Journal of Biomedical Informatics 182 (2026) 105093

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Quality-aware multi-source data fusion and enhancement for Medical Concept Normalization using Large Language Models☆

Yuhan Zhou [a], Ruochi Li [b], Ana Cleveland [a], Junhua Ding [c], Kewei Sha [c], Haihua Chen [c,*]

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

ARTICLE INFO

Dataset link: GitHub Repository

Keywords:
Medical Concept Normalization (MCN)
Data quality
Data augmentation
Multi-source dataset
Data fusion
Large language model

ABSTRACT

Objective: Medical Concept Normalization (MCN) maps informal health phrases to formal clinical concepts. It is a critical task for pharmacovigilance, patient record analysis, and health-related text mining. Extensive MCN research has mainly relied on single-source datasets and overlooked data quality (DQ) issues. This study aims to develop a quality-aware data fusion framework for MCN using Large Language Models (LLMs).
Methods: The methods consist of DQ evaluation and enhancement, LLM-based data augmentation, and multi-source MCN dataset fusion using enriched concept-phrase pairs. We analyze six widely used MCN datasets—AskAPatient, CADEC, COMETA, PsyTAR, TwADR-S, and TwiMed—each collected from social media and mapped to SNOMED-CT. Our evaluation metrics include correctness, concept validity, coverage, semantic variety, and class imbalance. For DQ enhancement and data augmentation, we use Gemini for zero-shot and few-shot learning to increase the semantic variety and phrase count for rare concepts. We then perform data fusion based on shared medical concepts.
Results: The DQ evaluation reveals substantial issues, including incorrect mappings, invalid concepts, low-variety redundant phrases, and long-tail concept-phrase distribution. After augmentation, phrase counts increase by 172.1%, and by 450.0% after fusion. To directly investigate model performance improvement on rare cases, we introduce concept-level macro metrics. SapBERT, KNN-BioEL, and KRISSBERT trained on the augmented and fused dataset achieve significant gains in accuracy, recall, precision, and F1 over single-source baselines, up to 38%.
Conclusion: Our study finds that existing MCN benchmarks present data quality issues and underexplored data fusion potential. Data quality enhancement and LLM-based controlled-variety data augmentation help alleviate overlapping phrases and long-tail issues. Moreover, quality-aware data fusion can expand conceptual coverage, improving MCN performance. This work highlights data quality evaluation and fusion strategies are effective in advancing MCN. We hope these contributions could support quality-guided large-scale MCN data generation with minimal label costs, strengthen reliable biomedical text mining and downstream applications.
Availability: https://github.com/yhZHOU515/DataFusion4MCN.

1. Introduction

Medical Concept Normalization (MCN) is the task of mapping informal expressions of health conditions, symptoms, and drug effects to standardized medical concepts defined in controlled vocabularies such as the Unified Medical Language System (UMLS) and SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms) [1]. By aligning patient-generated language with formal clinical terminology, MCN enables large-scale healthcare analytics, particularly in settings where health information is extracted from noisy, informal text sources such as social media and online forums [2].

The lack of a comprehensive data quality (DQ) evaluation represents a critical research gap in MCN [3]. Extensive evidence from machine learning research shows that data quality factors—such as correctness, coverage, semantic diversity, and class imbalance—have a direct and

☆ This research is partially supported by the National Science Foundation (NSF) grant #2225229.
* Corresponding author.
E-mail addresses: yuhanzhou@my.unt.edu (Y. Zhou), rli14@ncsu.edu (R. Li), ana.cleveland@unt.edu (A. Cleveland), junhua.ding@unt.edu (J. Ding), kewei.sha@unt.edu (K. Sha), haihua.chen@unt.edu (H. Chen).

https://doi.org/10.1016/j.jbi.2026.105093
Received 27 January 2026; Received in revised form 16 July 2026; Accepted 13 August 2026
Available online 21 August 2026

