Journal of Biomedical Informatics 178 (2026) 105005

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

A comprehensive survey on medical concept normalization: Datasets, techniques, applications, and future directions☆

Haihua Chen [a,*,1], Yuhan Zhou [b,1], Ruochi Li [c,1], Aryan Murthy Illa [b], Ana Cleveland [b], Junhua Ding [a]

[a] Anuradha and Vikas Sinha Department of Data Science, University of North Texas, Denton, 76203, TX, USA
[b] Department of Information Science, University of North Texas, Denton, 76203, TX, USA
[c] Department of Computer Science, North Carolina State University, Raleigh, 27695, NC, USA

ARTICLE INFO

Keywords:
Medical concept normalization
Biomedical entity linking
Machine learning
Deep learning
Transfer learning
Large language model

ABSTRACT

Medical concept normalization (MCN) involves mapping informal medical terms or phrases to standardized medical concepts, serving a crucial role in medical text analysis, healthcare intelligence, and other applications. Previous studies have primarily focused on developing datasets, refining machine learning algorithms, and applying them to downstream tasks, but a comprehensive survey covering all key aspects of MCN has been lacking. This survey addresses this gap by providing a complete overview of MCN, including task definitions, annotation schemes, datasets, normalization techniques, state-of-the-art performance, and applications. A comparative analysis of benchmark datasets and leading normalization methods is conducted, along with a quantitative evaluation of CNN-based, RNN-based, transformer-based, GNN-based, LLM-based, and other models on popular benchmarks. Guidelines are also provided for selecting the most effective models and techniques. In addition, the applications of MCN in electronic health records (EHRs) management, clinical decision support (CDS), precision medicine, health information exchange (HIE), and others are explored. Finally, key challenges and potential directions for future research are discussed. This survey offers the first comprehensive review of the core components of MCN, serving as a valuable resource for both researchers and practitioners. Resources related to this survey can be accessed on GitHub at: https://github.com/haihua0913/awesome-mcn.

Contents

☆ This research is partially supported by The United States National Science Foundation (NSF) grants #2231519, #2244259 and #2225229.
* Corresponding author.
E-mail addresses: haihua.chen@unt.edu (H. Chen), yuhanzhou@my.unt.edu (Y. Zhou), rli14@ncsu.edu (R. Li), AryanMurthyIlla@my.unt.edu (A.M. Illa), ana.cleveland@unt.edu (A. Cleveland), junhua.ding@unt.edu (J. Ding).
[1] Equal contribution.

https://doi.org/10.1016/j.jbi.2026.105005
Received 19 October 2025; Received in revised form 18 January 2026; Accepted 25 February 2026
Available online 2 March 2026

1. Introduction

Medical concept normalization (MCN), also commonly termed medical entity normalization (MEN), or biomedical named entity normalization (BNEN), or biomedical entity linking (BEL), involves the process of mapping informal medical terms or phrases from social media, or biomedical documents, or other online platforms to formal medical concepts in a controlled vocabulary, ontology, or standardized medical database [1–4]. Examples of MCN are illustrated in Fig. 1, and show how an informal phrase can be mapped to multiple medical concepts and multiple informal phrases can also be mapped to the same medical concept. In addition, standardized medical concept databases (referred to as "annotation schemes" in this paper), such as the Unified Medical Language System (UMLS) [5], the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) [6], or Medical Subject

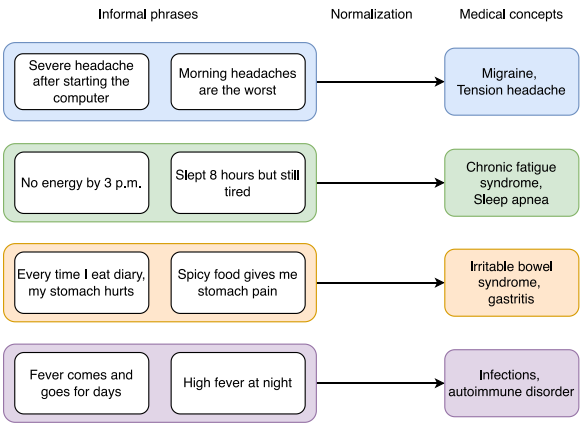

Fig. 1. MCN examples. Texts on the left are informal phrases, the arrows indicate MCN tasks, and the terms on the right are the corresponding medical concepts of the informal phrases. Each color represents an MCN task. These example mappings are not intended to imply definitive clinical associations and require additional contextual information in practice.

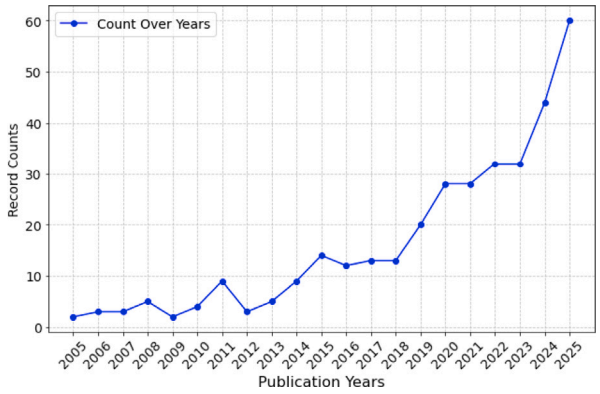

Fig. 2. Paper counts over years (2005–2025). Retrieved from Web of Science Core Collection, using the query "medical concept normalization" OR "medical entity normalization" OR "biomedical named entity normalization" OR "medical entity linking" OR "biomedical entity linking" to search the title, abstract, keyword plus, and author keywords fields. Data retrieved on December 31, 2025. The number of papers related to MCN in 2025 has been continuously growing.

Headings (MeSH) [7], among others [8,9], usually contain hundreds of medical concepts. However, many formal medical concepts lack corresponding informal phrases on social media, leading to MCN as a challenging long-tailed extreme multi-label text classification (XML) task [10,11].

MCN is a fundamental natural language processing (NLP) task in the medical domain [12,13]. It is critical for improving the efficiency, accuracy, and effectiveness of various healthcare applications, including electronic health record (EHR) management, clinical decision support (CDS) systems, health information exchange (HIE), precision medicine (PM), drug discovery and development, clinical trial retrieval, and automated patient messaging systems. MCN also plays an important role in broader clinical text mining and pharmacovigilance settings, where extracted mentions such as adverse drug reactions (ADR) and related clinical entities need to be grounded to standardized medical terminologies for reliable aggregation and downstream analysis. Ultimately, MCN has a direct impact on both patient care and healthcare administration [14].

Over the past two decades, there has been a growing focus within the research community on MCN, especially after the transformer architectures [15] were introduced in 2017 and deep pre-trained models (PTMs), such as GPT [16] and BERT [17], were proposed for NLP tasks in 2018, as illustrated in Fig. 2. We collected all survey papers on MCN over the past decade and conducted a comparative analysis on the time frame, survey paper types, number of articles included, scope, and whether annotation schemes, datasets, data quality, algorithms, experimental results, and applications were discussed in these surveys, as presented in Table 1.

Previous surveys primarily emphasized different components of MCN and only investigated a small portion of the literature. For example, although French and McInnes conducted a systematic review on BEL related to history, applications, datasets, performance, and shared tasks based on the literature from 1980 to August 2022, only 60 papers were included [4]. Shi et al. provided a comprehensive review on rule-based, machine learning, and deep learning models for BEL with a special focus on knowledge graph (KG)-based methods [18]; however, only four types of KG-based methods and six datasets were investigated. The two latest surveys encountered the same issues: (1) only a few articles (less than 20%) from the last ten years were included, a large number of papers using deep learning-based, transformer-based, and large language model-based techniques were missing in the investigation; and (2) detailed introductions and comparisons on the datasets and techniques were lacking. Kartchner et al. performed a comprehensive quantitative evaluation of nine BEL models across seven datasets, aiming to identify the characteristics, reproducibility, and usability of different models [19]. Other earlier surveys either focused on datasets [21], or targeted entity normalization (or recognition) in the general domain with little discussion on medical concept normalization [3,20,22]. Furthermore, the absence of an explicit framework requires future research in this area. With the growing interest and work on this topic (see Fig. 2), it is time for a survey such as our paper, which aims to:

- define the research framework for MCN tasks and introduce annotation schemes for MCN dataset construction;
- sort out popular and useful MCN datasets and investigate the quality of these datasets;
- depict the history of techniques for MCN over time and summarize the performance of competitive models on benchmark datasets to compare the advantages and disadvantages of different methods;
- outline the applications of MCN in various contexts and downstream tasks; and
- identify key challenges and future directions to motivate and orient interests in this area effectively.

To summarize, this paper is the first comprehensive survey on medical concept normalization based on more than 219 articles from 2016 to 2025. It covers all key stages and components of MCN, including task definition, annotation schemes, datasets, techniques (and toolkits), applications, challenges, and future directions.

Statement of Significance	
Problem	Medical concept normalization (MCN) maps informal medical expressions in clinical text, biomedical literature, and social media to standardized medical concepts. Despite extensive research, the literature is scattered across datasets, methods, and applications, making it difficult to obtain a comprehensive view of progress in this field.

Table 1
Surveys related to MCN within the last ten years. The **Survey** column lists the citations of the previous surveys. **Year** is the year of publication. **Time frame** is the period that the survey covers. **#IA** indicates the total number of articles included in the survey. **Scope** describes the scope or area of MCN in the survey. **AS** represents annotation schemes or medical concept categories. **DS** indicates datasets. **DQ** indicates a summary of the dataset quality. **AL** refers to algorithms or techniques for MCN. **ER** represents a summary of the experimental results. **AP** is a summary of the application. ✓ indicates being included while ✗ indicates not being included.

Survey	Year	Time frame	Type	#IA	Scope	AS	DS	DQ	AL	ER	AP
[4]	2023	1980–2022	Systematic	60	Overview of BEL	✗	✓	✗	✗	✓	✓
[18]	2023	Unknown	Comprehensive	–	KG for BEL	✗	✓	✗	✓	✓	✗
[19]	2023	2019–2022	General	–	Comparative evaluation on 9 models over 7 datasets	✗	✓	✗	✓	✓	✗
[3]	2022	2015–2021	Systematic	55	General domain neural EL models	✗	✗	✗	✓	✓	✓
[20]	2021	2018–2021	Systematic	23	Clinical NER and RE	✗	✓	✗	✓	✓	✗
[21]	2020	Unknown	General	–	BNER & BNEN datasets	✗	✓	✓	✗	✗	✗
[22]	2016	Before 2015	General	–	Temporal concept normalization in clinical domain	✓	✗	✗	✓	✗	✗
Ours	2025	2016–2025	Comprehensive	219	Covers all aspects and phases of MCN	✓	✓	✓	✓	✓	✓

Notes: Paper [3] only focused on neural network-based models for entity linking in the general domain, which might be different from the medical domain. Paper [20] mainly focused on techniques for named entity recognition (NER) and relationship extraction (RE) but mentioned little about medical entity linking (MEL). Paper [21] specifically focused on datasets for BNER & BNEN; however, only ten datasets related to BNEN were included. Paper [22] focused on temporal data representation, normalization, extraction, and reasoning in the clinical domain and only used two paragraphs to discuss the normalization of time and time-related concepts in clinical texts.

What is already known	Prior studies have introduced datasets, machine learning and deep learning models, and applications for MCN. Existing reviews typically focus on limited aspects such as specific datasets, algorithms, or time periods, leaving a lack of systematic understanding of datasets, data quality issues, algorithmic development, and real-world applications.
What this paper adds	This paper provides a comprehensive survey of MCN research from 2016–2025. It proposes a unified analytical framework, reviews datasets and data quality issues, summarizes the evolution of algorithms including neural and large language model approaches, analyzes benchmark performance, surveys open-source toolkits, and discusses practical applications and future research directions.
Who would benefit from the new knowledge	Researchers in biomedical NLP, clinical text mining, and health informatics can use this study as a consolidated reference for MCN datasets, algorithms, and evaluation strategies. Practitioners developing healthcare information systems can also leverage these insights to design more effective normalization systems and identify future research opportunities.

2. Methodology

2.1. Data collection

Search strategy and information sources. We conducted a systematic literature search in PubMed Central, Web of Science, and Scopus using the query: (“medical concept normalization” OR “medical entity normalization” OR “biomedical named entity normalization” OR “medical entity linking” OR “biomedical entity linking”). Records from all databases were exported and merged. Duplicates were removed using a hierarchical matching strategy prioritizing DOI when available, and otherwise normalized title and publication year.

Eligibility criteria and screening. We included English-language peer-reviewed research articles and regular conference papers focused on medical/biomedical concept or entity normalization and entity linking. We excluded non-research publication types such as review articles, editorials, tutorials, workshop papers, posters, and demos. Title and abstract-based relevance screening was performed using the inclusion topic terms above and close variants.

Identification from other sources. In addition to database searches, we conducted backward and forward snowballing to identify additional eligible studies. Studies not previously identified were included and flagged as records identified from other sources through snowballing.

A total of 219 studies were included in the final analysis, including 117 records from database search and 102 records from snowballing. The complete study identification, screening, eligibility assessment, and inclusion process is illustrated in the PRISMA flow diagram in Fig. 3.

2.2. Research framework for medical concept normalization

As introduced above, medical concept normalization (MCN) is the task of mapping informal mentions in text to medical concepts in a given standardized medical database [4,18]. The informal mentions are typically sourced from X (formerly Twitter), Reddit, AskaPatient, the biomedical literature, or electronic health records, while the target medical databases include the Unified Medical Language System (UMLS) [5], the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [6], or Medical Subject Headings (MeSH) [7], among others [8,9]. MCN contains several major components: (1) task definitions, where the mapping between informal phrases and standardized medical concepts needs to be defined; (2) dataset construction, where a set of informal phase-medical concept pairs needs to be labeled based on the task definition; (3) model development and evaluation, where normalization systems can be built either through supervised training and fine tuning of machine learning or deep learning models, or through training free inference using large language models in zero shot or few shot settings (for example, via in context learning with demonstration examples); in both cases, metrics such as accuracy, precision, recall, or F1 score are used for evaluation and (4) applications, where MCN is implemented for different applications, such as clinical decision support, precision medicine, automated patient messaging, and so on. The workflow is demonstrated in Fig. 4.

The MCN task can be defined (adapted from [18]) as follows:

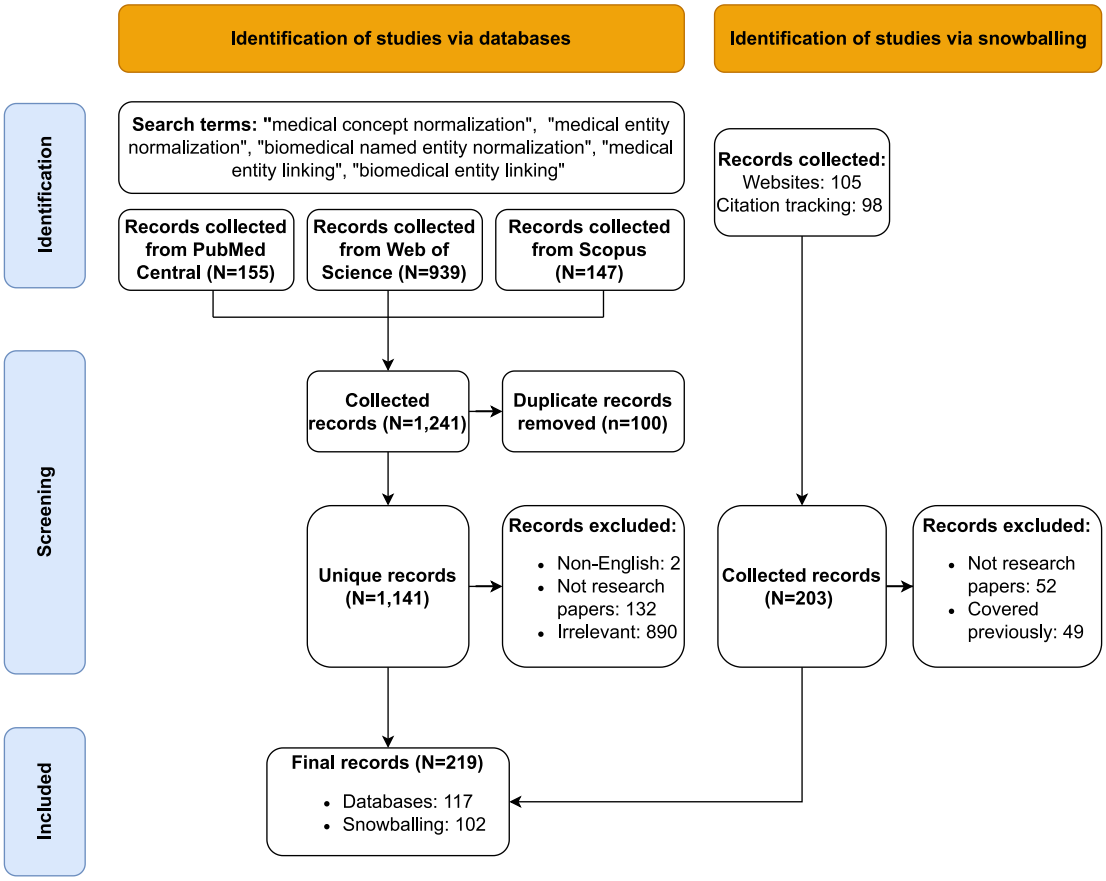

Fig. 3. PRISMA flow diagram (Data collected from January 2016 to December 2025).

Definition 2.1 (*Medical Concept Normalization*)**.** Given a standardized medical database in the medical or biomedical field consisting of N concepts $\mathrm{C} = \{c_1, c_2, c_3, \ldots, c_N\}$, a social media text or electronic health record or biomedical text document T contains a set of recognized informal phrases or entity mentions $\mathrm{M} = \{m_1, m_2, m_3, \ldots, m_M\}$, and the task is to find the concept $c_i \in C$ to which $m_j \in M$ refers.

2.3. Survey structure

MCN is usually considered as a text classification task. A fundamental step is to create a dataset by annotating the informal phrases with medical concepts. Therefore, we investigate the most popular standardized medical databases (referred to as "annotation schemes" in this survey) for data annotation. Subsequently, we conduct a comprehensive introduction to well-established datasets across different sources, languages, and years in Section 3. We also summarize and discuss their data quality which has been revealed in the existing literature.

Over the last decade, various machine learning techniques have been proposed for MCN and deep learning models have proven more effective than traditional machine learning models. Therefore, this survey focuses on deep learning techniques for MCN. Section 4 investigates CNN-based, RNN-based, transformer-based, GNN-based, LLM-based, and other widely adopted models.

In Section 5, we then analyze the architectures, advantages, and disadvantages, and conduct a comprehensive investigation on their performance across multiple benchmarks. We also present the open source toolkits for MCN in Section 6.

Based on the above investigations, we outline several important applications in Section 7, and identify key challenges and future directions in Section 8.

