

Some of the studies use proprietary or institution-specific EHR data gathered in hospitals or clinical partners [38]. Although these datasets frequently capture realistic deployment conditions, they are limited in their availability, which limits reproducibility and ability to compare across studies. Absence of common standards is one of the biggest barriers to comparative appraisal.

Problems related to Clinical Datasets: Current datasets have several challenges despite their usefulness. Clinical acronym annotation is not domain-neutral, and it is time-consuming, which limits the size of datasets. Also, the use of abbreviations is highly inconsistent between institutions, specialties, and even individual clinicians, which causes bias in datasets and decreased generalization [39]. Privacy policies also limit the data sharing, which contributes to the lack of publicly available, big-scale annotated datasets.

Evaluation Metrics: Measures of clinical acronym and abbreviation disambiguation are typically borrowed, in NLP, classification tasks. Nevertheless, the choice of metrics and reporting making vary among studies, thus making comparisons challenging.

- **Accuracy:** The most reported measure is accuracy, which is the percentage of words of the correctly disambiguated acronyms. Intuitive as it is, it can be inaccurate on dataset with skewed distributions of senses [40].
- **Precision, Recall, and F1-Score:** Precision, recall and F1-score give a more detailed analysis especially when it is a multi-class scenario and it is imbalanced. These measures are common in the determination of the model robustness and error trade-offs [41].
- **Top-k Accuracy:** Other studies indicate top-k accuracy, which is a factor that determines the presence of the correct expansion among the top-k predictions. This measure is especially pertinent in a clinical decision support case, where a myriad of candidate expansions can be offered to clinicians [42].
- **Metrics of Computational Efficiency and Scalability:** Recent research also puts growing emphasis on computational metrics, including, but not limited to, training time, inference latency, and memory consumption, of graph-based and deep learning methods [43]. The measures play a key role in evaluating practical deployability in clinical settings.

Limitations to Existing Evaluation Practices: Most of the current assessments involve intrinsic metrics and fail to monitor the downstream effect of acronym disambiguation on clinical NLP systems including entity recognition or decision support. Also, the absence of standardized benchmarks and cohesive reporting methods hinder reproducibility and comparison between studies [44]. The limitations are crucial and should be tackled to further develop the field to a clinically meaningful evaluation.

5. Research Gaps and Future Directions

Although the use of clinical acronym and abbreviations in Electronic Health Records has not been completely studied, its disambiguation is a problem that is not solved yet and has become a controversial issue. This review presents some of the research gaps as critical problems hindering the clinical applicability and generalizability of current methods and a future research opportunity.

Small Cross-domain, and Cross-institution generalization: Most of the available research assesses their approach using a single dataset or in one institution, e.g. i2b2 or MIMIC-derived corpora. But the usage of abbreviations differs greatly in different hospitals, different departments in the clinical, as well as in different geographical areas. Algorithms trained on data unique to the institutions tend to show a significant drop in performance when transferred to unknown areas [45]. In the future, cross-dataset assessment, domain adaptation, and transfer learning methods should be the center of attention to enhance cross-system robustness in heterogeneous EHR environments.

The reliance of the authors on annotated clinical data: The literature is dominated by supervised learning techniques that have high requirements in terms of manually annotated datasets, expensive, time-consuming, and domain knowledge. The reliance restricts scalability, especially to uncommon abbreviations and low-resource clinical environments. Whereas medical ontologies have been employed in weak supervision and distant supervision, those strategies are not exploited [46]. Further research on semi-supervised, self-supervised, and active learning paradigms should be conducted in future to decrease the annotation load without compromising performance.

Weak Modelling of Clinical Semantics and Context: Although deep learning models can learn contextual information, they do not tend to learn structured clinical semantics, including hierarchical relationship, temporal relationship, and emerging medical knowledge. Most of the existing systems handle acronyms as individual classification problems, but not as a part of more extensive clinical reasoning processes [47]. Research in the future ought to consider the models that combine the contextual text representations with the structured clinical knowledge and temporal patient data.

Scalability and Computational Constraints: More complicated models, especially graph-based and hybrid models, have a huge computational overhead with the intricate representations and inference. This restricts their usability in a real-time environment in a clinical context where latency and resource usage matter [48]. Studies on model compression, lightweight architectures, and efficient inference strategies are required to trade-off between performance and deployability.

Explainability and Clinical Trust: Another significant obstacle in adoption by clinicians is explainability. Medicine is a field of expertise where clinical practitioners need clear and understandable lines of reasoning to be confident in automated systems, particularly in critical decision-making contexts. Even though there are current methods to give attention-based or rule-derived explanations, they are not satisfactory to make them a routine clinical tool [49]. Research of explainable AI frameworks specific to clinical NLP work should be pursued in the future.

Inconsistent Standards and Assessment Procedures: The lack of standardized datasets, metrics of evaluation and reporting practices makes it hard to compare method across studies. Preprocessing, sense inventory, and evaluation arrangement differences cause bias and reduce the reproducibility [50]. It should be noted that setting common standards, open sets of data and community-based assessment campaigns are an essential way forward.