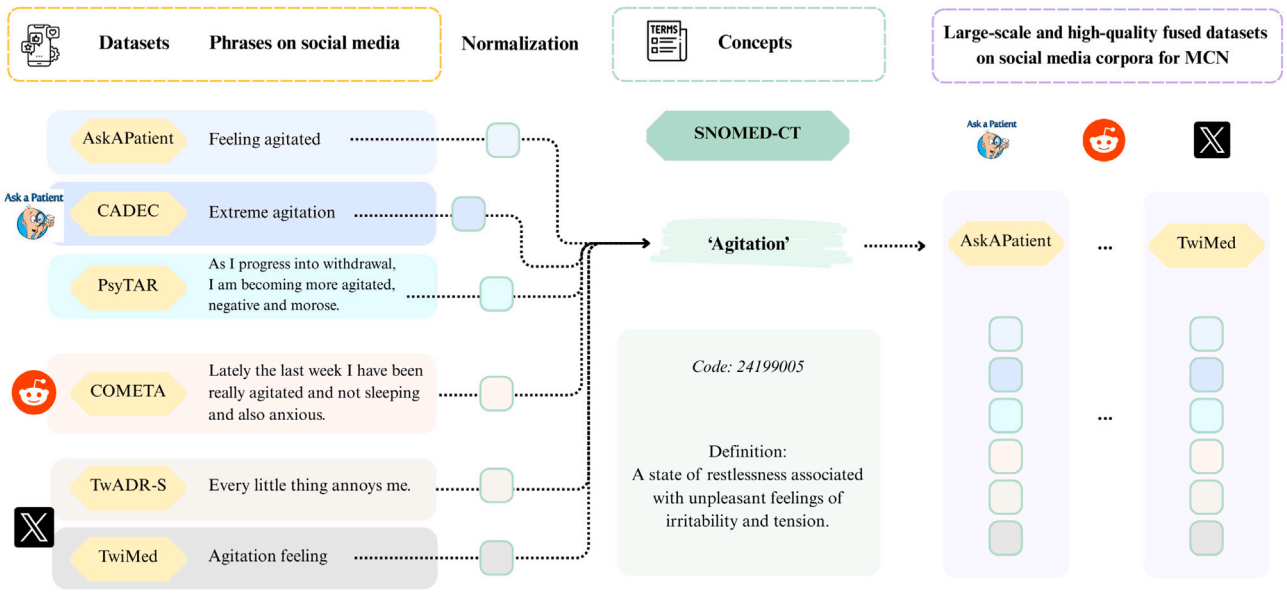


Fig. 1. Illustration of quality-aware multi-source data fusion and enhancement framework for MCN, using one SNOMED-CT concept 'Agitation' as an example. Three datasets are sourced from a patient forum, AskAPatient, and the other three are sourced from general social media platforms. Under normalization, each colored square denotes the example phrase-concept pair. Since these datasets share this concept, after data quality enhancement and fusion, each fused dataset will be enriched with all the phrases mapped to this concept from other datasets, as shown on the right.

often decisive impact on model performance and generalization [4]. In healthcare NLP, low-quality data can lead to unstable predictions and biased decision signals, particularly because patient expressions frequently diverge from standardized clinical terminology [5,6]. Several MCN datasets have already been reported to suffer from such issues. For example, AskAPatient has been shown to exhibit severe redundancy between training and test splits, leading to inflated performance estimates and model overfitting [1,7]. Other datasets, including MedNorm, COMETA, and MIMIC-IV-derived resources, demonstrate pronounced class imbalance, meaning medical concepts with limited phrases dominate the datasets [8–11]. A systematic, multi-dataset data quality evaluation across the MCN landscape is still missing, limiting both dataset enhancement and diagnostic insight.

Another gap lies in the exclusive reliance on single-source datasets for MCN model training and evaluation. Most MCN corpora are constructed from specific data sources, such as patient forums, social media platforms, or biomedical literature, each reflecting a narrow linguistic and contextual scope. Prior studies in biomedical NLP and information extraction suggest that multi-source corpora offer richer semantic diversity and improved generalization by capturing complementary expression patterns across domains and platforms [12–14]. Yet, MCN research has not fully exploited this potential. Consequently, semantic complementarity across data sources remains underexplored, and current MCN benchmarks fail to reflect the variability, constraining the MCN models trained on them [15].

Taken together, effective multi-source fusion for MCN must be quality-aware. Therefore, we propose a quality-aware multi-source data fusion framework for MCN, focusing on social media-derived datasets as a representative and impactful use case. Fig. 1 illustrates the core idea. We evaluate six MCN datasets derived from social media and mapped to the same controlled vocabulary, SNOMED-CT. While multi-source fusion naturally enriches phrase coverage for shared medical concepts, qualitative inspection reveals substantial data quality issues. For example, the phrase *"Agitation feeling"* from TwiMed is nearly identical to its target concept *"Agitation"* and is semantically redundant with *"Feeling agitated"* from AskAPatient. Such intra- and inter-dataset redundancy reflects limited semantic variety and highlights the risk of naïve datasets. We aim to address the following research questions:

- **RQ1:** How to define and evaluate data quality metrics for MCN datasets?
- **RQ2:** How can large language models (LLMs) be leveraged to improve data quality across MCN datasets?
- **RQ3:** What is the impact of data quality on MCN model performance?
- **RQ4:** How to construct larger-scale and high-quality datasets for MCN using a quality-aware multi-source fusion framework?