3. Overview of the datasets for medical concept normalization

3.1. Annotation schemes for MCN dataset construction

Investigating annotation schemes or ontologies for medical concepts is essential for categorizing MCN datasets by their specific vocabularies. Medical ontologies provide standardized representations of clinical, biological, and pharmacological knowledge, enabling interoperability, data integration, and reproducible dataset construction. These ontologies vary in scope, granularity, purpose, and language. In the following, we review the ontologies that are most commonly employed in the construction and annotation of existing MCN datasets.

The UMLS (Unified Medical Language System),[2] one of the most extensive medical Metathesaurus, integrates controlled vocabularies and classifications. It covers an extensive list of around 13 million distinct normalized concept names for its 3,454,294 concepts, and over 12 million relations among them [23]. In the latest 2025 version, it has 29 languages contributing concept names, with English taking up 61.2%.[3]

SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms)[4] is a comprehensive and multilingual clinical healthcare terminology database [24]. It contains over 357,000 unique healthcare concepts with formal logic-based definitions. Other than the English and Spanish versions managed by SNOMED International, 12 translated versions are also available of varying sizes and scopes. Access to these

[2] https://www.nlm.nih.gov/research/umls/index.html
[3] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html
[4] https://www.nlm.nih.gov/healthit/snomedct/index.html

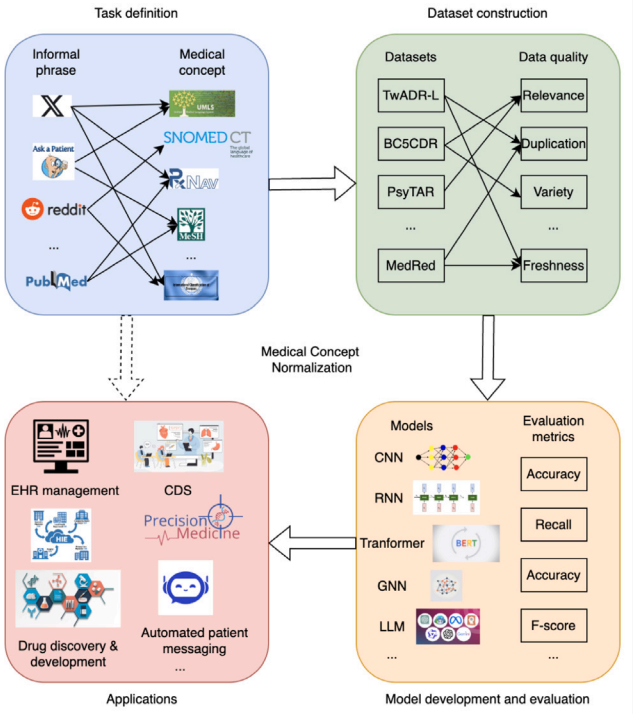

Fig. 4. The workflow and main components of Medical Concept Normalization. We list only some examples of informal phrases and medical concepts in the task definition, datasets, and data quality in dataset construction. This shows mapping in task definition and dataset construction only for demonstration purposes instead of mapping in real-world scenarios.

translations can be retrieved through the relevant National Release Center.

RxNorm,[5] a standardized nomenclature for clinical drugs and devices, has been instrumental in drug utilization analysis [25]. Its integration with UMLS and SNOMED-CT enables comprehensive drug-disease associations [9,26,27]. It serves as a national standard for coding clinical drugs in the U.S.

MedDRA (Medical Dictionary for Regulatory Activities)[6] is used primarily in the biopharmaceutical and regulatory industries for registration, documentation, and safety monitoring of medical products. MedDRA covers terms across five hierarchical levels, from broad categories to granular details. With over 80,000 terms, MedDRA enables detailed categorization of adverse events, symptoms, diagnoses, and medical history, making it essential for constructing high-quality annotated datasets [28,29]. MedDRA is now accessible in 24 languages in its latest release in early 2025.

MeSH (Medical Subject Headings)[7] is a comprehensive controlled vocabulary to facilitate the indexing, cataloging, and searching of biomedical and health-related information. It contains over 30,000 descriptors arranged in a hierarchical tree structure, covering medical domains, such as diseases, anatomical structures, chemicals, procedures, and biological processes. MeSH descriptors are enriched with unique identifiers, entry terms, and supplementary concept records for specific entities like drugs or chemicals. It has been translated into other languages [30].

OMIM (Online Mendelian Inheritance in Man)[8] is a comprehensive, authoritative database focused on the relationships between human genes and genetic disorders. OMIM contains detailed entries on over 27,825 genes and phenotypes, encompassing monogenic, polygenic, and complex traits. Each entry provides curated information on gene function, inheritance patterns, associated diseases, clinical features, and molecular mechanisms. OMIM's unique six-digit identifier for disorders and genes enables precise annotation and cross-referencing. For the more recent versions, it is available in English.

ICD (International Classification of Diseases)[9] is a globally recognized system for standardizing the classification and coding of diseases, injuries, and other health-related conditions. Currently, the 11th revision (ICD-11) contains 17,000 diagnostic categories, with over 100,000 medical diagnostic index terms, organized hierarchically to support various levels of granularity. It supports ten languages now and will expand its linguistic reach in support of its global mission.[10]

Wikipedia[11] serves as a valuable resource in MCN tasks due to its extensive and diverse knowledge base. It provides a vast repository of structured and semi-structured information on diseases, symptoms, drugs, anatomical terms, and procedures [31]. Each medical concept on Wikipedia is often associated with a unique page, containing definitions, synonyms, relationships, and contextual details with hyperlinks and redirects. Additionally, its multilingual nature supports cross-lingual normalization [32,33].

The selection of annotation schemes needs to align with language considerations, domain requirements, and task context. For example, RxNorm is suitable for U.S. drugs, and OMIM is specialized for genes and genetic disorders. While ICD, MedDRA, and SNOMED-CT can be adopted in more general and common use cases, they also offer stronger multilingual support. Among all, UMLS links over 200 source vocabularies across 29 languages, including ICD, SNOMED-CT, and MeSH. Therefore, it could serve as a comprehensive and integrated controlled concept database that suits a wide range of MCN tasks.

3.2. Existing MCN datasets

MCN tasks rely on robust data collection from diverse sources, including social media platforms, the medical literature, and clinical records, providing broad coverage of medical terminology and its usage across various contexts. , Table 3, and Table 4 present MCN datasets collected from three corpora types: biomedical, clinical, and social media, respectively. The frequently used datasets since 2016 are introduced below:

TwADR-L dataset focuses on Twitter[12] messages related to medication and negative drug reactions (NDR). It comprises 1436 Twitter phrases mapped to one or more of the 2220 medical concepts [34]. This mapping process results in a dataset containing 50,730 records, with each record including both the informal language and its corresponding normalized medical concept.

AskaPatient [34] is derived from blog posts on the AskaPatient website,[13] which contains patient-reported experiences with prescription medications. By extracting non-clinical medical terms from social media and mapping them to clinical ontologies like SNOMED-CT, AskaPatient offers valuable research resources.

The BC5CDR dataset [31] includes chemical-disease relationships from 1500 PubMed abstracts. It includes 4409 chemicals, 5818 diseases, and 3116 chemical-disease interactions, making it valuable for training and evaluating machine learning models in biomedical relation extraction [35].

MIMIC-III [49] is a large, publicly accessible dataset of de-identified electronic health records (EHRs) from over 40,000 critically ill patients

[5] https://www.nlm.nih.gov/research/umls/rxnorm/index.html
[6] https://ich.org/page/meddra
[7] https://www.nlm.nih.gov/mesh/meshhome.html
[8] https://www.omim.org/
[9] https://icd.who.int/en
[10] https://www.who.int/news/item/08-02-2024-icd-11-2024-release
[11] https://www.wikipedia.org/
[12] https://x.com/
[13] https://www.askapatient.com/

Table 2
Medical Concept Normalization datasets: Biomedical corpora.

Dataset	Year	Language(s)	# Phrases	Types of Phrases	# Concepts	Sources of Concepts
NCBI Disease [36]	2014	English	793	PubMed abstracts	6892	MeSH, OMIM
QUAERO [37]	2014	French	26,409	Biomedical research papers	10	UMLS
Mantra GSC [38]	2015	English, German, French, Spanish, and Dutch	1000	Journal titles, drug labels, patents	5530	UMLS, SNOMED-CT, MedDRA
BC5CDR [31]	2016	English	1500	PubMed articles	13,343	MeSH, UMLS, Wikipedia
TAC 2017 [39]	2017	English	200	Drug-safety labels	13,501	MedDRA
SPL-ADR-200db [40]	2018	English	200	DailyMed drug labels	5098	UMLS, MedDRA
MedMentions [41]	2019	English	4392	PubMed abstracts	352,496	UMLS
DrugProt [42]	2021	English	3500	PubMed abstracts	89,529	Expert annotations
NEREL-BIO [43]	2023	Russian	756	PubMed abstracts	4544	UMLS
WikiMed-DE [32]	2023	German	53,981	Wikipedia articles	35,012	UMLS, MeSH, DOID
BioWiC [44]	2024	English	20,156	Aggregated biomedical corpora	7413	UMLS, MeSH, OMIM
WALVIS [33]	2024	Dutch	2307	Wikipedia sentences	1086	UMLS

Table 3
Medical Concept Normalization datasets: Clinical corpora.

Dataset	Year	Language	# Phrases	Types of Phrases	# Concepts	Sources of Concepts
i2b2/VA [45]	2010	English	826	Patient health records	72,846	RPDR [46]
ShARe/CLEF [47]	2014	English	298	Clinical notes	4626	UMLS
SemEval-2015 T14 [48]	2015	English	531	Clinical documents from ShARe	2500	UMLS, SNOMED-CT
MIMIC-III [49]	2016	English	112,000	Clinical reports	1159	ICD-9
MADE [50]	2019	English	1089	Cancer clinical notes	43,000	MedDRA, SNOMED-CT
nc2c 2019 [29]	2019	English	100	Discharge summaries	10,919	SNOMED-CT, RxNorm
Cantemist [51]	2020	Spanish	1301	Oncological case reports	16,030	ICD-10
CMeIE [52]	2020	Chinese	28,008	Medical textbooks and clinical practices	11	MeSH, ICD-10, ATC [53]
CHIP-CDN [54]	2021	Chinese	18,192	Clinical diagnoses	9587	ICD-10
nc2c 2022 [55]	2022	English	500	Clinical notes	9013	Medication mentions
RuCCoN [56]	2022	Russian	16,028	Medical histories	2409	UMLS
SemClinBr [57]	2022	Portuguese	1000	Clinical notes	65,117	UMLS
FRASIMED [58]	2023	French	2051	Synthetic clinical cases	24,037	ICD-O, SNOMED-CT
MedNERF [59]	2023	French	100	Typewritten prescriptions	406	n2c2 [60]
DisTEMIST-linking [61]	2023	Spanish	2800	Clinical cases	7303	SNOMED-CT
MIMIC-IV [62]	2024	English	204	Clinical reports	75,000	ICD-9, ICD-10
SympTEMIST [63]	2024	Spanish	1000	Clinical reports	12,196	SNOMED-CT
TCMNER [64]	2024	Chinese	9036	Traditional Chinese medicine diagnoses	5	Expert annotations
RuCCoD [65]	2025	Russian	865,539	Patient health records	3546	ICD-10

(2001–2012) who were at Beth Israel Deaconess Medical Center.

The SMM4H-2017 dataset [66] consists of health-related information extracted from tweets, including data on diseases, symptoms, treatments, and sentiment, with annotations derived from social media posts.

PsyTAR [67], the Psychiatric Treatment Adverse Reactions dataset, contains patient reviews about the efficacy and side effects of psychotropic drugs. The initial phase includes 891 drug reviews from visit.askapatient.com, each with demographic information, treatment duration, and detailed patient assessments. The second phase classifies sentences into 6009 labels across categories like Adverse Drug Reaction (ADR), Withdrawal Symptoms (WDs), and others.

MADE corpus [50] contains 1089 EHR notes, with 876 for training and 213 for testing, annotated for medication, indication, and adverse drug events. These annotations map to MedDRA and SNOMED-CT terms from the UMLS Metathesaurus, with the training set containing 35,000 mentions and the test set 8000 mentions.

nc2c 2019 track 3 dataset [29] is instrumental in clinical text normalization. It was developed for the Fourth i2b2/VA Shared Task [45] and includes 100 discharge summaries containing 3792 concepts with a total of 10,919 mentions from two controlled vocabularies.

MedMentions [41] is a dataset designed for named entity recognition (NER) in biomedical scientific papers. It includes titles and abstracts from 4392 papers published on PubMed in 2016, covering a range of medical entities, such as diseases, chemicals, and genes. This dataset is commonly used for training and evaluating machine learning models and NLP systems in biomedical contexts [68].

MedRed dataset [69] analyzed 1980 Reddit[14] posts from 18 subreddits, selecting 110 posts from each. It uses expert-annotated CADEC data [70] for quality control.

3.3. Observations

Although there are many available MCN datasets, their data quality (DQ) remains a concern [14,26,80]. Data quality dimensions in machine learning include comprehensiveness, consistency, duplication, and other issues [81]. Some studies have adopted a subset of these to evaluate MCN datasets. The commonly used dimensions are introduced as follows, and Table 5 presents the investigation details provided by both the data curation description and experimental results in the existing literature.

Duplication refers to the same phrases being repeatedly mapped to identical medical concepts [1,14,82]. Duplication issues can exist in the training corpus or across the training, test, and validation datasets. Largely overlapped training and test data could cause the inflation of the normalization performance [1,14,80,82]. Each validation dataset and the test dataset should contain a significant amount of new data compared to the corresponding training dataset [81].

Comprehensiveness means each phrase is supposed to be mapped to one or more medical concepts. However, only 273 out of 2220 concepts in the TwADR-L dataset are linked to informal phrases. Therefore, a significant portion of the medical concepts lack informal expressions, highlighting a data quality challenge of breadth of coverage [14]. This data quality issue is noted in [26].

[14] https://www.reddit.com/

Table 4
Medical Concept Normalization datasets: Social media corpora.

Dataset	Year	Language(s)	# Phrases	Types of Phrases	# Concepts	Sources of Concepts
Cadec [70]	2015	English	6754	AskaPatient posts	1029	SNOMED-CT, MedDRA
TwADR-S [34]	2016	English	201	Twitter posts	58	SNOMED-CT
TwADR-L [34]	2016	English	1436	Twitter posts	2220	SIDER
AskaPatient [34]	2016	English	17,324	ADR blog posts	1036	SNOMED-CT
SMM4H-2017 [66]	2017	English	9000	Twitter posts	726	MedDRA
TwiMed [71]	2017	English and Spanish	2000	Twitter, PubMed sentences	3144	SNOMED-CT
SMM4H-2019 [72]	2019	English	12,800	Twitter posts (ADR mentions)	1548	MedDRA
PsyTAR [67]	2019	English	6556	Psychiatric drug reviews	618	UMLS, SNOMED-CT
HMC2019 [73]	2019	English	15,393	Twitter posts	10	Disease names
MedNorm [74]	2019	English	23,292	Aggregated social media corpora	3607	MedDRA, SNOMED-CT
COMETA [75]	2020	English	800,000	Reddit posts	20,015	SNOMED-CT
MedRed [69]	2020	English	1980	Reddit posts	3511	CADEC [70]
SMM4H-2021 [76]	2021	English	18,833	Twitter posts (ADE mentions)	2237	MedDRA
RHMD [77]	2022	English	10,015	Reddit posts	15	Disease names
SocialDisNER [78]	2022	Spanish	7500	Twitter posts	19,425	MedDRA
NHMD [79]	2023	English and Nigerian	7763	Nairaland Forum	4	Disease names

Table 5
Data quality evaluation of MCN datasets.

Dataset	Data Quality Evaluation
TwADR-L [34]	(1) For comprehensiveness, 1947 out of 2220 concepts are irrelevant to Twitter phrases [34,80]. (2) For duplication, 8.62% of the test set overlaps with training in the same fold [1]. (3) For class imbalance, many concepts have only one example [1].
AskaPatient [34]	On average, 35.82% of the test data overlap with training in the same fold [1].
ShARe/CLEF [47] SemEval-2015 T14 [48]	Institution-specific corpus limits the generalizability of models trained on these data [60].
n2c2 2019 [29]	(1) 2.7% of mentions could not be mapped to any concept; mappings sometimes inconsistent across SNOMED-CT hierarchies [29]. (2) Ranking of candidate concepts biases annotator choices toward top results [29]. (3) 505 prediction errors due to ambiguous (e.g., misspellings, unknown semantic types) or complex mentions [26]. (4) Single-word mentions comprise 55% of the test set and are hardest to normalize [26]. (5) Provides full clinical notes for richer context compared to other social media-derived corpora [27].
BC5CDR [31]	718 out of 1337 disease mentions in the test set never appear in training or development sets [31].
SPL-ADR-200db [40]	83 ADRs remained under-specified; 19 received alternate mappings by editors [40].
MedMentions [41]	(1) 42% of test concepts do not occur in training; 38% occur in neither training nor development [41]. (2) MedMentions adopts a subset of UMLS for annotation [41,44].
MedNorm [74]	40% of concepts are under-reported with only one instance [74].
COMETA [75]	Each concept has at least five example sentences on average (median of 3) [75].
BioWiC [44]	It covers almost 80% of UMLS semantic types (99 out of 127) across splits [44].
MIMIC-IV [62]	2162 out of 5336 concepts appear only once in the annotated data [85].
WALVIS [33]	Manual checking found 28% of 100 sampled mentions linked to related but different concepts [33].

Class imbalance refers to a substantial disparity in the number of samples between different classes and impacts the model accuracy predicting underrepresented diseases and groups [65,83]. For example, MCN datasets that lack demographic diversity can perpetuate biases in downstream applications like clinical decision support [84].

The mentioned data quality challenges in MCN underscore critical limitations and future research directions as follows.

1. **The development of a DQ evaluation framework.** First, while prior work has emphasized algorithmic innovation, systematic evaluation of DQ dimensions remains understudied [14]. Future efforts should involve the development of a conceptual or methodological framework to evaluate these dimensions across a broader range of datasets or models. Such frameworks should also quantify the impacts of data quality on model performance, as demonstrated by the techniques proposed in Section 4.
2. **Domain-specific evaluation strategy.** MCN datasets are developed for specific domains and departments, such as social media content or cancer notes in clinical practices. DQ frameworks and dimensions cannot fit all and need to be adapted to balance between generalizability and sensitivity based on different applications [26].
3. **Data quality of multilingual resources.** English corpora currently dominate MCN research. However, neglecting non-English data risks models that underdiagnose or misinterpret conditions in vast populations, perpetuating information gaps and limiting real-world deployability. Fusing multilingual resources faces cross-lingual inconsistencies. Annotation guidelines diverge across languages, ontologies differ in conceptual granularity, and literal term translations often lose language-specific nuance [43,58,86,87]. Low-resource language datasets need further evaluation in the above aspects.
4. **Data quality of multimodal resources.** Multimodal data in patient records could offer more information, yet suffer from low quality issues, such as noise, incomplete modality, and imbalance [88,89]. Normalizing multi-modal data in healthcare also requires further research [90].
5. **Leveraging LLMs for DQ evaluation and improvement.** As manual quality checks are impractical for large-scale or dynamically updated datasets, automated pipelines and leveraging LLMs could help with faster and more cost-effective quality control approaches, such as augmenting MCN data with high-quality phrase-concept pairs [14].