6. Discussion

This systematic review demonstrates that the methodology of clinical acronym and abbreviation disambiguation studies has undergone a visible change through the years to encompass the evidence-based systems of rule-based systems to include data-driven systems of learning, and the more recent, relational and hybrid paradigms. Every change of method indicates an effort to overcome the weaknesses of previous methods, especially in the areas of scale, contextual interpretation and semantic argument.

Although rule-based and dictionary-driven approaches are interpretable, they cannot be used in modern EHR settings that are linguistically diverse and need terms to evolve quickly. Early methods in machine learning enhanced flexibility but were limited by extra engineering of features and by the narrow semantic model of features. The approaches to deep learning had a significant impact on the progress of the field because they were able to learn deep contextual representations, but presented novel problems regarding data dependency, interpretability, and generalization.

Later graph-based and hybrid methods show evidence of making improvements in modeling relationships between clinical concepts and contextual factors. They are however more complex of nature and this brings into question their effectiveness and feasibility in computation as well as application. Notably, the review notes that the majority of studies test acronym disambiguation as a standalone activity, without regard to its subsequent contribution to more comprehensive clinical NLP systems like decision support, clinical summarization, or patient outcome analysis.

The other important finding is the lack of connection between the methodological performance and clinical relevance. The high intrinsic evaluation scores may not always lead to a better clinical decision-making process or workflow. This loophole highlights the necessity to develop assessment frameworks to measure clinical impact in the real world, user trust, and integration viability.

Altogether, it can be concluded that methodology innovation has become faster, but the possibility to transfer research developments into clinically significant and reliable systems is also something the research has not figured out yet.

7. Conclusion

This systemic review has been able to provide an elaborate synthesis of studies relating to clinical acronym and abbreviation disambiguation in Electronic Health Records. The structure and organization of existing studies into a systematic taxonomy, data analysis, and data evaluation practices, and uncovering gaps in research revealed by multiple studies contribute to the review as a consolidated knowledge of the current field.

The results show that despite the tremendous advancements achieved, especially on the learning based and relationship methods, there are still considerable bottlenecks concerning the generalization, dependency on data, explainability, scalability, and standard evaluation. To resolve the challenges, it is necessary to deploy acronym disambiguation systems successfully in the real clinical situation.

Conclusively, further studies ought to transcend limited methodological enhancement and embrace holistic, clinically based worldviews that have the capacity of focusing on robustness, transparency, and practical impact. The review is supposed to serve as a reference base of researchers and practitioners conducting their work in clinical NLP and healthcare informatics, as well as to contribute to the achievements of the next generations of the systems capable of providing reliable support of clinical decision-making and healthcare analytics advancement.

References

1. N. Hastie, R. Tibshirani, J. Friedman, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-21606-5.
2. A. Alhelbawy and R. Gaizauskas, "Collective Named Entity Disambiguation using Graph Ranking and Clique Partitioning Approaches," pp. 1544–1555, Accessed: Jan. 08, 2026. [Online]. Available: https://lucene.apache.org/
3. A. Le Glaz et al., "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," J Med Internet Res 2021;23(5):e15708 https://www.jmir.org/2021/5/e15708, vol. 23, no. 5, p. e15708, May 2021, doi: 10.2196/15708.
4. A. Ben Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," Journal of Biomedical Semantics 2011 2:5, vol. 2, no. 5, pp. S4-, Oct. 2011, doi: 10.1186/2041-1480-2-S5-S4.
5. C. Zhang, D. Biś, X. Liu, and Z. He, "Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks," BMC Bioinformatics 2019 20:16, vol. 20, no. 16, pp. 502-, Dec. 2019, doi: 10.1186/S12859-019-3079-8.
6. E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings," pp. 72–78, Jul. 2019, doi: 10.18653/V1/W19-1909.
7. O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," Nucleic Acids Res, vol. 32, no. suppl_1, pp. D267–D270, Jan. 2004, doi: 10.1093/NAR/GKH061.
8. D. Moussallem, A. C. N. Ngomo, P. Buitelaar, and M. Arcan, "Utilizing knowledge graphs for neural machine translation augmentation," K-CAP 2019 - Proceedings of the 10th International Conference on Knowledge Capture, pp. 139–146, Sep. 2019, doi: 10.1145/3360901.3364423;TOPIC:TOPIC:CONFERENCE-COLLECTIONS.
9. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," IEEE Trans Neural Netw Learn Syst, vol. 32, no. 1, pp. 4–24, Jan. 2021, doi: 10.1109/TNNLS.2020.2978386.
10. G. Pons, B. Bilalli, and A. Queralt, "Knowledge Graphs for Enhancing Large Language Models in Entity Disambiguation," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) , vol. 15231 LNCS, pp. 162–179, 2025, doi: 10.1007/978-3-031-77844-5_9.
11. Z. Zhong, A. Barkova, and D. Mottin, "Knowledge-augmented Graph Machine Learning for Drug Discovery: A Survey," ACM Comput Surv, vol. 57, no. 12, Jul. 2025, doi: 10.1145/3744237.
12. S. Pakhomov, G. Finley, R. McEwan, Y. Wang, and G. Melton, "Corpus domain effects on distributional semantic modeling of medical terms," Bioinformatics, vol. 32, no. 23, pp. 3635–3644, 2016.
13. J. Wu, X. Xu, Y. Zhang, and F. Xu, "Clinical abbreviation disambiguation using neural word embeddings," Journal of Biomedical Informatics, vol. 73, pp. 96–106, 2017.
14. A. Névéol, C. Grouin, J. Leixa, A. Rosset, and P. Zweigenbaum, "Clinical information extraction at the CLEF eHealth evaluation lab," Journal of Biomedical Informatics, vol. 100, pp. 103–117, 2019.