To answer these research questions, (1) we first define and evaluate data quality in MCN by assessing six widely used datasets along five metrics: correctness, concept validity, coverage, semantic variety, and class imbalance. (2) We then adopt Gemini to augment data, improving semantic variety and mitigating long-tail class imbalance. Phrase counts increase by 172.1% after augmentation, and by 450.0% after fusion. (3) We adopted SapBERT, KNN-BioEL, and KRISSBERT models to quantify the impact of data quality on MCN performance through comparative experiments on original, augmented, and fused datasets. Across all three models, F1 scores increase under augmentation on every dataset, up to 38%. We also investigated the redundancy issue in AskAPatient and CADEC revealed by the decreased accuracy scores. (4) Finally, this framework enables us to construct larger-scale, high-quality MCN datasets with little label costs.

Introducing data quality and fusion into MCN has broad implications. For the MCN research community, it provides a systematic lens to diagnose foundational dataset issues and encourages more rigorous, data-centric development practices. For biomedical NLP, constructing larger, higher-quality corpora reduces annotation costs while improving model generalization and reliability. For healthcare applications, improved MCN performance supports more accurate interpretation of patient expressions, strengthens pharmacovigilance and safety surveillance, and ultimately contributes to more informed clinical and regulatory decision-making [3,16].

Our main contributions are summarized in Table 1 and as follows:

1. We propose the first quality-aware multi-source data fusion and enhancement framework for MCN. This framework contains data quality evaluation, investigation of data quality impacts on model performance, LLM augmentation, and data fusion for large-scale MCN dataset curation with less label cost.

Table 1
Statement of significance.

Aspect	Summary
Problem or Issue	Existing Medical Concept Normalization (MCN) research primarily relies on single-source datasets and often neglects systematic evaluation of data quality (DQ).
What is Already Known	Prior MCN datasets have reported notable DQ issues. Multi-source data fusion has been recognized as a way to alleviate data scarcity in health informatics, but quality-aware approaches remain limited.
What this Paper Adds	This paper introduces a novel quality-aware data fusion framework with five DQ metrics and LLM-based data augmentation strategies, demonstrated with six widely used MCN datasets. It presents the first quality-aware multi-source data fusion and enhancement framework for MCN.
Who Would Benefit	Biomedical NLP researchers and practitioners who seek to build high-quality health-related datasets more efficiently and at lower cost.

2. We design 5 DQ metrics for MCN datasets and analyze DQ issues in 6 widely used MCN datasets through experiments.
3. We develop LLM-based data augmentation strategies to enhance semantic variety and alleviate long-tail class imbalance in MCN corpora.
4. We construct a larger-scale, high-quality MCN dataset through quality-aware fusion. We demonstrate that the fused datasets improve model performance over single-source baselines.

The rest of the paper is organized as follows. Section 2 summarizes related work in MCN, data quality metrics, LLMs in data augmentation, and multi-source data fusion. Section 3 introduces our framework and workflow in detail, including datasets, data quality evaluation metrics, LLM prompt design, and fusion methods. Section 4 presents the DQ evaluation results, and Section 5 discusses how DQ issues, data augmentation and fusion affect model performance. Finally, Section 6 concludes the findings and puts forward future directions.

2. Related work

This section reviews related studies from four perspectives: (1) Medical Concept Normalization, covering definitions, datasets, and modeling approaches; (2) Data quality, emphasizing its impact on model performance; (3) Large Language Models (LLMs) in data augmentation, addressing how synthetic data can mitigate imbalance while raising quality concerns; and (4) Multi-source data fusion, exploring how integrating diverse datasets and knowledge bases enhances robustness through quality-aware fusion strategies.

2.1. Medical concept normalization

Medical concept normalization is also known as medical entity normalization/linking (MEN/MEL) [16,17], biomedical entity linking (BEL) [18], biomedical named entity normalization (BNEN) [19], biomedical terminology normalization (BTN) [20], and extreme multi-label text classification (XMTC) task [21].

Prior research on MCN has largely focused on improving model architectures, including domain-adapted transformer encoders, contrastive learning strategies, and multi-stage ranking pipelines [22,23]. Common solutions include convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer-based models, graph neural networks (GNNs), and large language models (LLMs). For example, SapBERT is a transformer-based model derived from BERT that specializes in biomedical concept representation through self-alignment of synonymous terms [24].