Table 6
Summary of major method architectures for medical concept normalization (MCN).

Architecture	Core modeling idea	Advantages	Weaknesses	Typical datasets or scenarios
CNN methods	Convolution and pooling over character or word embeddings to capture loca semantics and morphology for mention encoding and mention candidate ranking	Robust to noisy and informal mentions, effective at capturing subword morphology and orthographic variants, computationally efficient with high parallelism, and well suited to feature based ranking in large candidate spaces	Limited capacity for long range contextual disambiguation and global coherence, pooling may blur fine grained contextual cues, and performance can depend on candidate generation quality, hence CNN based pipelines are often augmented with richer context or sequence level constraints	Noisy user generated text, ADR and symptom normalization, drug label pipelines, multilingual and cross lingual settings with character level variation, Chinese clinical narratives with dictionary features.
RNN methods	Gated recurrent sequence encoding of mention context, optionally with attention, for context sensitive mention representation and concept mapping	Captures word order and local contextual cues that are critical for disambiguation in short clinical or social text, handles variable length inputs naturally, and can perform competitively with limited supervision when combined with domain embeddings	Weaker long context and global coherence than transformers, less parallelizable, depends on candidate generation at scale, often used as a component in hybrid pipelines	ADR and symptom normalization in social media, clinical notes, rapid domain shifts such as COVID, multilingual or transliteration scenarios
Transformer encoder methods	Domain pretrained transformers with self attention for contextual mention encoding and mention–concept matching via classification, ranking, or dense retrieval	Strong semantic representation, robust contextual disambiguation, scalable via retrieval and re ranking, supports metric learning alignment	High computational cost for fine tuning and inference, especially for cross encoder reranking, and sensitivity to candidate generation quality and ontology coverage when scaling to large vocabularies	UMLS and SNOMED CT linking, RxNorm and ICD mapping, EHR and clinical notes, multilingual normalization
GNN methods	Graph neural message passing over ontology or heterogeneous graphs that connect mentions, concepts, and contextual nodes, integrating textual embeddings with multi relational structure to support collective disambiguation and coherence aware linking	Explicitly leverages relational structure and global context to improve coherence across mentions and documents, strengthens disambiguation when surface forms are ambiguous, and can enhance rare entity linking by propagating evidence through concept neighborhoods	Graph construction and edge quality sensitive, scaling and memory overhead, often depends on strong text encoders for local semantics	UMLS based linking, coherence driven clinical EL, heterogeneous doc word concept graphs, cross lingual linking with graph enriched pretraining
LLM methods	Instruction tuned LLMs perform MCN by prompting the model to generate the target concept or identifier, can be guided by retrieved ontology candidates and adapted via LoRA style fine tuning	Excellent zero few shot transfer, handles noisy context well, useful for data augmentation and text cleaning, can provide explanations	High latency and cost, output variability and limited controllability, risks of hallucination or ungrounded normalization decisions without explicit candidate grounding	Low resource settings, multilingual clinical text, abbreviation disambiguation, text cleansing, data augmentation and prompt driven disambiguation
Other techniques	Rules, dictionaries, classical ML, hybrid pipelines with handcrafted similarity or heuristics	Interpretable and controllable decisions, low computational cost, strong precision for exact or near exact matches, and practical for rapid deployment or for domains where curated vocabularies and heuristics are reliable	Limited recall under paraphrases and unseen synonyms, brittle to noisy or highly variable expressions, and requires ongoing manual maintenance of dictionaries and heuristics; performance can degrade when terminology evolves or coverage is incomplete	Abbreviation disambiguation, dictionary driven clinical text, ADR normalization baselines, cross lingual translation plus matching pipelines

4. Evolution of medical concept normalization techniques and toolkits

Medical concept normalization (MCN) has progressed through a diverse range of deep learning architectures, each addressing distinct challenges in BEL and standardization. Fig. 5 presents a unified framework that integrates convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer-based models, graph neural networks (GNNs), and large language models (LLMs). Table 6 summarizes the core modeling idea, advantages, weaknesses, and usage scenarios of these major method architectures.

CNNs facilitate local feature extraction, while RNNs enhance sequential modeling by capturing contextual dependencies. Transformer-based models, especially for domain-specific models, such as BioBERT and PubMedBERT, leverage self-attention mechanisms to improve contextual representation, and are further refined through retrieval-based ranking and metric learning techniques. GNNs extend these capabilities by incorporating relational reasoning, modeling biomedical concepts as structured graphs to enhance entity disambiguation. More recently, LLMs, including GPT-4 and Llama, have introduced fine-tuning, zero/few-shot learning, and retrieval-augmented generation, demonstrating robust generalization in medical text normalization.

This unified perspective underscores the evolution of MCN methodologies, highlighting a shift from localized feature-based models to context-aware and knowledge-enriched frameworks that leverage both structured and unstructured information for enhanced normalization accuracy.

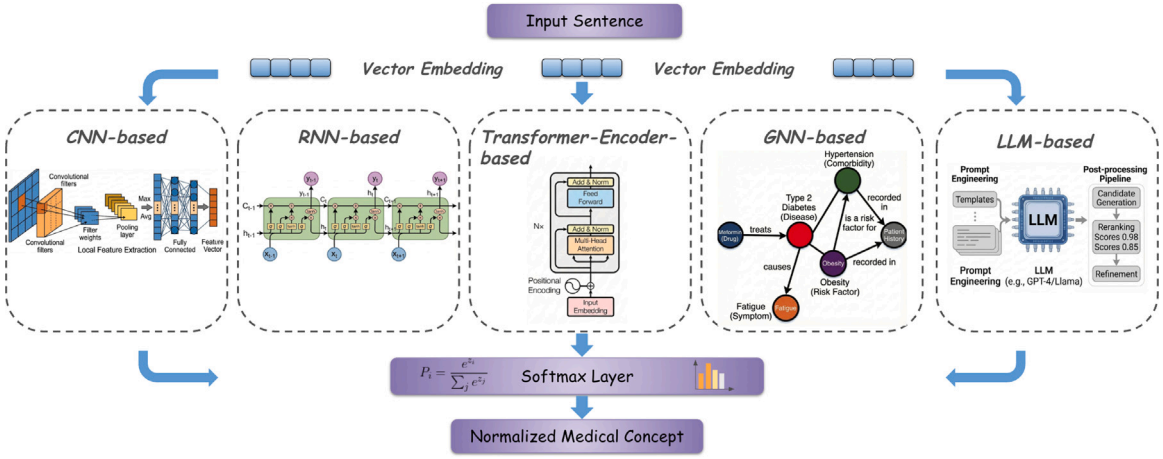

Fig. 5. Unified framework for Medical Concept Normalization approaches.

4.1. CNN-based methods

4.1.1. Modeling intuition

Convolutional Neural Networks (CNNs) have played a pivotal role in advancing medical concept normalization (MCN) by enabling the extraction of local semantic patterns from medical text. Their ability to process textual data at multiple levels of characters, words, and sentences has allowed researchers to tackle challenges such as noisy input, informal language, and domain-specific terminology. Over time, CNN-based methods have evolved, both independently and in combination with other architectures, to address specific challenges in MCN.

4.1.2. Representative methods

Early CNN-based approaches focused on capturing semantic patterns to normalize medical terms in noisy and unstructured text. Limsopatham and Collier [34] utilized convolutional layers to extract local semantic dependencies from word embeddings, applying max pooling to summarize these dependencies into fixed-size vectors that represented the entire message, which helped handle informal and colloquial phrases by focusing on contextual meaning. Li et al. [91] developed a CNN-based ranking framework that combined word embeddings with morphologically similar features, leveraging convolutional filters to encode the joint representation of entity mentions and candidate concepts and integrating custom morphological similarity vectors to enhance candidate ranking precision. To address the limitations of standalone CNNs in capturing sequential dependencies and broader contextual relationships, subsequent work developed hybrid CNN architectures. Lee et al. [1] combined CNNs for extracting local semantic patterns with Gated Recurrent Units (GRUs) to model sequential dependencies in user-generated medical texts, while Roller et al. [92] extended this idea to cross-lingual normalization by using CNNs with bidirectional GRUs (Bi-GRUs) to handle linguistic variations across languages.

4.1.3. Strengths and limitations

CNN-based methods are effective at extracting localized semantic and morphological cues that are prominent in noisy and informal mentions, and the convolution plus pooling design provides a compact representation that can support efficient matching and ranking. At the same time, standalone CNN encoders are limited in modeling longer contextual dependencies and broader relational signals, which motivates hybrid designs and richer supervision. In practice, this has led to multi-view and multi-task extensions that incorporate complementary contextual modeling and task signals, as well as domain-specific adaptations that introduce additional constraints or fairness objectives.

4.1.4. Typical datasets and usage scenarios

CNN-based MCN is commonly applied to settings characterized by noisy input and informal language, particularly user-generated medical texts where local surface cues are informative. These models are also used in multilingual and cross-lingual normalization scenarios where linguistic variations amplify the value of character- and word-level pattern extraction. Beyond these representative examples, multi-view and multi-task CNN frameworks such as MTMV-CNN [93] and F-DMTMV-CNN [94] process interactions between mentions and entities at the string, character, word, and sentence levels to improve robustness across diverse inputs and tasks. CNN hybrids are further adapted to domain-specific pipelines, including ADR normalization in drug labels [95] and normalization in Chinese clinical text [96]. More recently, fairness-aware variants integrate CNN-based backbones with Fair Identity Normalization (FIN) layers to mitigate demographic biases while preserving predictive performance, including applications on 2D RNFLT maps and 3D OCT scans [97].

4.2. RNN-based methods

4.2.1. Modeling intuition

Recurrent Neural Networks (RNNs) have significantly contributed to medical concept normalization (MCN) by capturing sequential dependencies and contextual relationships in biomedical text. Over time, these models have evolved to incorporate attention mechanisms, hybrid architectures, and knowledge-enhanced frameworks, addressing challenges, such as noisy data, language variability, and large-scale concept spaces.

4.2.2. Representative methods

A representative direction augments bidirectional recurrent encoders with attention to focus on informative spans in noisy narratives. Tutubalina et al. [98] employed Bi-LSTM and GRU models with attention to dynamically weight key segments in social media text, leveraging HealthVec and PubMedVec embeddings and cosine similarity features to enhance concept matching. Sarker et al. [66] similarly applied a bidirectional GRU with attention for ADR mention normalization, improving robustness to noisy and informal expressions. Beyond attention weighting, robustness to misspellings and rare words has also been strengthened through contextualized embeddings, for example, by combining Bi-LSTMs with ELMo representations for normalization in noisy social media settings [99]. Data-centric enhancements further improve RNN robustness by augmenting training instances with paraphrasing and character-level edits for layperson MEL [100].

Another representative line develops hybrid pipelines that complement recurrent contextual modeling with lexical normalization and

knowledge-driven ranking. Luo, Sun, and Rumshisky [101,102] combined Bi-LSTM encoders with exact matching and edit distance to handle typographically diverse yet semantically similar terms that rule-based methods alone could not resolve. Ruas, Lamurias, and Couto [35] introduced REEL, integrating Bi-LSTM-based relation extraction with Personalized PageRank to exploit semantic relations and ontology links for candidate ranking. Loureiro and Jorge [103] further extended this hybrid paradigm with a Bi-LSTM-CRF sequence labeling architecture that integrates neural contextual embeddings with dictionary matching, improving scalability and precision in practical entity linking pipelines.

4.2.3. Strengths and limitations

RNN-based models naturally encode word order and local contextual dependencies, which is beneficial when short surrounding context determines the correct concept in MCN. Attention mechanisms further improve interpretability and effectiveness by emphasizing salient spans in noisy inputs. However, standalone RNN encoders are limited in modeling long-range dependencies and global coherence, and their training is less parallelizable than transformer encoders. In large-scale normalization settings, end-to-end performance often depends on candidate generation quality and auxiliary constraints, which motivates hybrid designs that integrate lexical signals, dictionary matching, or knowledge-driven ranking.

4.2.4. Typical datasets and usage scenarios

RNN-based MCN methods are frequently applied to noisy user-generated medical text, including ADR and symptom normalization in social media, where sequential cues and attention-based focusing are valuable. Hybrid RNN pipelines have also been used in rapidly evolving domains such as COVID-19 symptom normalization [104] and in data-scarce settings supported by distant supervision [105]. Beyond English-centric scenarios, RNN architectures have been adapted to multilingual normalization, including transliteration into UMLS concepts [106] and Chinese MCN pipelines that incorporate medical dictionaries and linguistic features [107].

4.3. Transformer-based models

4.3.1. Modeling intuition

Transformer-based architectures have become the dominant paradigm for medical concept normalization (MCN) because self-attention enables long-range contextual encoding and produces mention representations that are sensitive to surrounding cues. In transformer encoder based MCN, a mention is encoded together with its surrounding context and then linked to a standardized concept by comparing the resulting representation against candidate concept names, descriptions, or identifiers. This comparison is commonly implemented as mention classification over a candidate set, mention–candidate scoring with a sentence-pair encoder, or embedding based retrieval with a bi-encoder followed by re-ranking. Beyond general-purpose transformers, domain-specific pretraining or continued pretraining on biomedical corpora such as PubMed, clinical narratives, or UMLS resources further improves terminology coverage and reduces semantic drift, while retrieval-based ranking and metric learning objectives enhance scalability and disambiguation when the ontology is large or the context is ambiguous.

4.3.2. Representative methods

A representative line of work focuses on domain-specific transformers and fine-tuning. Biomedical variants such as BioBERT, SciBERT, ClinicalBERT, PubMedBERT, BlueBERT, and UMLSBERT are fine-tuned for MCN as text classification or pairwise ranking, encoding the mention context and candidate descriptions with a shared encoder [108–113]. EhrBERT fine-tunes an encoder pretrained on EHR notes for MCN classification [114], while BertMCN adds a Highway Network to filter irrelevant information when encoding colloquial and clinical mentions [115]. To mitigate sparse training data, SciBERT with layer overwriting initializes output parameters from entity representations and augments training with synthetic instances derived from SNOMED-CT and RxNorm, using cosine similarity based alignment objectives [116].

Another representative direction emphasizes candidate retrieval and re-ranking to handle large ontologies. Typical pipelines use BM25 or dictionary based lookup (often enhanced with synonym resources) to generate high-recall candidates, then fine-tune a transformer to re-rank them [27,117–119]. Dual-encoder designs further improve retrieval efficiency, as illustrated by BENNERD, which applies BERT-based encoders for candidate retrieval and downstream linking in biomedical texts [120]. Complementing ranking based pipelines, metric learning approaches align mention and concept embeddings with contrastive or triplet objectives: SapBERT mines synonym triples from UMLS and optimizes a multi-similarity loss [121], while BioSyn combines sparse TF–IDF matching with dense transformer representations for robust retrieval [82,122]. Hard-negative sampling and hierarchy aware negatives are explored in DILBERT and KRISSBERT to improve fine-grained disambiguation and recall on rare entities [123–125]. Hybrid and knowledge-enriched designs incorporate structured signals from biomedical ontologies to strengthen disambiguation when context is insufficient. Some approaches inject ontology structure or hierarchical constraints into encoder representations, improving semantic type consistency and concept separability [126–128]. Another line adopts generative or constrained decoding formulations, where the model generates concept identifiers under ontology-aware constraints to reduce invalid or implausible predictions [129,130]. More recent work explores prompting and memory-style mechanisms to better utilize entity-type cues and cached mention–context evidence during inference [85,131–133].

4.3.3. Strengths and limitations

Transformer encoders provide strong context-aware disambiguation and are effective for handling lexical variation, abbreviations, and polysemy, particularly when coupled with biomedical pretraining and synonym alignment objectives. Their representations integrate naturally with retrieval and re-ranking pipelines, making them practical for large-scale ontologies. However, cross-encoder re-ranking can be computationally expensive, and overall performance often depends on candidate generation quality and ontology coverage. Model robustness may also degrade under domain shift or low-resource language settings, motivating multilingual adaptation, synthetic data augmentation, and knowledge-enriched objectives.

4.3.4. Typical datasets and usage scenarios

Transformer-based MCN is widely used in clinical narratives and EHR notes, biomedical literature, and noisy user-generated medical text, where long-range contextual cues are important for disambiguation. It is particularly suitable for ontology-scale normalization over UMLS-linked resources, where retrieval plus re-ranking or metric learning is required for efficiency. Multilingual and cross-lingual MCN increasingly relies on multilingual transformers and UMLS translation resources to support non-English clinical narratives [37,134–139]. Joint learning settings further combine normalization with related tasks such as NER or adverse event extraction to improve robustness under limited supervision [2,80,140–142].

4.4. GNN-based models

4.4.1. Modeling intuition

Graph Neural Networks (GNNs) introduce an explicit structural view for BEL and MCN by integrating textual encodings with graph-based knowledge representations. By propagating information over edges that encode biomedical relations, GNNs complement transformer

encoders with relational and global coherence signals, and are particularly useful when local textual evidence is insufficient for disambiguation. Concretely, GNN-based MCN methods treat biomedical concepts, mentions, and contextual evidence as nodes in a graph, and model their interactions through message passing. Transformer encoders such as BERT, BioBERT, or SapBERT typically provide initial node features for mentions or concept names, while graph modules such as GCN, GAT, GraphSAGE, or R-GCN refine these features by aggregating information from neighbors under typed relations. This graph-aware refinement supports collective disambiguation by encouraging predictions that are consistent with ontology structure, relational constraints, and document-level coherence.

4.4.2. Representative methods

Heterogeneous graph construction is a common design choice to jointly model mentions, entities, and contextual evidence under multiple relation types. Dai et al. [143] combined BioBERT embeddings with heterogeneous graphs to compute semantic affinities among mentions and entities, and used a minimum spanning forest based clustering strategy to enforce unique linking while maintaining semantic coherence. D'Auria et al. [144] leveraged BLINK for candidate generation and applied R-GCNs together with DistMult for multi-relational modeling, demonstrating that explicit relation modeling can strengthen embedding-based linking in clinical settings.

Attention-based message passing further improves disambiguation by weighting informative neighbors and injecting document-level coherence signals. Bo and Zhang [145] employed GATs to capture node-level interactions and integrate local mention–candidate similarity with global document coherence for improved ranking. Sakhovskiy et al. [146] extended SapBERT with a graph encoder (GraphSAGE or GAT) over UMLS, aligning textual and graph representations through multiple contrastive losses in GEBERT. Collective disambiguation has also been formulated as a clustering problem, where graph structure guides coherent grouping of mentions and entities, as in the BERT-encoded graph framework with hierarchical clustering proposed by Angell et al. [147]. Graph prompting has been explored in a related setting of biomedical synonym reasoning: GraphPrompt combines hierarchical graph-based prompts with BioBERT to improve synonym prediction using structured cues [148].