15. D. Moher *et al.*, “Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,” *PLoS Medicine*, 2009.
16. B. Kitchenham and S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” EBSE Technical Report, 2007.
17. J. Higgins et al., Cochrane Handbook for Systematic Reviews of Interventions, Wiley, 2019.
18. B. Kitchenham et al., “Systematic literature reviews in software engineering – A tertiary study,” Information and Software Technology, 2010.
19. S. Névéol et al., “Clinical natural language processing in languages other than English,” Journal of Biomedical Informatics, 2018.
20. R. Islamaj Dogan et al., “Biomedical natural language processing: A survey,” Briefings in Bioinformatics, 2014.
21. A. Xu et al., “A semantic approach to clinical abbreviation disambiguation,” Journal of Biomedical Informatics, 2007.
22. T. Liu et al., “Disambiguating clinical abbreviations with contextual information,” AMIA Symposium Proceedings, 2001.
23. S. Pakhomov et al., “Automated abbreviation disambiguation in clinical text,” AMIA Annual Symposium, 2002.
24. W. Wu et al., “Supervised learning for clinical abbreviation expansion,” Bioinformatics, 2015.
25. J. Moon et al., “Machine learning-based abbreviation disambiguation in clinical narratives,” BMC Medical Informatics and Decision Making, 2014.
26. S. Joshi et al., “Feature-based methods for medical acronym disambiguation,” Artificial Intelligence in Medicine, 2016.
27. J. Li et al., “Neural network models for medical abbreviation disambiguation,” IEEE Journal of Biomedical and Health Informatics, 2018.
28. J. Lee et al., “BioBERT: A pre-trained biomedical language representation model,” Bioinformatics, 2020.
29. E. Alsentzer et al., “Publicly available clinical BERT embeddings,” NAACL, 2019.
30. A. Holzinger et al., “Explainable AI in medicine,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017.
31. M. Dehghan et al., “Graph-based methods for clinical concept disambiguation,” Journal of Biomedical Informatics, 2020.
32. Y. Zhang et al., “Graph neural networks for biomedical text mining,” Briefings in Bioinformatics, 2021.
33. H. Peng et al., “Hybrid neural and graph-based models for clinical text understanding,” Artificial Intelligence in Medicine, 2022.
34. Ö. Uzuner et al., “Evaluating the state of the art in automatic de-identification,” Journal of the American Medical Informatics Association, 2011.
35. S. Pakhomov et al., “The clinical abbreviation sense inventory,” AMIA Annual Symposium Proceedings, 2006.
36. A. Johnson et al., “MIMIC-III, a freely accessible critical care database,” Scientific Data, 2016.
37. A. Johnson et al., “MIMIC-IV: A freely accessible electronic health record dataset,” Scientific Data, 2023.

38. J. Wu et al., “Abbreviation disambiguation in clinical text using domain-specific corpora,” BMC Medical Informatics and Decision Making, 2019.
39. S. Moon et al., “Variability of clinical abbreviation usage across institutions,” Journal of Biomedical Informatics, 2014.
40. T. Saito and M. Rehmsmeier, “The precision–recall plot is more informative than the ROC plot,” PLoS One, 2015.
41. C. Manning et al., Introduction to Information Retrieval, Cambridge University Press, 2008.
42. J. Demner-Fushman et al., “Evaluating clinical NLP systems,” Journal of Biomedical Informatics, 2009.
43. Y. Wu et al., “Efficiency-aware deep learning for clinical NLP,” Artificial Intelligence in Medicine, 2021.
44. A. Névéol et al., “Reproducibility in clinical NLP,” Briefings in Bioinformatics, 2020.
45. J. Wu et al., “Cross-domain challenges in clinical NLP,” Journal of Biomedical Informatics, 2020.
46. S. Ratner et al., “Data programming: Creating large training sets,” NeurIPS, 2016.
47. J. Demner-Fushman et al., “Clinical NLP: Current challenges,” Journal of Biomedical Informatics, 2017.
48. Y. Tay et al., “Efficient deep learning for healthcare,” IEEE JBHI, 2021.
49. A. Holzinger et al., “What do we need to build explainable AI systems for the medical domain?” arXiv, 2019.
50. A. Névéol et al., “Reproducibility and reuse in clinical NLP,” Briefings in Bioinformatics, 2020.