Multi-objective graph augmented pretraining provides another practical route to inject structural supervision while keeping inference lightweight. BERGAMOT combines pre-trained language models with GNN variants (GraphSAGE, RGCN, GAT) and optimizes contrastive and graph-infomax style objectives; graph representations are used during pretraining, yielding a graph-enriched transformer encoder for downstream BEL, including cross-lingual settings [149].

4.4.3. Strengths and limitations

GNN-based methods explicitly leverage ontology structure and relational dependencies, which can improve disambiguation under ambiguity, enforce global coherence, and benefit rare or sparsely expressed concepts through neighborhood information. They are naturally compatible with transformer encoders, which provide strong textual semantics while GNNs inject structured constraints. However, graph construction and message passing introduce additional engineering complexity, and performance can be sensitive to graph quality, relation coverage, and candidate generation. Large graphs may also increase training cost, and some designs rely on pretraining-time graph access that may not be available in all deployment settings.

4.4.4. Typical datasets and usage scenarios

GNN-based MCN is most beneficial when structured biomedical resources are available, such as UMLS-linked ontologies and knowledge graphs, and when entity linking requires coherence beyond local mention text. Typical scenarios include clinical entity linking with rich relational structure, collective disambiguation across documents, and settings where candidate sets are large and structurally confusable. Heterogeneous graph designs are also used in downstream biomedical applications that connect text to concepts, such as adverse drug event detection with document–word–concept graphs [150]. Cross-lingual or multilingual BEL can further benefit from graph-enriched pretraining that transfers structural supervision across languages [149].

4.5. LLM-based methods

4.5.1. Modeling intuition

Large language models (LLMs) bring a generative, instruction-driven view to BEL and MCN. Instead of only encoding mentions for classification or retrieval, LLM-based pipelines cast MCN as following natural-language instructions to identify the correct standardized concept, often by reasoning over mention context, surface-form variants, and candidate descriptions. In practice, LLMs are used in two main ways: (i) as a decision module that selects or justifies a concept from a provided candidate list (for example, retrieved by BM25, SapBERT, or other encoders), and (ii) as a domain-adapted generator via parameter-efficient fine-tuning, where lightweight updates such as LoRA tailor a general LLM to clinical terminology and normalization conventions.

4.5.2. Representative methods

Parameter-efficient fine-tuning has been adopted to specialize instruction-following LLMs for biomedical normalization. Wang et al. [151] fine-tuned Llama2 with LoRA for rare disease normalization, improving robustness to unseen synonyms, typographical errors, and lay expressions. Liu et al. [152] applied LoRA to Gemma on biomedical datasets, showing that competitive normalization performance can be obtained while updating only a small subset of parameters. Complementary work uses LLMs to align or enrich representations with external resources. Hosseini and Javidan [153] improved clinical abbreviation disambiguation by augmenting a BlueBERT-based setting with BioGPT-generated data. Remy et al. [154] combined GPT-3.5 with clinical knowledge graphs in BioLORD-2023, aligning text and graph signals to support domain and multilingual generalization.

Prompting-based use of LLMs emphasizes zero-shot and few-shot adaptation through natural-language instructions. Chen et al. [14] used ChatGPT for zero-shot and few-shot data generation to improve MCN training data quality, which in turn benefits downstream normalizers. BioLinkerAI [155] employed GPT-4 and Llama2 with enriched prompts together with rule-based components to perform context-aware disambiguation in a neuro-symbolic manner. Abdulnazar et al. [156] used GPT-4 for zero-shot text cleansing in German clinical narratives to reduce terminological noise before linking, and Borchert et al. [157] applied few-shot prompting to simplify long mentions, improving downstream linking for Spanish biomedical text.

4.5.3. Strengths and limitations

LLM-based methods are particularly useful when MCN depends on subtle contextual cues, instruction-driven resolution, or the need to produce explanations that clinicians can inspect. They also adapt well in low-supervision settings, either through carefully designed prompts or via parameter-efficient fine-tuning that injects domain constraints without retraining the full model. Their limitations are mainly practical: inference is typically more expensive than encoder-only pipelines, performance can be sensitive to prompt phrasing and how candidates are presented, and the model may hallucinate or give overconfident decisions when the candidate list is incomplete or when ontology constraints are not enforced explicitly.

4.5.4. Typical datasets and usage scenarios

LLM-based MCN is commonly used in low-resource or rapidly evolving settings where labeled data is scarce and adaptation via prompting or LoRA is practical, such as rare disease normalization [151]. Prompt-based cleansing and mention rewriting has been explored for non-English clinical narratives, including German text normalization [156] and Spanish biomedical mentions [157]. LLMs are also used as a controllable component in hybrid neuro-symbolic pipelines that combine retrieval, rules, and LLM-based reasoning for disambiguation [155].

4.6. Other techniques

4.6.1. Modeling intuition

Although deep neural architectures have become mainstream for MCN, traditional machine learning and hybrid pipelines remain practically important. These methods emphasize explicit lexical matching and lightweight decision rules, and are often used as strong baselines or as complementary components for candidate generation, lexical cleanup, and language adaptation, especially when labeled data is limited, terminology is highly variable, or multilingual transfer is required. Typical designs rely on curated dictionaries, heuristic normalization rules, and edit distance patterns to map mentions to canonical forms, or use feature-driven statistical classifiers to select among candidate concepts. A common strategy is to combine symbolic candidate construction with semantic similarity signals from distributional representations, so that exact matching handles high precision cases while similarity-based scoring covers paraphrases, misspellings, and informal expressions. In multilingual scenarios, these pipelines frequently introduce translation or language detection modules before performing dictionary alignment and candidate scoring.

4.6.2. Representative methods

Rule-oriented normalization remains effective for constrained subproblems. Mowery et al. [158] combined rule-based abbreviation mappings with CRFs and SVMs for clinical abbreviation disambiguation, leveraging complementary strengths of deterministic rules and learned models. Kate [159,160] proposed edit distance-based normalization that generalizes spelling variation patterns through exact matching and subconcept decomposition, improving robustness to lexical variation. Traditional feature-based learning is also used when deep models are unnecessary or infeasible. Emadzadeh et al. [161] employed SVMs together with latent semantic analysis and hybrid semantic measures to normalize adverse drug reaction mentions in noisy social media text. Semantic and dictionary-based pipelines integrate distributional semantics into candidate ranking. Cho et al. [162] combined dictionary matching with Word2Vec similarity to link mentions to biomedical concepts, while Karadeniz and Özgür [163] refined candidate ranking by incorporating syntactic signals alongside embedding-based similarity. Hybrid multilingual pipelines highlight the role of translation and language-specific preprocessing. Seinen et al. [164] validated Dutch clinical concept extraction tools via translation of English annotated corpora while preserving labels, and Almeida et al. [165] integrated language detection, text mining, and Levenshtein similarity to harmonize clinical data across languages.

4.6.3. Strengths and limitations

These techniques are attractive for their interpretability, low computational cost, and data efficiency. Rule and dictionary-driven components provide transparent decisions that are easy to audit, and they can offer strong performance when the terminology space is narrow or when high-quality lexicons are available. However, they tend to be brittle under paraphrasing and context-dependent ambiguity, and their coverage is limited by dictionary completeness and manual curation. Feature-based statistical models also require careful feature engineering and often struggle to generalize across domains or institutions without additional adaptation.

4.6.4. Typical datasets and usage scenarios

Other techniques are commonly used for clinical abbreviation disambiguation and short mention normalization, where explicit rules and lexicons are reliable [158,159]. They are also frequently applied to noisy user generated text, such as adverse drug reaction mentions in social media, where edit distance and lightweight classifiers provide robust baselines [161]. In cross-lingual and multilingual settings, translation-assisted annotation transfer and language-aware dictionary alignment are typical usage patterns for adapting MCN pipelines to new languages [164,165].

5. Model performance on MCN benchmarks

Building on Section 4, this section synthesizes state-of-the-art performance on representative MCN benchmarks. Rather than re-introducing each dataset, we focus on what the reported results reveal about model families, including discriminative rankers, generative sequence-to-sequence models, graph-enhanced architectures, and LLM-assisted pipelines, under different corpus conditions. Fig. 6 summarizes the co-evolution of model paradigms and benchmark construction.

Reported benchmark scores should be interpreted with caution. As discussed in Section 3 and Table 5, train-test leakage, duplication, class imbalance, and annotation inconsistencies can inflate headline metrics and obscure generalization gaps. In addition, metrics differ across benchmarks (Accuracy, F1, Recall), and candidate set construction varies substantially. Therefore, Table 9 is primarily used to compare trends within each benchmark family rather than to rank models across datasets.

Table 9 summarizes leading results on clinical corpora (MIMIC-III, ShARe/CLEF, CHIP-CDN), biomedical literature benchmarks (BC5CDR, NCBI Disease), social media and user-generated corpora (TwiMed, AskaPatient, TwADR-L, SMM4H-2017), and multilingual or domain-specific settings (Cantemist, DisTEMIST-linking subtrack, NEREL-BIO). For benchmarks with more than five reported results, we retain only the top five to emphasize competitive approaches.

5.1. Clinical text normalization datasets

Clinical benchmarks evaluate normalization in institution-specific notes and discharge summaries, where abbreviations and context-dependent disambiguation are common. In this setting, performance is typically driven by domain-adapted encoders and effective candidate ranking, while robustness is sensitive to label imbalance and the coverage of controlled terminologies.

5.1.1. MIMIC-III and MIMIC-IV

As introduced in Section 3, MIMIC benchmarks test clinical concept normalization against ICD codes. Reported results remain markedly lower than those on curated literature datasets, with an encoder–decoder model reaching 70.5% on MIMIC-III [166] and SNOBERT achieving 43.0% on MIMIC-IV [85]. This performance gap reflects the difficulty of mapping heterogeneous real-world phrasing to standardized codes. Moreover, the severe class imbalance in MIMIC-IV (Table 5) suggests that accuracy can be biased toward frequent concepts, and improvements should be interpreted together with imbalance-aware evaluation.

5.1.2. ShARe/CLEF

On ShARe/CLEF, biomedical transformer based approaches dominate the leaderboard, indicating that domain specific pretraining and strong contextual representations are critical for disambiguation in clinical notes. B LBConA [167] reports 94.2% accuracy, while KRISS-BERT [125] demonstrates that ontology aware contrastive objectives can further improve sense separation under abbreviations and synonymy. BLINKout [168] suggests that prompt driven linking and candidate reranking remain effective when paired with strong encoders.

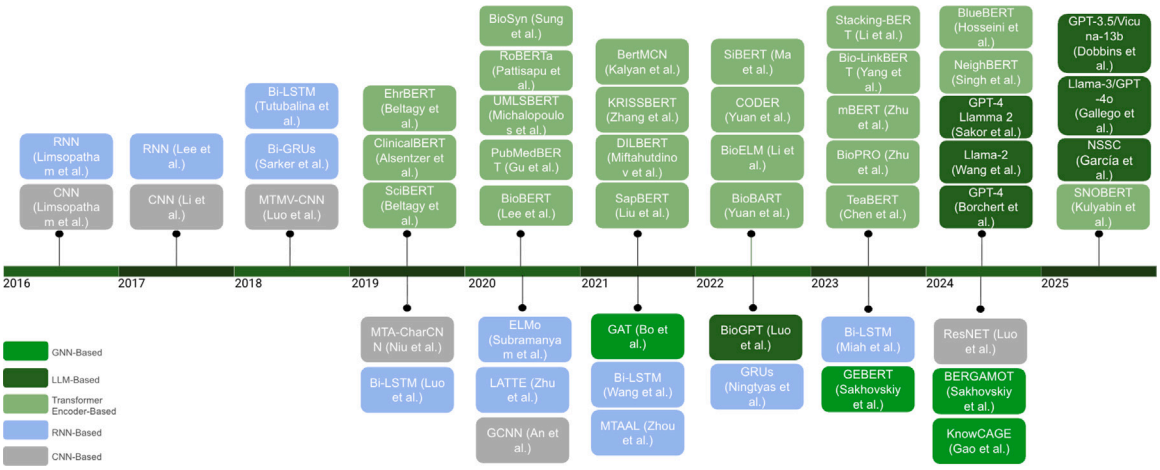

Fig. 6. The history and evolution of Medical Concept Normalization (MCN) models.

Table 7
Comparison of model sizes on the BC5CDR dataset..

Model	Backbone Model	Category	Model Size(Params)	Metric(Accuracy/F1/Recall)
BioLinkerAI [155]	LlaMA-2/GPT-4	LLM	70B (LLaMA-2)/GPT-4	93.3%
GenBioEL [173]	BART-Large	Transformer	406M	93.3%**
ANGEL [130]	BART-large	Transformer	406M	94.7%
BioPRO [131]	SAPBERT	Transformer	110M	94.4%
BioELM [126]	PubMedBERT	Transformer	110M	94.7%
GAT [145]	Graph Attention Network	Graph-Based	28.9M	89.3%
NormCo [174]	RNN-Based Model	RNN	9.1M	83.4%*

Note. In the Metric column, unmarked values denote Accuracy, values marked with * denote F1-score, and values marked with ** denote Recall. Different papers report different metrics, and these values should be interpreted under their corresponding metric definitions rather than being directly compared across metrics.

Hybrid neural baselines such as CNN plus BiLSTM [169] remain competitive, which supports a broader conclusion that careful candidate ranking and mention modeling can matter as much as increasing model scale on this benchmark.

5.1.3. CHIP-CDN

For CHIP-CDN, the best reported results remain substantially lower than English clinical benchmarks, reflecting both language-specific variability and the difficulty of aligning Chinese diagnosis expressions to standardized codes. BERT-based systems reach 64.2% accuracy [170], while graph-based methods [171] achieve 63.1%, indicating that injecting structured relations can help but does not fully resolve synonym and abbreviation variation. Contrastive representation learning, such as SimCSE [172] reports 57.6%, suggesting that representation alignment alone may be insufficient without improved candidate construction and more comprehensive coverage.

5.2. Biomedical literature normalization datasets

This category includes datasets extracted from scientific publications, research abstracts, and medical ontologies, where models must standardize disease, chemical, and gene mentions into structured biomedical vocabularies.

5.2.1. BC5CDR

On BC5CDR, transformer based pipelines achieve the strongest reported accuracy, with ANGEL [130] and BioELM [126] both reporting 94.7%. BioELQA [175] reports 94.5%, supporting the view that sequence-to-sequence modeling can be competitive when paired with strong biomedical pretraining and candidate control. GenBioEL [173], and BioPRO [131] further indicate that generative or hybrid designs can remain strong, especially when candidate recall and synonym coverage are improved.

At the same time, smaller architectures illustrate a meaningful efficiency frontier. NormCo [174] reaches 83.4% F1 with a compact RNN-based design, and GAT [145] reaches 89.3% accuracy with a moderate parameter budget, suggesting that targeted modeling of coherence or relational structure can provide competitive performance without large-scale encoders.

Recent work also explores large language models as decision modules. BioLinkerAI [155] combines rule-guided extraction with LLM-based disambiguation, reporting 93.3% accuracy. While this is strong, the comparison in Table 7 highlights a practical trade off between compute requirements and incremental gains over strong transformer baselines.

Finally, the split shift noted in Table 5, where many test disease mentions do not appear in training or development, implies that BC5CDR results are sensitive to candidate coverage and distribution mismatch. This supports evaluating methods not only by headline accuracy, but also by robustness to unseen mention variants.

5.2.2. NCBI disease

On NCBI Disease, SapBERT with prompt tuning [176] reports 94.5% accuracy, reinforcing the effectiveness of retrieval oriented learning for disease name normalization. B-LBConA [167] and BioELQA [175] report above 93%, suggesting that strong domain-specific encoders and careful ranking remain key. M^{3}E [132] reports 92.3% F1, indicating that memory and knowledge augmentation mechanisms can help with disambiguation, particularly when concepts share overlapping surface forms.

5.3. Social media and user-generated content dataset

This category includes datasets derived from Twitter, Reddit, patient forums, and online health discussions, where medical concept normalization (MCN) models must handle noisy, informal language, abbreviations, misspellings, and layperson terminology. Unlike datasets

Table 8
Explainability methods and fine-tuning impact on the AskaPatient dataset.

Model	Explainability method	Before Fine-tune (%)	After Fine-tune (%)	Relative improvement (%)
BERT (BioBERT) [27]	Generate-and-Rank with Semantic Type Regularization	87.52	87.46	−0.069
CNN+Attention [184]	Attention Mechanism with Character-Level Encoding	84.22	84.65	0.51
BART [173]	Knowledge Base-Guided Pretraining & Synonym Awareness	88.8	89.3	0.56
ANGEL [130]	Negative Sample Learning	89.3	90.2	1.01
KNN-BioEL [182]	Retrieval-Augmented Learning	89.1	90.3	1.35
BioBERT+MTL [185]	Multi-Task Learning with Context Utilization	83.88	85.37	1.78
ProSyno [183]	Prompt Learning with Wiktionary Descriptions	–	90.22	–
SapBERT [121]	Contrastive Learning with Negative Sample Augmentation	70.5	89.0	26.2
SapBERT+LLM [14]	LLM-Based Data Augmentation	86.66	88.01	1.56

Note. Accuracy (%) is used as the evaluation metric for all results.

derived from the biomedical literature or clinical records, these corpora require models to bridge the gap between colloquial expressions and standardized medical concepts, making normalization particularly challenging.

5.3.1. TwiMed

TwiMed shows high performance for strong domain adapted transformers, with FARM BERT reporting 95.9% F1 [177]. The competitiveness of multinomial naive Bayes at 91.0% accuracy [178] suggests that this benchmark still contains many lexically accessible cases. Deep learning variants such as DSN [179] and BiLSTM based designs [180] remain competitive, while BioLORD [181] indicates that contrastive pretraining may be beneficial but can be limited when mention noise and candidate ambiguity dominate. Overall, TwiMed results suggest that strong pretraining plus robust token level modeling can yield near ceiling performance when the concept space is relatively constrained.

5.3.2. AskaPatient

AskaPatient performance highlights the importance of both candidate retrieval and synonym-aware modeling for patient language. BART-based methods [173] report 93.3% F1, while KNN BioEL [182] reaches 90.3%, showing that nearest neighbor retrieval can effectively recover plausible candidates from informal mentions. ProSyno [183] and ANGEL [130] report about 90.2%, suggesting that prompt learning and negative sampling strategies can be effective when they improve alignment between lay expressions and standardized terms. BioBERT [27] remains a strong baseline at 87.5%, indicating that encoder strength alone can capture a substantial portion of the mapping.

To complement leaderboard numbers, Table 8 summarizes representative explainability-oriented designs and the effect of finetuning. The results highlight that interpretability is often improved through explicit token weighting, semantic type constraints, or retrieval augmented reasoning, which is especially important when user-generated mentions are ambiguous.

Importantly, the overlap between training and test splits reported in Table 5 suggests that headline scores on AskaPatient can overestimate generalization to truly unseen paraphrases. This motivates evaluating models under stricter split policies when claiming robustness to real-world patient language.

5.3.3. TwADR-L

TwADR L remains one of the most challenging social media benchmarks, with the best reported accuracy at 51.5% by ProSyno [183]. BioBERT [27] reports 47.0% and BERTMCN [115] reports 48.3%, while CNN with attention [184] and hybrid BERT plus RNN [80] are lower. The narrow separation among methods suggests that the dominant limitation is not encoder capacity alone, but candidate scarcity and extreme tail distributions.

Consistent with the data quality analysis in Table 5, TwADR L suffers from severe imbalance, measurable overlap between training and test data, and limited concept coverage. These issues imply that progress on TwADR L is likely to depend on improved data augmentation, better candidate expansion for rare concepts, and stricter evaluation splits that reduce leakage.

5.3.4. SMM4H-2017

On SMM4H 2017, BERT based systems report 69.64 [186], while a BioBERT plus BioSyn ranking approach reports 60.5 [82]. The gap suggests that representation strength and adaptation can be decisive when informal phrasing interacts with limited context, but that robust reranking remains difficult under noisy mention forms. The overall performance indicates that social media normalization for adverse events continues to require stronger candidate construction and better disambiguation signals beyond surface similarity.

5.4. Multilingual and domain-specific datasets

This category includes datasets designed for non-English medical concept normalization and domain-specific entity linking, where models must generalize across different languages, specialized medical subfields, and unique terminologies. These benchmarks are particularly challenging due to linguistic variations, diverse annotation standards, and differences in concept mapping strategies across languages and medical domains.

5.4.1. Cantemist, DisTEMIST-linking subtrack and NEREL-BIO

On Spanish benchmarks, RoBERTa reports 41.2% F1 on DisTEMIST linking [187], indicating that clinical entity linking can remain difficult even with strong encoders when candidate ambiguity is high. On the Cantemist benchmark family, BETO [188] reports 86.4% F1, showing that language-specific pretraining can yield strong normalization performance when the concept inventory and annotation scheme are well aligned. For Russian biomedical normalization, BERT achieves 72.7% F1 on NEREL BIO [189], highlighting growing progress in multilingual biomedical NLP, while leaving room for improvement in cross-vocabulary consistency and low-resource coverage.

5.5. Bias detection and mitigation in MCN benchmark evaluation

Several widely used MCN benchmarks contain data quality issues that introduce bias, such as train-test leakage, duplication, or class imbalance, as discussed in previous sections. This is also reflected in Table 9, where representative benchmarks are annotated with their dominant quality concerns. Therefore, beyond reporting the best numbers, a key question is how to detect such biases and design evaluation protocols that better reflect model generalization.

Recognizing these challenges, Agarwal et al. [136] systematically investigated dataset bias in medical concept normalization, focusing on issues, such as data leakage between training and test splits and excessive overlap with candidate dictionaries. To detect and address

Table 9
Model performance on MCN benchmarks. For benchmarks with more than five model testing results, we only keep the top five.

Dataset	Quality Issue	Fine-Grained Model	Category	Metrics	Code Source	Year
TwiMed	—	FARM-BERT [177]	Transformer	95.9*	–	2021
		Multinomial NB [178]	Machine Learning	91.0	–	2023
		DSN [179]	Deep Learning	89.8*	–	2024
		BiLSTM+MSAM+position [180]	Neural Network	85.3*	–	2019
		BioLORD [181]	Transformer	70.6	–	2023
BC5CDR	Comprehensiveness	ANGEL [130]	Transformer	94.7	link	2024
		BioELM [126]	Transformer	94.7	–	2022
		BioELQA [175]	Transformer	94.5	–	2024
		GenBioEL [173]	Transformer	93.3**	link	2022
		BioPRO [131]	Transformer	94.4	link	2023
NCBI	—	BERT (SapBERT, PubMedBERT) [176]	Transformer	94.5	link	2022
		B-LBConA [167]	Transformer	93.6	–	2023
		BioELQA [175]	Transformer	93.5	–	2024
		BART [173]	Transformer	93.3	link	2022
		M^3E [132]	Transformer	92.3*	link	2024
SMM4H-2017	—	BERT (SapBERT, PubMedBERT), GPT-2 [186]	Transformer	69.64	link	2022
		BERT (BioBERT, BioSyn) [82]	Transformer	60.5	link	2020
HMC2019	—	BERT [190]	Transformer	95.0*	–	2024
		DeBERTa+RWP+CL [191]	Transformer, Contrastive learning	93.9*	–	2023
		RoBERTa-MTL [192]	Transformer	90.2*	link	2023
AskaPatient	Duplication	BART [173]	Transformer	89.3*	–	2022
		KNN-BioEL [182]	Transformer	90.3	link	2024
		ProSyno [183]	Transformer	90.2	–	2025
		ANGEL [130]	Transformer	90.2	link	2024
		BERT (BioBERT) [27]	Transformer	87.5	link	2020
ShARe/CLEF	Comprehensiveness	B-LBConA [167]	Transformer	94.2	–	2023
		KRISSBERT [125]	Transformer	90.4	link	2021
		BLINKout [168]	Transformer	91.2	link	2023
		CNN, RNN (Bi-LSTM) [169]	Neural Network	90.1	link	2021
COMETA	Comprehensiveness	BART [173]	Transformer	93.3	–	2022
		KNN-BioEL [182]	Transformer	85.7	link	2024
		BioELQA [175]	Transformer	85.2	–	2024
		BERT (SapBERT), Prompt base encoder [193]	Transformer	83.7	link	2023
		ANGEL [130]	Transformer	82.8	link	2024
		GenBioEL [173]	Transformer	82.8	link	2022
MedMentions	Comprehensiveness	M^3E [132]	Transformer	88.7*	link	2024
		BioLinkerAI [155]	LLM	81.3	–	2024
		BioFEG [194]	Transformer	76.7	link	2023
		Joint Learning [184]	Transformer	75.1	–	2019
Cadec	—	BERT (SapBERT, PubMedBERT), GPT-2 [186]	Transformer	84.1	link	2022
		ProSyno [183]	Transformer	60.6	–	2025
PsyTAR	—	BERT (BioBERT)+Multitask Learning [185]	Transformer	85.4	–	2022
MIMIC-III	—	encode–decode [166]	Neural Network	70.5	–	2018
TwADR-L	Imbalance, Duplication and Comprehensiveness	ProSyno [183]	Transformer	51.5	–	2025
		BERT (BioBERT) [27]	Transformer	47.0	link	2020
		BertMCN [115]	Transformer	48.3	–	2021
		CNN+Attention [184]	Neural Network	46.5	–	2019
		BERT+RNN (ULMFit – > AWD-LSTM) [80]	Transformer, RNN	41.7	link	2021
CHIP-CDN*	—	BERT [170]	Transformer	64.2*	link	2024
		GNN [171]	Neural Network	63.1*	–	2023
		SimCSE [172]	Contrastive learning	57.6*	–	2022
Catemist*	—	BETO [188]	Transformer	86.4*	link	2021
		Bi-LSTM [195]	Neural Network	73.7	–	2020
BioWiC	—	Llama2 [44]	LLM	78.0	link	2024
MedRed	—	RoBERTa [69]	Transformer	73.0	–	2020
NEREL-BIO*	—	BERT [43]	Transformer	72.7	link	2024
MIMIC-IV	Imbalance	SNOBERT [85]	Transformer	43.0	link	2025
MADE	—	EhrBERT [114]	Transformer	41.0	–	2019
DisTEMIST-linking subtrack*	—	RoBERTa [187]	Transformer	41.2*	link	2022

Note. Unmarked values denote Accuracy, * denotes F1-score, and ** denotes Recall. Metrics vary across papers and are not directly comparable.

these biases, they propose a comprehensive filtering framework that removes from the test set any entity mentions also present in the training data or highly similar to dictionary entries, based on surface form or lexical similarity. This approach is applied across several multilingual clinical datasets, ensuring that evaluation is not confounded by trivial matches or memorization. By constructing such filtered test sets, the authors provide a more realistic assessment of model generalization and establish practical guidelines for creating unbiased evaluation settings in this field. In the context of Table 9, such overlap and duplication control is particularly relevant for duplication-prone benchmarks (e.g., AskaPatient and TwADR-L), where strong Accuracy or F1 scores may be optimistically estimated and partially reflect repeated surface forms rather than robust generalization.

In a similar direction, Sood and Imran [196] on biomedical information retrieval highlighted the necessity of benchmark datasets with gold-standard annotations, careful curation, and thorough documentation of potential sources of bias. They recommend mapping both queries and candidates to standardized concepts using controlled vocabularies and ontologies, such as UMLS, MeSH, and SNOMED CT. This approach helps to mitigate problems of term mismatch and semantic drift. Additionally, they suggest employing corpus-based query expansion strategies and filtering mechanisms to prevent retrieval bias or query drift caused by noisy or irrelevant augmentation. These recommendations are especially important for benchmarks with broader concept coverage and larger candidate spaces (e.g., BC5CDR, MedMentions, COMETA, and ShARe/CLEF in Table 9), where the evaluation outcome is more sensitive to candidate construction, annotation conventions, and the long tail of concepts.

Building on these principles, recent work has developed a multistage framework for biomedical concept normalization by leveraging large language models alongside traditional rule-based systems [197]. In this approach, LLMs, such as GPT-3.5-turbo and Vicuna-13b, are first prompted to generate synonyms and alternative phrasings, which maximizes recall during the candidate generation phase. In the second stage, LLMs are used to prune and rank candidate concepts, thereby improving precision. These results show that LLMs can be effectively combined with existing normalization systems to boost performance, particularly in low-resource scenarios, provided that evaluation is conducted on high-quality datasets that are free from leakage and other biases. From the perspective of Table 9, such multi-stage designs can be beneficial under highly imbalanced settings (e.g., TwADR-L and MIMIC-IV), where naive ranking can be dominated by frequent concepts and performance on rare concepts is more brittle.

Together, these approaches provide a foundation for more transparent and reliable MCN evaluation, and highlight the ongoing need for rigorous dataset quality control and bias mitigation in future research.

6. Toolkits for MCN

6.1. Existing MCN toolkits

Open-source toolkits for medical concept normalization (MCN) have become essential for enabling scalable, robust, and reproducible normalization across diverse biomedical datasets and multilingual settings. These toolkits provide modular architectures, standardized interfaces, and integrated data scripts, empowering both research innovation and real-world clinical applications. Table 10 summarizes existing MCN toolkits with language and deployment support. The latest toolkits after 2024 are introduced as follows.

HunFlair2 [198] is a comprehensive toolkit for BNER and BNEN, directly integrated with the Flair NLP framework. Leveraging its All-in-One NER (AIONER) architecture, HunFlair2 supports a broad spectrum of biomedical entities and incorporates advanced neural models, such as LinkBERT and SapBERT, to handle synonym variability and species-specific normalization. Its proven scalability and generalization across benchmarks like MedMentions and BioID have established HunFlair2 as a cornerstone resource for biomedical NLP.

Preon [199] is a specialized toolkit for precision oncology, focusing on the normalization of drug names and cancer types. Preon employs a multi-step matching pipeline, including combining exact, token-based, and edit-distance methods, to maximize precision and efficiency. It links drug names to ChEMBL and cancer types to Disease Ontology (DO), demonstrating superior performance to existing solutions, such as MetaKB, on datasets including CIViC and ClinicalTrials.gov.

Thera-Py [200] is a Python-based toolkit and web API designed for the normalization of therapeutic concepts. Utilizing a graph-based merging strategy, Thera-Py unifies entities from major resources, such as RxNorm, NCIt, and DrugBank, thereby enhancing cross-resource alignment and normalization coverage for datasets like PharmGKB and OncoKB.

xMEN [87] is a modular toolkit addressing multilingual MCN challenges through a generate-and-rank pipeline that integrates TF–IDF, SapBERT embeddings, and cross-encoders. Its architecture is especially effective for low-resource languages, leveraging weak supervision and data augmentation scripts to provide a systematic and reproducible solution for multilingual normalization tasks.

To facilitate evaluation standardization and interoperability, benchmarking toolkits, such as BigBIO [201],[15] supply unified data formats and standardized task definitions across a wide range of biomedical NLP datasets, supporting fair and comprehensive MCN assessment.

Collectively, these toolkits reflect the growing maturity and diversity of open source MCN resources, supporting both methodological research and downstream clinical applications. Looking ahead, there remains a strong need for further standardization of evaluation metrics and enhanced integration of toolkits into clinical workflows, as discussed in Section 7.

6.2. Observations

The reviewed MCN toolkits differ in scope, methods, and intended use. Early systems such as cTAKES, MetaMapLite, and QuickUMLS rely mainly on rule-based or dictionary matching over UMLS resources. They are easy to deploy, transparent, and stable, which makes them suitable for clinical pipelines and exploratory analysis. However, they often struggle with ambiguous mentions and unseen terminology. Newer toolkits such as TaggerOne, BioSyn, BERN2, MedCAT, and HunFlair2 use machine learning or hybrid approaches. These systems generally achieve higher normalization accuracy, especially in biomedical literature and complex clinical text. Their disadvantages include higher computational cost, greater configuration effort, and dependence on training data. Multilingual support is still limited in most tools, although xMEN and HunFlair2 provide stronger support for non-English settings.

Several toolkits are designed for specific application domains and show clear advantages in those contexts. Preon focuses on precision oncology; Thera-Py targets therapeutic concepts and integrates multiple drug and cancer knowledge bases through a graph-based strategy. These specialized tools often outperform general-purpose systems on domain-specific datasets. Therefore, toolkit selection should depend on the target entity types, the language and domain of the data, and practical constraints such as deployment and maintenance. General UMLS-based tools are suitable for broad and heterogeneous datasets, while specialized toolkits are better choices when high precision and domain consistency are required.

[15] https://github.com/bigscience-workshop/biomedical

Table 10
Overview of representative open-source medical concept normalization (MCN) toolkits, organized by year of release.

Toolkit	Year	Ontology	Domain	Language Support	Deployment
cTAKES	2010	UMLS, SNOMED-CT, RxNorm	Biomedical, clinical, social media	English	Java
TaggerOne	2016	Custom	Biomedical	English	Java
MetaMapLite	2016	UMLS	Clinical	English	Java
QuickUMLS	2016	UMLS	Biomedical, clinical, social media	Multilingual	Python/API
MedLinker	2020	UMLS	Biomedical	English	Python
BERN2 [202]	2020	UMLS	Biomedical	English	API/Server
BioSyn [203]	2020	UMLS, MeSH, OMIM, MedDRA	Biomedical	English	Python
MedCAT	2020	SNOMED-CT, UMLS, Human Phenotype Ontology (HPO)	Clinical	English	Python/API
KAZU [204]	2022	GO, HPO	Biomedical	English	Python/JVM
HunFlair2	2023	Multiple	Biomedical	English, German, Dutch, Spanish	Pytorch
Thera-Py [200]	2023	RxNorm, NCIt, HemOnc, DrugBank, Drugs@FDA, IUPHAR Guide to Pharmacology, ChEMBL, ChemIDplus, Wikidata	Biomedical, clinical	English	Python/API
Preon [199]	2024	ChEMBL, Disease Ontology (DO)	Biomedical	English	Python
xMEN [87]	2024	UMLS, Custom	Biomedical	English, Spanish, French, German, Dutch	Python

7. MCN applications

Medical concept normalization (MCN) bridges the gap between heterogeneous clinical texts and standardized terminologies, laying the foundation of various healthcare applications. For example, automated MCN models and tools can reduce the reliance on manual annotation and coding by clinical staff, saving time and minimizing errors [85]. By mapping diverse expressions of the same medical entity to a unified concept identifier, MCN enables seamless data integration across disparate systems, bolsters interoperability, and enhances the accuracy of subsequent healthcare data management and decision-making processes. This section summarizes the key applications of MCN, ranging from streamlining electronic health records (EHRs) management, improving clinical decision support (CDS) systems, facilitating medical research to enhancing health information exchange (HIE), and supporting automated patient messaging systems, among others.

7.1. EHR management

In electronic health records (EHRs), different terms or abbreviations are frequently used to refer to the same condition, which easily cause ambiguity and confusion. Medical concept normalization (MCN) helps ensure that both machines and humans operate on a consistent, unambiguous representation of key medical concepts [183] by automatically uncovering and aligning synonymous expressions across diverse clinical documents [183]. For example, expressions, such as “heart attack”, “myocardial infarction”, and “MI”, from multiple sources (e.g., combining hospital records, lab reports, and imaging reports) can be linked to a single standardized identifier, enabling clearer patient profiles and facilitating smooth data integration and exchange between EHR systems. Recently, cross-lingual MCN has been proposed to aligning clinical entities across different languages to a unified terminology, which enable seamless information organization, retrieval, and analysis from multilingual medical documents [87].

7.2. Clinical decision support (CDS)

MCN plays a critical role in clinical decision-making by providing a standardized medical lexicon, enabling the accurate analysis of clinical notes, laboratory results, and patient information. With normalized medical concepts, decision-support algorithms can more effectively interpret and interact with the data [205]. For instance, when patient symptoms, diagnoses, and test results are consistently coded, clinical decision support (CDS) systems can automatically present treatment plans and recommendations. MCN also helps identify patterns and trends across patient populations, supporting personalized care [117]. Additionally, MCN reduces the risk of errors arising from the misinterpretation of medical terms, preventing misdiagnoses or incorrect treatments. Consistent terminology ensures clear communication among healthcare providers, promoting optimal healthcare delivery [37]. CDS systems can also better search for guidelines and recommendations if the medical terms in patient records are normalized to a known vocabulary [87].

7.3. Medical research

MCN can facilitate medical research from different perspectives. For instance, researchers depend on MCN to harmonize and integrate data from clinical trials, patient cohorts, and other studies [35]; MCN is also valuable for meta-analyses and systematic reviews, as it maintains consistency in indicators and uniformity in outcomes [166]; and normalized data based on MCN (e.g., SNOMED CT terms) also make it easier to run statistical analyses, track disease trends, or summarize outcomes at scale. When entities in clinical notes are normalized to SNOMED CT, researchers can quickly identify patients with specific diagnoses or conditions, even if there are multiple ways of describing them in the clinical text [206]. In summary, a standardized medical vocabulary is essential for research by pooling large sets of EHR data, ensuring data consistency across studies, and enabling the aggregation and comparison of results [206].

7.4. Health information exchange

MCN also finds important applications in health information exchange (HIE), particularly in scenarios involving large-scale and heterogeneous data sources. One such example is the analysis of social media data for detecting drug abuse patterns [207]. Patient or consumer posts often include colloquial or brand names for drugs — sometimes spelled incorrectly or misused — making direct mapping to standardized terminologies challenging. By normalizing these mentions to official drug codes, public health authorities and researchers can more accurately monitor emerging trends and identify at-risk populations [208,209]. Similarly, symptom or disease normalization in public health surveillance can significantly improve outbreak detection [71,210], enabling algorithms to aggregate and interpret symptom or disease mentions from diverse sources — ranging from community forums and social media, to emergency department reports — under a unified set of concepts [104,210].

7.5. Precision medicine and clinical trial retrieval

One of the main components of constructing an accurate precision medicine and clinical trial retrieval system is query expansion [124,211,212]. MCN can significantly enhance query expansion by aligning free-text queries with standardized terminologies and thereby uncovering relevant synonyms, hierarchies, and related concepts [213]. For instance, a query for "EGFR mutation–positive tumor" can be expanded to include variations, such as "epidermal growth factor receptor–positive cancer", brand-specific medications targeting EGFR, and corresponding ICD or SNOMED codes. This broader query scope ensures that otherwise disparate representations — commonly found in clinical notes, research articles, or EHR repositories — are comprehensively retrieved. By tapping into curated ontologies like UMLS, SNOMED CT, or MeSH, MCN-based query expansion can integrate disease subtypes, genetic profiles, and treatment options under a single standardized umbrella, thus improving recall and precision in search results [211,214]. In precision medicine and clinical trial retrieval contexts, such an approach is especially valuable when searching for clinically relevant treatments or studies tied to specific biomarkers, as it automatically accounts for the variations and synonyms that naturally arise in medical documentation [196,215]. The result is a more robust retrieval pipeline that helps clinicians, researchers, and patients alike to identify pertinent data without missing crucial information hidden under alternate terminologies or abbreviations.

7.6. Automated patient messaging systems

MCN is also beneficial for automated patient messaging systems, including virtual health assistants and chatbots, by ensuring consistent understanding of medical terms and conditions across a wide range of user inputs [216,217]. For example, patients might describe their ailments in colloquial language or abbreviations (e.g., "I have a pain in my tummy" versus "I have abdominal pain"), leading to ambiguity if not mapped to standardized medical concepts. By normalizing these diverse expressions to well-defined terminologies (e.g., ICD-10, SNOMED CT), chatbots can more accurately identify patient needs and provide relevant health information or care recommendations [216–218]. MCN-enhanced patient messaging systems can also flag urgent or severe conditions by matching user-reported symptoms to potential diagnoses with greater reliability, thus directing patients to timely medical attention. In turn, this consistent concept mapping supports downstream tasks, such as clinical triage, medication reminders, and integrated remote monitoring, making virtual health assistants more robust, trustworthy, and efficient in diverse real-world usage scenarios.

8. Challenges and possible future directions

8.1. Challenges

Although medical concept normalization (MCN) has gained momentum over the past decade, it remains a formidable task due to the sheer variety of textual expressions — acronyms, abbreviations, misspellings, and domain-specific synonyms — which complicate consistent mapping to standardized terminologies [9,151,219]. This complexity intensifies in specialized domains, such as rare diseases and pharmacovigilance, where scarce training data and frequent typographical inconsistencies further hinder accurate normalization [9,151,220]. In addition, composite mentions (often involving coordinate structures with "and"/"or") present a unique challenge by requiring multiple knowledge base (KB) codes for a single textual mention — an aspect frequently overlooked by existing studies [157].

Another challenge lies in the lack of large-scale, high-quality data sets, which serve as the backbone for training deep learning-based MCN models. Constructing such datasets fully through human effort is cost-prohibitive, and while many publicly available MCN datasets span different ontologies, several suffer from data quality issues, including duplication, limited coverage, class imbalance, and other inconsistencies [1,14,80]. These shortcomings can bias model performance and produce misleading evaluation results [80], especially in low-resource settings — for instance, datasets for minority languages or underrepresented populations — where data paucity and limited label diversity become even more pronounced [97,221]. Existing datasets also often contain a narrow set of diseases and symptoms, further limiting generalizability to real-world clinical scenarios.

The third challenge relates to end-to-end evaluation pipelines. Many benchmarks separately assess recognition and normalization, rather than providing combined results for named entity recognition (NER) and normalization (NEN) [141]. When upstream entity recognition is inaccurate, MCN performance suffers dramatically, creating cascading errors in downstream tasks like relation extraction [198]. Even though MCN models show strong performance on common, highly researched entities, they frequently fail on concepts that appear infrequently in the literature [198]. Developing more robust, end-to-end benchmarks and architectures could help address these pipeline vulnerabilities.

The interpretability of deep learning-based MCN models poses another challenge. While deep neural networks, especially transformer-based architectures, deliver state-of-the-art performance, they often operate as "black boxes" whose internal logic is opaque. This lack of explainability can be problematic for clinical applications, where healthcare providers need to understand the rationale behind model outputs, particularly when patient outcomes and treatment decisions are at stake.

Even though LLMs have become the dominating approaches for MCN, the performance improvement of LLM-based methods has been constrained by a lack of high-quality domain knowledge. This challenge is urgent as LLM-based methods rely heavily on knowledge bases that may be incomplete or slow to update [154]. Many entities in medical ontologies are sparsely defined, leading to indistinguishable embeddings for similar concepts. Moreover, domain knowledge evolves rapidly, so newly discovered or emergent concepts may not be included in existing ontologies. This gap undermines the model's ability to accurately differentiate between closely related entities or to map novel concepts to standardized codes.

Moreover, LLM-based MCN is also facing a critical challenge on hallucinations, when LLMs invent or confidently assert incorrect information [156]. In MCN workflows, an LLM might be prompted with a clinical term or phrase to produce a standardized concept (e.g., ICD-10 or SNOMED-CT code). However, the model's generative nature can lead it to suggest fictitious codes or erroneously map to superficially similar concepts. Detection of such errors is difficult without systematic validation, given that LLMs often output fluent, confident-sounding text. In clinical contexts, even a minor mismatch between a mention and a standardized code can have serious repercussions for patient care, raising both ethical and safety concerns.

Last, but not least, computational costs present another major challenge for LLM-based MCN. Methods involving large-scale transformers typically demand more computational resources and longer processing times than simpler dictionary- or rules-based approaches. These higher costs can be prohibitive in real-time clinical applications, where rapid turnaround is often crucial. Practical considerations, such as budget, infrastructure, and the specific nature of the target terminology (e.g., well-known versus rarely researched codes), determine the most suitable normalization method [222].

8.2. Possible future directions

Building on the progress made in medical concept normalization (MCN) and the current challenges, several possible future directions deserve exploration:

Leveraging LLMs with Data Augmentation. As large language models (LLMs) continue to deliver state-of-the-art results for MCN, data

augmentation represents a critical step to address data scarcity and enhance both accuracy and robustness. Recent work has shown that prompting LLMs (e.g., zero-shot and few-shot approaches) can generate synthetic training samples with minimal computational overhead [183,223–225]. Applied in clinical and biomedical domains, this technique has demonstrated promise in expanding limited datasets and reducing labeling costs [14,226,227]. Future research could refine prompting strategies, incorporate domain-specific constraints, and systematically evaluate the quality of generated examples to further boost MCN performance.

Enhancing Domain Knowledge Integration. The effectiveness of LLMs in MCN relies heavily on high-quality domain knowledge. Consequently, acquiring and curating large-scale, up-to-date medical knowledge bases (KBs) and knowledge graphs (KGs) is essential [107,153,183,228]. Next steps include developing domain-specific LLMs and fine-tuning general LLMs on specialized corpora, as well as exploring KG-enhanced architectures. These knowledge-rich approaches could significantly improve normalization quality, especially for less common or rapidly evolving medical concepts.

Addressing Safety and Privacy in LLM-based MCN. Although LLMs offer powerful language-understanding capabilities, concerns remain regarding hallucinations, black-box decision-making, and data privacy. In clinical contexts, misclassifications can have serious repercussions for patient care [156]. One promising direction involves retrieval-augmented generation (RAG), whereby an LLM's outputs are validated or supplemented with external KB lookups. This could be particularly valuable for automatically identifying adverse drug reactions (ADRs) in electronic health records (EHRs), where structured data may be incomplete. Likewise, cloud-based or commercial LLM deployments must ensure robust data encryption and compliance with healthcare privacy regulations. Incorporating user feedback loops — allowing clinicians to quickly correct errors — can continuously improve model performance and instill trust in automated recommendations [222].

Bridging Health Disparities in Rural, Under-served, and Non-English Settings. MCN for underrepresented populations remains a pressing area of concern. Current algorithms often overlook the unique needs of rural or under-served communities, perpetuating health disparities [97,221]. Similarly, multilingual and cross-lingual MCN can broaden access to standardized terminologies for large non-English-speaking populations [229,230]. Key efforts include collecting and annotating minority-language clinical corpora at high granularity, and harnessing multilingual transformer-based architectures. Exploring LLM-driven synthetic corpus creation [164] could also mitigate data shortages in these low-resource contexts, facilitating more inclusive MCN solutions.

Multimodal MCN. Another promising direction involves incorporating non-textual data, such as radiology images, pathology slides, and lab results, into the normalization pipeline [231,232]. By aligning textual descriptions with imaging evidence, models can better disambiguate closely related medical entities, thereby improving overall accuracy. Recent strides in multimodal vision–language models suggest that integrating disparate data streams is increasingly feasible, paving the way for more holistic concept mappings that reflect the patient's full clinical profile.

Real-World Clinical Integration. Finally, embedding MCN directly into EHR systems and clinical workflows can unlock meaningful improvements in decision support, medical coding, and patient documentation. Providing real-time feedback at the point of care — rather than as a back-end process — enables clinicians to verify or refine model outputs, thereby boosting both efficiency and trust. Ensuring transparency and interpretability of MCN models will be critical, so that healthcare providers can understand the rationale behind system recommendations.

9. Conclusion

This paper presents a comprehensive survey on medical concept normalization (MCN) in the last decade. We review existing studies by categorizing them into task definition, dataset construction, model development, performance evaluation, applications, and summarizing current challenges and limitations. We observe that there has been extensive research on MCN with a significant number of benchmark datasets. However, more datasets are expected to be developed by expanding the set of different concepts (e.g., diseases, symptoms, and drugs) and different languages. We sort out different machine learning and deep learning techniques for MCN and also present a summary of the experimental results of the state-of-the-art models on benchmark datasets using these techniques. Deep learning models, especially transformer-based language models, have demonstrated prestigious performance on MCN under different scenarios. A thorough review of the practical applications that utilize MCN in real-world clinical settings is also provided. We summarize the current challenges from different perspectives and highlight some potential directions for future research. Overall, this study provides a better understanding of the current research on MCN and assists readers in advancing this field.

CRediT authorship contribution statement

Haihua Chen: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Yuhan Zhou:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Ruochi Li:** Writing – original draft, Visualization, Validation, Methodology. **Aryan Murthy Illa:** Writing – original draft, Software, Methodology. **Ana Cleveland:** Writing – review & editing, Data curation. **Junhua Ding:** Writing – review & editing, Funding acquisition, Data curation, Conceptualization.

Declaration of generative AI usage

After completing the paper, we employ ChatGPT-o1 to identify writing typos. Subsequently, manual review and revision are performed to address these typos.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Abbreviations and acronyms used in this article (Table 11)

Data availability

I have shared the data in GitHub.

Table 11
Abbreviations and acronyms used in this article.

Abbreviations	Full name of concepts
ADE	Adverse Drug Effects
ADR	Adverse Drug Reactions
ANNSA	Automated Named Entity Normalization with Self-Attention
BART	Bidirectional and Auto-Regressive Transformers
BC5CDR	BioCreative V Chemical-Disease Relation
BEL	Biomedical Entity Linking
BERT	Bidirectional Encoder Representation from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BioBERT	Biomedical BERT
BioCDR	Biological Chemical-Disease Relations
BioLORD	Biological Literature Object Relation Detection
BioSyn	Biomedical Synonym Disambiguation
CADEC	Computer-Assisted Diagnosis of Emergency Cases
CNN	Convolutional Neural Network
CODER	Clinical Outcome Data Extraction and Reporting
CRF	Conditional Random Fields
EHR	Electronic Health Record
EhrBERT	Electronic Health Record BERT
FSL	Few-Shot Learning
GNN	Graph Neural Network
GCN	Graph Convolutional Networks
GAT	Graph Attention Networks
GRU	Gated Recurrent Units
KNN	K-Nearest Neighbor
KnowCAGE	Knowledgeable Contextualized Adversarial Generation for Entity Linking
LSTM	Long Short-Term Memory
LLMs	Large Language Models
LoRA	Low-Rank Adaptation
MCN	Medical Concept Normalization
MedDRA	Medical Dictionary for Regulatory Activities
MEL	Medical Entity Linking
MTMV-CNN	Multi-View Convolutional Neural Network
NER	Name Entity Recognition
PsyTAR	Psychiatric Treatment Adverse Reactions
RNN	Recurrent Neural Networks
RoBERTa	Robustly optimized BERT approach
SapBERT	Self-alignment Pretraining BERT
SHAP	SHapley Additive exPlanations
ShARe/CLEF	Shared Annotated Resources and Evaluation at the Cross-Language Evaluation Forum
SIDER	Side Effect Resource
SMM4H	Social Media Mining for Health
SNOMED-CT	Systematized Nomenclature of Medicine or Clinical Terms
SVM	Support Vector Machine
TwADR	Twitter Adverse Drug Reactions
UMLS	Unified Medical Language System
ZSL	Zero-Shot Learning

References

[1] K. Lee, S.A. Hasan, O. Farri, A. Choudhary, A. Agrawal, Medical concept normalization for online user-generated texts, in: 2017 IEEE International Conference on Healthcare Informatics, ICHI, IEEE, 2017, pp. 462–469.

[2] N. Pattisapu, V. Anand, S. Patil, G. Palshikar, V. Varma, Distant supervision for medical concept normalization, J. Biomed. Inform. 109 (2020) 103522.

[3] Ö. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko, C. Biemann, Neural entity linking: A survey of models based on deep learning, Semant. Web 13 (3) (2022) 527–570.

[4] E. French, B.T. McInnes, An overview of biomedical entity linking throughout the years, J. Biomed. Inform. 137 (2023) 104252.

[5] X. Jing, The unified medical language system at 30 years and how it is used and published: systematic review and content analysis, JMIR Med. Inform. 9 (8) (2021) e20675.

[6] E. Chang, J. Mostafa, The use of SNOMED CT, 2013–2020: a literature review, J. Am. Med. Informatics Assoc. 28 (9) (2021) 2017–2026.

[7] N. Baumann, How to use the medical subject headings (MeSH)., Int. J. Clin. Pract. 70 (2) (2016) 171–174.

[8] J.E. Harrison, S. Weber, R. Jakob, C.G. Chute, ICD-11: an international classification of diseases for the twenty-first century, BMC Med. Inform. Decis. Mak. 21 (2021) 1–10.

[9] H. Le, R. Chen, S. Harris, H. Fang, B. Lyn-Cook, H. Hong, W. Ge, P. Rogers, W. Tong, W. Zou, Rxnorm for drug name normalization: a case study of prescription opioids in the FDA adverse events reporting system, Front. Bioinform. 3 (2024) 1328613.

[10] R. Zhang, Y.-S. Wang, Y. Yang, D. Yu, T. Vu, L. Lei, Long-tailed extreme multi-label text classification by the retrieval of generated pseudo label descriptions, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1092–1106.

[11] B. Portelli, S. Scaboro, G. Serra, Improving multi-lingual medical term normalization to address the long-tail problem, in: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, 2023, pp. 809–818.

[12] H. Peng, Y. Xiong, Y. Xiang, H. Wang, H. Xu, B. Tang, Biomedical named entity normalization via interaction-based synonym marginalization, J. Biomed. Inform. 136 (2022) 104238.

[13] H. Xu, D.D. Fushman, Natural language processing in biomedicine: A practical guide, Springer, 2024.

[14] H. Chen, R. Li, A. Cleveland, J. Ding, Enhancing data quality in medical concept normalization through large language models, J. Biomed. Inform. 165 (2025) 104812.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 6000–6010.

[16] A. Radford, Improving language understanding by generative pre-training, 2018, URL https://openai.com/index/language-unsupervised/.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[18] J. Shi, Z. Yuan, W. Guo, C. Ma, J. Chen, M. Zhang, Knowledge-graph-enabled biomedical entity linking: a survey, World Wide Web 26 (5) (2023) 2593–2622.

[19] D. Kartchner, J. Deng, S. Lohiya, T. Kopparthi, P. Bathala, D. Domingo-Fernández, C.S. Mitchell, A comprehensive evaluation of biomedical entity linking models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2023, NIH Public Access, 2023, p. 14462.

[20] P. Bose, S. Srinivasan, W.C. Sleeman IV, J. Palta, R. Kapoor, P. Ghosh, A survey on recent named entity recognition and relationship extraction techniques on clinical texts, Appl. Sci. 11 (18) (2021) 8319.

[21] M.-S. Huang, P.-T. Lai, P.-Y. Lin, Y.-T. You, R.T.-H. Tsai, W.-L. Hsu, Biomedical named entity recognition and linking datasets: survey and our recent development, Brief. Bioinform. 21 (6) (2020) 2219–2238.

[22] M. Madkour, D. Benhaddou, C. Tao, Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain, Comput. Methods Programs Biomed. 128 (2016) 52–68.

[23] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270.

[24] K. Donnelly, et al., SNOMED-CT: The advanced terminology and coding system for ehealth, Stud. Health Technol. Inform. 121 (2006) 279.

[25] S. Liu, W. Ma, R. Moore, V. Ganesan, S. Nelson, Rxnorm: prescription for electronic drug information exchange, IT Prof. 7 (5) (2005) 17–23.

[26] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C.P. Rosé, E. Fosler-Lussier, Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets, J. Am. Med. Inform. Assoc. 28 (3) (2021) 516–532.

[27] D. Xu, Z. Zhang, S. Bethard, A generate-and-rank framework with semantic type regularization for biomedical concept normalization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8452–8464.

[28] E.G. Brown, L. Wood, S. Wood, The medical dictionary for regulatory activities (meddra), Drug Saf. 20 (2) (1999) 109–117.

[29] Y.-F. Luo, W. Sun, A. Rumshisky, MCN: a comprehensive corpus for medical concept normalization, J. Biomed. Inform. 92 (2019) 103132.

[30] M. Dusenne, K. Billey, F. Desgrippes, A. Benis, S.J. Darmoni, J. Grosjean, WikiMeSH: multi lingual MeSH translations via wikipedia, in: 32nd Medical Informatics Europe Conference (MIE2022), vol. 294, IOS Press, 2022, pp. 403–404.

[31] J. Li, Y. Sun, R.J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A.P. Davis, C.J. Mattingly, T.C. Wiegers, Z. Lu, BioCreative v CDR task corpus: a resource for chemical disease relation extraction, Database 2016 (2016).

[32] Y. Wang, C. Dima, S. Staab, WikiMed-DE: Constructing a silver-standard dataset for german biomedical entity linking using wikipedia and wikidata, in: Proceedings of the Wikidata Workshop 2023, 2023, pp. 145–151.

[33] F. Hartendorp, T. Seinen, E. van Mulligen, S. Verberne, Biomedical entity linking for dutch: Fine-tuning a self-alignment bert model on an automatically generated wikipedia corpus, in: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024, 2024, pp. 253–263.

[34] N. Limsopatham, N. Collier, Normalising medical concepts in social media texts by learning semantic representation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1014–1023.

[35] P. Ruas, A. Lamurias, F.M. Couto, Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature, J. Cheminform. 12 (2020) 1–11.

[36] R.I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.

[37] P. Wajsbürt, A. Sarfati, X. Tannier, Medical concept normalization in french using multilingual terminologies and contextual embeddings, J. Biomed. Inform. 114 (2021) 103684.

[38] J.A. Kors, S. Clematide, S.A. Akhondi, E.M. Van Mulligen, D. Rebholz-Schuhmann, A multilingual gold-standard corpus for biomedical concept recognition: the mantra GSC, J. Am. Med. Inform. Assoc. 22 (5) (2015) 948–956.

[39] K. Roberts, D. Demner-Fushman, J.M. Tonning, Overview of the TAC 2017 adverse reaction extraction from drug labels track, Theory Appl. Categ. (2017) URL https://api.semanticscholar.org/CorpusID:46894148.

[40] D. Demner-Fushman, S.E. Shooshan, L. Rodriguez, A.R. Aronson, F. Lang, W. Rogers, K. Roberts, J. Tonning, A dataset of 200 structured product labels annotated for adverse drug reactions, Sci. Data 5 (1) (2018) 1–8.

[41] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with umls concepts, in: Proceedings of the 1st Conference on Automated Knowledge Base Construction, AKBC, 2019, pp. 1–13.

[42] A. Miranda, F. Mehryary, J. Luoma, S. Pyysalo, A. Valencia, M. Krallinger, Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations, in: Proceedings of the Seventh BioCreative Challenge Evaluation Workshop, 2021, pp. 11–21.

[43] N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical concept normalization over nested entities with partial umls terminology in Russian, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 2383–2389.

[44] H. Rouhizadeh, I. Nikishina, A. Yazdani, A. Bornet, B. Zhang, J. Ehrsam, C. Gaudet-Blavignac, N. Naderi, D. Teodoro, A dataset for evaluating contextualized representation of biomedical concepts in language models, Sci. Data 11 (1) (2024) 455.

[45] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.

[46] R. Nalichowski, D. Keogh, H.C. Chueh, S.N. Murphy, Calculating the benefits of a research patient data repository, in: AMIA Annual Symposium Proceedings, vol. 2006, 2006, p. 1044.

[47] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D.L. Mowery, S. Velupillai, W.W. Chapman, D. Martinez, G. Zuccon, et al., Overview of the share/clef ehealth evaluation lab 2014, in: Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings 5, Springer, 2014, pp. 172–191.

[48] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, G. Savova, SemEval-2015 task 14: Analysis of clinical text, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 303–310.

[49] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[50] A. Jagannatha, F. Liu, W. Liu, H. Yu, Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0), Drug Saf. 42 (2019) 99–111.

[51] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323.

[52] T. Guan, H. Zan, X. Zhou, H. Xu, K. Zhang, Cmeie: Construction and evaluation of Chinese medical information extraction dataset, in: Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9, Springer, 2020, pp. 270–282.

[53] W.H. Organization, et al., The anatomical therapeutic chemical classification system with defined daily doses-ATC/DDD, 2009.

[54] N. Zhang, M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, K. Yin, C. Tan, J. Xu, F. Huang, et al., Cblue: A chinese biomedical language understanding evaluation benchmark, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 7888–7915.

[55] D. Mahajan, J.J. Liang, C.-H. Tsou, Toward understanding clinical context of medication change events in clinical narratives, in: AMIA Annual Symposium Proceedings, vol. 2021, 2022, p. 833.

[56] A. Nesterov, G. Zubkova, Z. Miftahutdinov, V. Kokh, E. Tutubalina, A. Shelmanov, A. Alekseev, M. Avetisian, A. Chertok, S. Nikolenko, RuCCoN: clinical concept normalization in Russian, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 239–245.

[57] L.E.S.e. Oliveira, A.C. Peters, A.M.P. Da Silva, C.P. Gebeluca, Y.B. Gumiel, L.M.M. Cintho, D.R. Carvalho, S. Al Hasan, C.M.C. Moro, SemClinBr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks, J. Biomed. Semant. 13 (1) (2022) 13.

[58] J. Zaghir, M. Bjelogrlic, J.-P. Goldman, S. Aananou, C. Gaudet-Blavignac, C. Lovis, FRASIMED: A clinical french annotated resource produced through crosslingual BERT-based annotation projection, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 7450–7460.

[59] F. Gaschi, X. Fontaine, P. Rastin, Y. Toussaint, Multilingual clinical NER: Translation or cross-lingual transfer? in: 5th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2023, pp. 289–311.

[60] Y.-F. Luo, S. Henry, Y. Wang, F. Shen, O. Uzuner, A. Rumshisky, The 2019 n2c2/umass lowell shared task on clinical concept normalization, J. Am. Med. Inform. Assoc. 27 (10) (2020) 1529–e1.

[61] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, 2022, Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings.

[62] A.E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, et al., MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023) 1.

[63] S. Lima-López, E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret, M. Krallinger, Overview of symptemist at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text, in: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models, 2023, p. 11.

[64] S. Cao, Q. Wu, MC-TCMNER: A multi-modal fusion model combining contrast learning method for traditional Chinese medicine NER, in: International Conference on Multimedia Modeling, Springer, 2024, pp. 341–354.

[65] A. Nesterov, A. Sakhovskiy, I. Sviridov, A. Valiev, V. Makharev, P. Anokhin, G. Zubkova, E. Tutubalina, RuCCoD: Towards automated ICD coding in Russian, 2025, arXiv preprint arXiv:2502.21263.

[66] A. Sarker, M. Belousov, J. Friedrichs, K. Hakala, S. Kiritchenko, F. Mehryary, S. Han, T. Tran, A. Rios, R. Kavuluru, et al., Data and systems for medication-related text classification and concept normalization from Twitter: insights from the social media mining for health (SMM4H)-2017 shared task, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1274–1283.

[67] M. Zolnoori, K.W. Fung, T.B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola, N.D. Shah, Y.S.S. Wu, C.E. Eldredge, J. Luo, et al., The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications, Data Brief 24 (2019) 103838.

[68] M. Wiatrak, J. Iso-Sipila, Simple hierarchical multi-task neural end-to-end entity linking for biomedical text, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020, pp. 12–17.

[69] S. Scepanovic, E. Martin-Lopez, D. Quercia, K. Baykaner, Extracting medical entities from social media, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 170–181.

[70] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A corpus of adverse drug event annotations, J. Biomed. Inform. 55 (2015) 73–81.

[71] N. Alvaro, Y. Miyao, N. Collier, et al., TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations, JMIR Public Health Surveill. 3 (2) (2017) e6396.

[72] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. Paul, G. Gonzalez, Overview of the fourth social media mining for health (SMM4h) shared tasks at ACL 2019, in: Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, 2019, pp. 21–30.

[73] R. Biddle, A. Joshi, S. Liu, C. Paris, G. Xu, Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter, in: Proceedings of the Web Conference 2020, 2020, pp. 1217–1227.

[74] M. Belousov, W.G. Dixon, G. Nenadic, Mednorm: A corpus and embeddings for cross-terminology medical concept normalisation, in: Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, 2019, pp. 31–39.

[75] M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A corpus for medical entity linking in the social media, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 3122–3137.

[76] A. Magge, A. Klein, A. Miranda-Escalada, M.A. Al-Garadi, I. Alimova, Z. Miftahutdinov, E. Farre, S. Lima-López, I. Flores, K. O'Connor, et al., Overview of the sixth social media mining for health applications (# SMM4h) shared tasks at NAACL 2021, in: Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021, pp. 21–32.

[77] U. Naseem, M. Khushi, J. Kim, A.G. Dunn, RHMD: a real-world dataset for health mention classification on reddit, IEEE Trans. Comput. Soc. Syst. 10 (5) (2022) 2325–2334.

[78] L.G. Sánchez, D.E. Zavala, E. Farré-Maduell, S. Lima-López, A. Miranda-Escalada, M. Krallinger, The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora, in: Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, 2022, pp. 182–189.

[79] O.T. Aduragba, J. Yu, A. Cristea, Y. Long, Improving health mention classification through emphasising literal meanings: A study towards diversity and generalisation for public health surveillance, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3928–3936.

[80] H. Chen, J. Chen, J. Ding, Data evaluation and enhancement for quality improvement of machine learning, IEEE Trans. Reliab. 70 (2) (2021) 831–847.

[81] Y. Zhou, F. Tu, K. Sha, J. Ding, H. Chen, A survey on data quality dimensions and tools for machine learning invited paper, in: 2024 IEEE International Conference on Artificial Intelligence Testing (AITest), IEEE, 2024, pp. 120–131.

[82] E. Tutubalina, A. Kadurin, Z. Miftahutdinov, Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6710–6716.

[83] T. Xia, T. Dang, J. Han, L. Qendro, C. Mascolo, Uncertainty-aware health diagnostics via class-balanced evidential deep learning, IEEE J. Biomed. Health Inform. 28 (11) (2024) 6417–6428.

[84] A. Zink, Z. Obermeyer, E. Pierson, Race adjustments in clinical algorithms can help correct for racial disparities in data quality, Proc. Natl. Acad. Sci. 121 (34) (2024) e2402267121.

[85] M. Kulyabin, G. Sokolov, A. Galaida, A. Maier, T. Arias-Vergara, SNOBERT: A benchmark for clinical notes entity linking in the SNOMED CT clinical terminology, in: International Conference on Pattern Recognition, Springer, 2025, pp. 154–163.

[86] H. Zong, R. Wu, J. Cha, W. Feng, E. Wu, J. Li, A. Shao, L. Tao, Z. Li, B. Tang, et al., Advancing Chinese biomedical text mining with community challenges, J. Biomed. Inform. (2024) 104716.

[87] F. Borchert, I. Llorca, R. Roller, B. Arnrich, M.-P. Schapranow, xMEN: a modular toolkit for cross-lingual medical entity normalization, JAMIA Open 8 (1) (2025) ooae147.

[88] T. Mota, M.R. Verdelho, D.J. Araújo, A. Bissoto, C. Santiago, C. Barata, MMIST-ccrcc: A real world medical dataset for the development of multi-modal systems, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024, pp. 2395–2403.

[89] Y. Wu, J. Chen, L. Hu, H. Xu, H. Liang, J. Wu, OmniFuse: A general modality fusion framework for multi-modality learning on low-quality medical data, Inf. Fusion 117 (2025) 102890.

[90] X. Xu, J. Li, Z. Zhu, L. Zhao, H. Wang, C. Song, Y. Chen, Q. Zhao, J. Yang, Y. Pei, A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis, Bioengineering 11 (3) (2024) 219.

[91] H. Li, Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang, D. Huang, CNN-based ranking for biomedical entity normalization, BMC Bioinformatics 18 (2017) 79–86.

[92] R. Roller, M. Kittner, D. Weissenborn, U. Leser, Cross-lingual candidate search for biomedical concept normalization, MultilingualBIO: Multiling. Biomed. Text Process. (2018) 16.

[93] Y. Luo, G. Song, P. Li, Z. Qi, Multi-task medical concept normalization using multi-view convolutional neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (1) 2018, pp. 5868–5875.

[94] G. Song, Q. Long, Y. Luo, Y. Wang, Y. Jin, Deep convolutional neural network based medical concept normalization, IEEE Trans. Big Data 8 (5) (2020) 1195–1208.

[95] M. Tiftikci, A. Özgür, Y. He, J. Hur, Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels, BMC Bioinformatics 20 (2019) 1–9.

[96] Y. An, J. Wang, L. Zhang, H. Zhao, Z. Gao, H. Huang, Z. Du, Z. Jiao, J. Yan, X. Wei, et al., PASCAL: a pseudo cascade learning framework for breast cancer treatment entity normalization in Chinese clinical text, BMC Med. Inform. Decis. Mak. 20 (2020) 1–12.

[97] Y. Luo, Y. Tian, M. Shi, L.R. Pasquale, L.Q. Shen, N. Zebardast, T. Elze, M. Wang, Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization, IEEE Trans. Med. Imaging (2024).

[98] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, V. Malykh, Medical concept normalization in social media posts with recurrent neural networks, J. Biomed. Inform. 84 (2018) 93–102.

[99] K.K. Subramanyam, S. Sangeetha, Deep contextualized medical concept normalization in social media text, Procedia Comput. Sci. 171 (2020) 1353–1362.

[100] A.M. Ningtyas, A. Hanbury, F. Piroi, L. Andersson, Data augmentation for layperson's medical entity linking task, in: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, 2021, pp. 99–106.

[101] Y.-F. Luo, W. Sun, A. Rumshisky, A hybrid method for normalization of medical concepts in clinical narrative, in: 2018 IEEE International Conference on Healthcare Informatics, ICHI, IEEE, 2018, pp. 392–393.

[102] Y.-F. Luo, W. Sun, A. Rumshisky, A hybrid normalization method for medical concepts in clinical narrative using semantic matching, AMIA Summits Transl. Sci. Proc. 2019 (2019) 732.

[103] D. Loureiro, A.M. Jorge, Medlinker: Medical entity linking with neural representations and dictionary matching, in: European Conference on Information Retrieval, Springer, 2020, pp. 230–237.

[104] J. Wang, N. Abu-el Rub, J. Gray, H.A. Pham, Y. Zhou, F.J. Manion, M. Liu, X. Song, H. Xu, M. Rouhizadeh, et al., COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model, J. Am. Med. Inform. Assoc. 28 (6) (2021) 1275–1283.

[105] A.M. Ningtyas, A. El-Ebshihy, G.B. Herwanto, F. Piroi, A. Hanbury, Leveraging wikipedia knowledge for distant supervision in medical concept normalization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022, pp. 33–47.

[106] Y. Bitton, R. Cohen, T. Schifter, E. Bachmat, M. Elhadad, N. Elhadad, Cross-lingual unified medical language system entity linking in online health communities, J. Am. Med. Inform. Assoc. 27 (10) (2020) 1585–1592.

[107] P. Han, X. Li, Z. Zhang, Y. Zhong, L. Gu, Y. Hua, X. Li, CMCN: Chinese medical concept normalization using continual learning and knowledge-enhanced, Artif. Intell. Med. 157 (2024) 102965.

[108] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[109] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3615–3620.

[110] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019, arXiv preprint arXiv:1904.05342.

[111] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Heal. (HEALTH) 3 (1) (2021) 1–23.

[112] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, BioNLP 2019 (2019) 58.

[113] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, A. Wong, UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 1744–1753.

[114] F. Li, Y. Jin, W. Liu, B.P.S. Rawat, P. Cai, H. Yu, et al., Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study, JMIR Med. Inform. 7 (3) (2019) e14830.

[115] K.S. Kalyan, S. Sangeetha, Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network, Artif. Intell. Med. 112 (2021) 102008.

[116] T. Tsujimura, M. Miwa, Y. Sasaki, Large-scale neural biomedical entity linking with layer overwriting, J. Biomed. Inform. 143 (2023) 104433.

[117] Q. Wang, Z. Ji, J. Wang, S. Wu, W. Lin, W. Li, L. Ke, G. Xiao, Q. Jiang, H. Xu, et al., A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes, J. Biomed. Inform. 105 (2020) 103418.

[118] Z. Ji, Q. Wei, H. Xu, Bert-based ranking for biomedical entity normalization, AMIA Summits Transl. Sci. Proc. 2020 (2020) 269.

[119] H. Cho, D. Choi, H. Lee, Re-ranking system with BERT for biomedical concept normalization, IEEE Access 9 (2021) 121253–121262.

[120] M.G. Sohrab, K. Duong, M. Miwa, G. Topić, I. Masami, T. Hiroya, BENNERD: A neural named entity linking system for COVID-19, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 182–188.

[121] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238.

[122] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3641–3650.

[123] Z. Miftahutdinov, A. Kadurin, R. Kudrin, E. Tutubalina, Drug and disease interpretation learning with biomedical entity representation transformer, in: European Conference on Information Retrieval, Springer, 2021, pp. 451–466.

[124] Z. Miftahutdinov, A. Kadurin, R. Kudrin, E. Tutubalina, Medical concept normalization in clinical trials with drug and disease representation learning, Bioinformatics 37 (21) (2021) 3856–3864.

[125] S. Zhang, H. Cheng, S. Vashishth, C. Wong, J. Xiao, X. Liu, T. Naumann, J. Gao, H. Poon, Knowledge-rich self-supervision for biomedical entity linking, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 868–880.

[126] Q. Li, G. Wu, T. You, Bioelm: Integrating biomedical knowledge into language model with entity-linking, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 763–766.

[127] C. Yan, Y. Zhang, K. Liu, J. Zhao, Y. Shi, S. Liu, Biomedical concept normalization by leveraging hypernyms, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3512–3517.

[128] Q. Jia, D. Zhang, S. Yang, C. Xia, Y. Shi, H. Tao, C. Xu, X. Luo, Y. Ma, Y. Xie, Traditional Chinese medicine symptom normalization approach leveraging hierarchical semantic information and text matching with attention mechanism, J. Biomed. Inform. 116 (2021) 103718.

[129] J. Yan, Y. Wang, L. Xiang, Y. Zhou, C. Zong, A knowledge-driven generative model for multi-implication chinese medical procedure entity normalization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 1490–1499.

[130] C. Kim, H. Kim, S. Park, J. Lee, M. Sung, J. Kang, Learning from negative samples in generative biomedical entity linking, 2024, arXiv preprint arXiv: 2408.16493.

[131] T. Zhu, Y. Qin, M. Feng, Q. Chen, B. Hu, Y. Xiang, Biopro: Context-infused prompt learning for biomedical entity linking, IEEE/ACM Trans. Audio, Speech, Lang. Process. 32 (2023) 374–385.

[132] G. Zhang, X. Peng, T. Shen, G. Long, J. Si, L. Qin, W. Lu, Extractive medical entity disambiguation with memory mechanism and memorized entity information, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 13811–13822.

[133] S. Garda, U. Leser, BELHD: improving biomedical entity linking with homonym disambiguation, Bioinformatics 40 (8) (2024) btae474.

[134] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 565–574.

[135] Y.-C. Lin, P. Hoffmann, E. Rahm, Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models, SN Comput. Sci. 3 (5) (2022) 387.

[136] A. Alekseev, Z. Miftahutdinov, E. Tutubalina, A. Shelmanov, V. Ivanov, V. Kokh, A. Nesterov, M. Avetisian, A. Chertok, S. Nikolenko, Medical crossing: a cross-lingual evaluation of clinical entity linking, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 4212–4220.

[137] L. Chen, Y. Qi, A. Wu, L. Deng, T. Jiang, Teabert: An efficient knowledge infused cross-lingual language model for mapping Chinese medical entities to the unified medical language system, IEEE J. Biomed. Health Inform. (2023).

[138] T. Zhu, Y. Qin, Q. Chen, X. Mu, C. Yu, Y. Xiang, Controllable contrastive generation for multilingual biomedical entity linking, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5742–5753.

[139] F. Gallego, G. López-García, L. Gasco-Sánchez, M. Krallinger, F.J. Veredas, Clinlinker: Medical entity linking of clinical concept mentions in spanish, in: International Conference on Computational Science, Springer, 2024, pp. 266–280.

[140] B. Zhou, X. Cai, Y. Zhang, X. Yuan, An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 6214–6224.

[141] J. Noh, R. Kavuluru, Joint learning for biomedical NER and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation, in: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2021, pp. 1–10.

[142] A. Magge, E. Tutubalina, Z. Miftahutdinov, I. Alimova, A. Dirkson, S. Verberne, D. Weissenbacher, G. Gonzalez-Hernandez, DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter, J. Am. Med. Inform. Assoc. 28 (10) (2021) 2184–2192.

[143] Q. Dai, J. Zhong, C. Wang, H. Yin, Q. Lei, X. Li, R. Li, Joint learning-based multiple documents heterogeneous graph inference for biomedical entity linking, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2023, pp. 2967–2974.

[144] D. D'Auria, V. Moscato, M. Postiglione, G. Romito, G. Sperlí, Improving graph embeddings via entity linking: a case study on Italian clinical notes, Intell. Syst. Appl. 17 (2023) 200161.

[145] M. Bo, M. Zhang, Learning dynamic coherence with graph attention network for biomedical entity linking, in: 2021 International Joint Conference on Neural Networks, IJCNN, IEEE, 2021, pp. 1–8.

[146] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Graph-enriched biomedical entity representation transformer, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 109–120.

[147] R. Angell, N. Monath, S. Mohan, N. Yadav, A. McCallum, Clustering-based inference for biomedical entity linking, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2598–2608.

[148] H. Xu, J. Zhang, Z. Wang, S. Zhang, M. Bhalerao, Y. Liu, D. Zhu, S. Wang, GraphPrompt: Graph-based prompt templates for biomedical synonym prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (9) 2023, pp. 10576–10584.

[149] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Biomedical entity representation with graph-augmented multi-objective transformer, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 4626–4643.

[150] Y. Gao, S. Ji, P. Marttinen, Knowledge-augmented graph neural networks with concept-aware attention for adverse drug event detection, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 9787–9798.

[151] A. Wang, C. Liu, J. Yang, C. Weng, Fine-tuning large language models for rare disease concept normalization, J. Am. Med. Inform. Assoc. (2024) ocae133.

[152] S. Liu, A. Wang, X. Xiu, M. Zhong, S. Wu, et al., Evaluating medical entity recognition in health care: Entity model quantitative study, JMIR Med. Inform. 12 (1) (2024) e59782.

[153] M. Hosseini, M. Hosseini, R. Javidan, Leveraging large language models for clinical abbreviation disambiguation, J. Med. Syst. 48 (1) (2024) 27.

[154] F. Remy, K. Demuynck, T. Demeester, Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights, J. Am. Med. Inform. Assoc. (2024) ocae029.

[155] A. Sakor, K. Singh, M.-E. Vidal, BioLinkerAI: Capturing knowledge using LLMs to enhance biomedical entity linking, in: International Conference on Web Information Systems Engineering, Springer, 2024, pp. 262–272.

[156] A. Abdulnazar, R. Roller, S. Schulz, M. Kreuzthaler, Large language models for clinical text cleansing enhance medical concept normalization, IEEE Access (2024).

[157] F. Borchert, I. Llorca, M.-P. Schapranow, Improving biomedical entity linking for complex entity mentions with LLM-based text simplification, Database 2024 (2024) baae067.

[158] D.L. Mowery, B.R. South, L. Christensen, J. Leng, L.-M. Peltonen, S. Salanterä, H. Suominen, D. Martinez, S. Velupillai, N. Elhadad, et al., Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF ehealth challenge 2013, task 2, J. Biomed. Semant. 7 (2016) 1–13.

[159] R.J. Kate, Normalizing clinical terms using learned edit distance patterns, J. Am. Med. Informatics Assoc. 23 (2) (2016) 380–386.

[160] R.J. Kate, Clinical term normalization using learned edit patterns and subconcept matching: system development and evaluation, JMIR Med. Inform. 9 (1) (2021) e23104.

[161] E. Emadzadeh, A. Sarker, A. Nikfarjam, G. Gonzalez, Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology, in: AMIA Annual Symposium Proceedings, vol. 2017, 2018, p. 679.

[162] H. Cho, W. Choi, H. Lee, A method for named entity normalization in biomedical articles: application to diseases and plants, BMC Bioinformatics 18 (2017) 1–12.

[163] I. Karadeniz, A. Özgür, Linking entities through an ontology using word embeddings and syntactic re-ranking, BMC Bioinformatics 20 (2019) 1–12.

[164] T.M. Seinen, J.A. Kors, E.M. van Mulligen, P.R. Rijnbeek, Annotation-preserving machine translation of english corpora to validate dutch clinical concept extraction tools, J. Am. Med. Inform. Assoc. 31 (8) (2024) 1725–1734.

[165] J.R. Almeida, J.L. Oliveira, Multi-language concept normalisation of clinical cohorts, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2020, pp. 261–264.

[166] J. Dai, M. Zhang, G. Chen, J. Fan, K.Y. Ngiam, B.C. Ooi, Fine-grained concept linking using neural networks in healthcare, in: Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 51–66.

[167] S. Yang, P. Zhang, C. Che, Z. Zhong, B-lbcona: a medical entity disambiguation model based on bio-linkbert and context-aware mechanism, BMC Bioinformatics 24 (1) (2023) 97.

[168] H. Dong, J. Chen, Y. He, Y. Liu, I. Horrocks, Reveal the unknown: out-of-knowledge-base mention discovery with entity linking, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 452–462.

[169] L. Chen, G. Varoquaux, F.M. Suchanek, A lightweight neural model for biomedical entity linking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, (14) 2021, pp. 12657–12665.

[170] Y. Fan, Y. Zhu, K. Xue, J. Liu, T. Ruan, Rrnorm: A novel framework for Chinese disease diagnoses normalization via LLM-driven terminology component recognition and reconstruction, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 9162–9175.

[171] Y. Zhang, K. Zhong, G. Liu, A novel method for medical semantic word sense disambiguation by using graph neural network, in: 2023 9th International Symposium on System Security, Safety, and Reliability, ISSSR, IEEE, 2023, pp. 263–272.

[172] M. Gu, W. Cui, X. Fu, BioSentEval: An evaluation toolkit for Chinese medical sentence representation, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 3107–3111.

[173] H. Yuan, Z. Yuan, S. Yu, Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 4038–4048.
[174] D. Wright, NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction, University of California, San Diego, 2019.
[175] Z. Lin, Z. Zhang, X. Wu, Y. Zheng, Biomedical entity linking as multiple choice question answering, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 2390–2396.
[176] T. Zhu, Y. Qin, Q. Chen, B. Hu, Y. Xiang, Enhancing entity representations with prompt learning for biomedical entity linking, in: IJCAI, 2022, pp. 4036–4042.
[177] S. Hussain, H. Afzal, R. Saeed, N. Iltaf, M.Y. Umair, Pharmacovigilance with transformers: A framework to detect adverse drug reactions using BERT fine-tuned with FARM, Comput. Math. Methods Med. 2021 (1) (2021) 5589829.
[178] O. Elbiach, H. Grissette, et al., Adverse drug reactions detection from social media: an empirical evaluation of machine learning techniques, in: 2023 14th International Conference on Intelligent Systems: Theories and Applications, SITA, IEEE, 2023, pp. 1–7.
[179] Y. Qiu, X. Zhang, W. Wang, H. Lin, Dual sentiment-aware networks for adverse drug reactions detection, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2024, pp. 2363–2368.
[180] T. Zhang, H. Lin, Y. Ren, L. Yang, B. Xu, Z. Yang, J. Wang, Y. Zhang, Adverse drug reaction detection via a multihop self-attention mechanism, BMC Bioinformatics 20 (2019) 1–11.
[181] F. Remy, S. Scaboro, B. Portelli, Boosting adverse drug event normalization on social media: General-purpose model initialization and biomedical semantic text similarity benefit zero-shot linking in informal contexts, in: Proceedings of the 11th International Workshop on Natural Language Processing for Social Media, 2023, pp. 47–52.
[182] Z. Lin, Z. Zhang, X. Wu, Y. Zheng, Improving biomedical entity linking with retrieval-enhanced learning, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 11461–11465.
[183] S. Zhang, L. He, D. Wang, H. Bao, S. Zheng, Y. Liu, B. Xiao, J. Li, D. Lu, N. Zheng, ProSyno: context-free prompt learning for synonym discovery, Front. Comput. Sci. 19 (6) (2025) 1–9.
[184] J. Niu, Y. Yang, S. Zhang, Z. Sun, W. Zhang, Multi-task character-level attentional networks for medical concept normalization, Neural Process. Lett. 49 (2019) 1239–1256.
[185] Y. Cao, L. Fang, Z. Zheng, Enriching pre-trained language model with multi-task learning and context for medical concept normalization, in: Proceedings of the 2022 International Conference on Intelligent Medicine and Health, 2022, pp. 79–83.
[186] B. Portelli, S. Scaboro, E. Santus, H. Sedghamiz, E. Chersoni, G. Serra, Generalizing over long tail concepts for medical term normalization, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 8580–8591.
[187] M. Chizhikova, J. Collado-Montañez, P. López-Úbeda, M.C. Díaz-Galiano, L.A.U. López, M.T.M. Valdivia, SINAI at CLEF 2022: Leveraging biomedical transformers to detect and normalize disease mentions., in: CLEF (Working Notes), 2022, pp. 265–273.
[188] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Transformers for clinical coding in spanish, IEEE Access 9 (2021) 72387–72397.
[189] N. Loukachevitch, S. Manandhar, E. Baral, I. Rozhkov, P. Braslavski, V. Ivanov, T. Batura, E. Tutubalina, NEREL-BIO: A dataset of biomedical abstracts annotated with nested named entities, Bioinformatics 39 (4) (2023) btad161.
[190] U. Naseem, J. Kim, M. Khush, A.G. Dunn, A linguistic grounding-infused contrastive learning approach for health mention classification on social media, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 529–537.
[191] P.I. Khan, M.N. Asim, A. Dengel, S. Ahmed, A unique training strategy to enhance language models capabilities for health mention detection from social media content, 2023, arXiv preprint arXiv:2310.19057.
[192] O.T. Aduragba, J. Yu, A. Cristea, Y. Long, Improving health mention classification through emphasising literal meanings: A study towards diversity and generalisation for public health surveillance, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3928–3936.
[193] Z. Xu, Y. Chen, B. Hu, Improving biomedical entity linking with cross-entity interaction, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (11) 2023, pp. 13869–13877.
[194] X. Sui, Y. Zhang, X. Cai, K. Song, B. Zhou, X. Yuan, W. Zhang, Biofeg: Generate latent features for biomedical entity linking, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 11584–11593.
[195] M. Jabreel, End-to-end neural coder for tumor named entity recognition., in: IberLEF@ SEPLN, 2020, pp. 358–367.
[196] S. Sood, H. Imran, Information retrieval and query expansion for biomedical data, in: Text Mining Approaches for Biomedical Data, Springer, 2024, pp. 193–235.
[197] N.J. Dobbins, Generalizable and scalable multistage biomedical concept normalization leveraging large language models, Res. Synth. Methods (2024) 1–12.
[198] M. Sänger, S. Garda, X.D. Wang, L. Weber-Genzel, P. Droop, B. Fuchs, A. Akbik, U. Leser, HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools, Bioinformatics 40 (10) (2024) btae564.
[199] A. Ermshaus, M. Piechotta, G. Rüter, U. Keilholz, U. Leser, M. Benary, Preon: Fast and accurate entity normalization for drug names and cancer types in precision oncology, Bioinformatics 40 (3) (2024) btae085.
[200] M. Cannon, J. Stevenson, K. Kuzma, S. Kiwala, J.L. Warner, O.L. Griffith, M. Griffith, A.H. Wagner, Normalization of drug and therapeutic concepts with thera-py, JAMIA Open 6 (4) (2023) ooad093.
[201] J. Fries, L. Weber, N. Seelam, G. Altay, D. Datta, S. Garda, S. Kang, R. Su, W. Kusa, S. Cahyawijaya, et al., Bigbio: A framework for data-centric biomedical natural language processing, Adv. Neural Inf. Process. Syst. 35 (2022) 25792–25806.
[202] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, J. Kang, BERN2: an advanced neural biomedical named entity recognition and normalization tool, Bioinformatics 38 (20) (2022) 4837–4839.
[203] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3641–3650.
[204] W. Yoon, R. Jackson, E. Ford, V. Poroshin, J. Kang, Biomedical NER for the enterprise with distillated BERN2 and the kazu framework, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2022, pp. 619–626.
[205] L. Chen, W. Fu, Y. Gu, Z. Sun, H. Li, E. Li, L. Jiang, Y. Gao, Y. Huang, Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking, J. Am. Med. Inform. Assoc. 27 (10) (2020) 1576–1584.
[206] Á. García-Barragán, A. Sakor, M.-E. Vidal, E. Menasalvas, J.C.S. Gonzalez, M. Provencio, V. Robles, NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes, Med. Biol. Eng. Comput. (2024) 1–24.
[207] K. Gueta, G. Chen, "You have to start normalizing": Identity construction among self-changers and treatment changers in the context of drug use normalization, Soc. Sci. Med. 275 (2021) 113828.
[208] E. Manousogiannis, S. Mesbah, A. Bozzon, R.-J. Sips, Z. Szlanik, S.B. Santamaría, Normalization of long-tail adverse drug reactions in social media, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020, pp. 49–58.
[209] A. Sboev, R. Rybka, A. Gryaznov, I. Moloshnikov, S. Sboeva, G. Rylkov, A. Selivanov, Adverse drug reaction concept normalization in Russian-language reviews of internet users, Big Data Cogn. Comput. 6 (4) (2022) 145.
[210] P.I. Khan, A. Dengel, S. Ahmed, Improving disease detection from social media text via self-augmentation and contrastive learning, 2024, arXiv preprint arXiv:2405.01597.
[211] H. Chen, J. Ding, J. Chen, G. Cao, Designing a novel framework for precision medicine information retrieval, in: International Conference on Smart Health, Springer, 2018, pp. 167–178.
[212] S. Frihat, N. Fuhr, Integration of biomedical concepts for enhanced medical literature retrieval, Int. J. Data Sci. Anal. (2025) 1–14.
[213] L. Tamine, L. Goeuriot, Semantic information retrieval on medical texts: Research challenges, survey, and open issues, ACM Comput. Surv. 54 (7) (2021) 1–38.
[214] N. Limsopatham, N. Collier, Normalising medical concepts in social media texts by learning semantic representation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1014–1023.
[215] S. Matos, J.P. Arrais, J. Maia-Rodrigues, J.L. Oliveira, Concept-based query expansion for retrieving gene related publications from MEDLINE, BMC Bioinformatics 11 (2010) 1–9.
[216] Z. Liu, Y. Quan, X. Lyu, M.J. Alenazi, Enhancing clinical accuracy of medical chatbots with large language models, IEEE J. Biomed. Health Inform. (2024).
[217] S. Zhang, J. Song, A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model, Sci. Rep. 14 (1) (2024) 17118.
[218] A. Azam, Z. Naz, M.U.G. Khan, CancerBot: A retrieval-augmented generation based cancer chatbot using large language models, in: 2024 18th International Conference on Open Source Systems and Technologies, ICOSST, IEEE, 2024, pp. 1–6.
[219] B. Pandey, D.K. Pandey, B.P. Mishra, W. Rhmann, A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions, J. King Saud Univ. Comput. Inf. Sci. 34 (8) (2022) 5083–5099.
[220] J. Wu, H. Dong, Z. Li, H. Wang, R. Li, A. Patra, C. Dai, W. Ali, P. Scordis, H. Wu, A hybrid framework with large language models for rare disease phenotyping, BMC Med. Inform. Decis. Mak. 24 (1) (2024) 289.
[221] E. Hassan, T. Abd El-Hafeez, M.Y. Shams, Optimizing classification of diseases through language model analysis of symptoms, Sci. Rep. 14 (1) (2024) 1507.

[222] J.S. Berkowitz, Y. Fatapour, J. Miguel Acitores Cortina, A. Srinivasan, N.P. Tatonetti, Biomedical text normalization through generative modeling, MedRxiv (2024) 2009–2024.
[223] W. Cui, X. Fu, S. Liu, M. Gu, X. Liu, J. Wu, I. King, Data augmentation techniques for Chinese disease name normalization, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2024, pp. 3134–3137.
[224] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 462–477.
[225] A. Abaskohi, S. Rothe, Y. Yaghoobzadeh, LM-CPPF: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023, pp. 670–681.
[226] A. Latif, J. Kim, Evaluation and analysis of large language models for clinical text augmentation and generation, IEEE Access 12 (2024) 48987–48996.
[227] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D.C. Comeau, et al., Opportunities and challenges for ChatGPT and large language models in biomedicine and health, Brief. Bioinform. 25 (1) (2024) bbad493.
[228] J. Li, Y. Li, Y. Pan, J. Guo, Z. Sun, F. Li, Y. He, C. Tao, Mapping vaccine names in clinical trials to vaccine ontology using cascaded fine-tuned domain-specific language models, J. Biomed. Semant. 15 (1) (2024) 14.
[229] T. Patzelt, Medical concept normalization in a low-resource setting, 2024, arXiv preprint arXiv:2409.14579.
[230] H. Rouhizadeh, A. Yazdani, B. Zhang, D. Vicente Alvarez, M. Hueser, A. Vanobberghen, R. Yang, I. Li, A. Walter, D. Teodoro, Large language models struggle to encode medical concepts-a multilingual benchmarking and comparative analysis, MedRxiv (2025) 2001–2025.
[231] S. Shi, Z. Xu, B. Hu, M. Zhang, Generative multimodal entity linking, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 7654–7665.
[232] C.-D.T. Nguyen, X. Wu, T.T. Nguyen, S. Zhao, K.M. Le, N.V. Anh, F. Yichao, L.A. Tuan, Enhancing multimodal entity linking with jaccard distance-based conditional contrastive learning and contextual visual augmentation, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 6695–6708.