On the dataset side, several benchmark corpora have been developed [12,25,26], alongside shared tasks and surveys that examine task formulations, evaluation protocols, and modeling challenges [27–29]. There are three main categories for MCN datasets based on the source type: social media platforms, the medical literature, and clinical records. Social media becomes a valuable resource because of its patient engagement [30]. It allows patients to describe symptoms, share experiences, and ask questions about health-related issues [31,32]. The language style is often informal and colloquial, expressed in slang [7,33,34]. The platforms include patient forums such as AskAPatient website,[1] and general social media platforms such as Reddit[2] and Twitter (now 'X'[3]) [4].

The controlled vocabulary, or ontology, contains standardized concepts. Widely adopted ones include UMLS (Unified Medical Language System),[4] ICD (International Classification of Diseases),[5] SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms),[6] MedDRA (Medical Dictionary for Regulatory Activities),[7] and MeSH (Medical Subject Headings).[8] Despite their comprehensiveness in covering common and rare terms, the overlapping concepts between each vocabulary, concept synonyms, incorrect concept names with typos, and inactive concepts require further quality checks [5,6].

We identify two research gaps in this field. Firstly, the majority of studies evaluate their models on a single dataset at a time, rather than performing multi-source integration and validation. Magge et al. observed that the adverse drug effect (ADE) mention normalization achieves the best F1 score only when incorporating semi-supervised learning and annotations from other corpora [27]. Furthermore, data quality evaluation for these datasets is often neglected despite its importance [1,35]. As a result, fundamental data-centric issues in widely used MCN datasets remain insufficiently understood and underexplored.

2.2. Data quality

Data quality (DQ) evaluation metrics, as specific measurements, often include correctness, duplication, class imbalance, variety, and so on [4,36]. Correctness refers to the fact that records in a dataset are correctly labeled if they are labeled records. Incorrectly labeled data leads to label noise [1]. For example, every mapping between a Reddit phrase and a medical concept in a dataset should be correct. Duplication refers to the same instances appearing multiple times in one dataset or several datasets [4]. Duplication within datasets can significantly inflate model

[1] https://www.askapatient.com/
[2] https://www.reddit.com/
[3] https://x.com/
[4] https://www.nlm.nih.gov/research/umls/index.html
[5] https://icd.who.int/en
[6] https://www.nlm.nih.gov/healthit/snomedct/index.html
[7] https://ich.org/page/meddra
[8] https://www.nlm.nih.gov/mesh/meshhome.html

performance [1]. Models trained on duplicated data demonstrate substantially higher but over-claimed accuracy compared to those trained on de-duplicated data [7]. Data that is not exactly the same but highly similar indicates a semantic variety issue. Drosou et al. have defined a distance-based diversity metric calculated by pairwise similarity, and a novelty-based diversity metric to determine how different a new sample is relative to the existing samples to reduce redundancy [37]. In MCN, the phrases often exhibit low semantic variety, characterized by synonym replacement or changes in word order. Research has also found the class imbalance issue in MCN datasets [27,38]. The concepts often follow a long-tail distribution [39,40]. These tail labels or rare labels are the concepts with limited training samples, and they are harder to predict than the frequently occurring ones [21]. However, simply removing the rare cases cannot address this issue [41], thus requiring further research on mitigating these difficult concepts.

Besides designing DQ evaluation metrics, there is also a need to tailor these metrics to specific tasks and integrate them into a systematic framework [42]. Schwabe et al. designed a data quality framework for medical training data using a systematic review. It contains five clusters: consistency, informativeness, representativeness, timeliness, and measurement process, with a total of 15 dimensions [43]. However, this kind of framework mainly focuses on the evaluation and lacks validation on DQ improvement. As long as fusing multi-source data is potential and promising, simply increasing the quantity of data does not always improve performance [1], emphasizing the importance of data quality in data augmentation and fusion.

2.3. LLM in data augmentation

Data augmentation is the generation of high-quality artificial data by manipulating existing data [44]. It can help mitigate data imbalance by generating new samples for minority groups [45]. Large Language Models (LLMs) exhibit strong capabilities in generating phrases, context, and mappings in MCN tasks [1,20] and others in the medical domain [46]. The strength of using LLMs lies in their accuracy and cost-efficiency. Phan et al. found out that Gemini could generate high-quality medical responses used as training data [47]. Based on Jahan et al.'s findings, LLMs could improve biomedical tasks that lack large annotated data [48]. Fan et al. employed LLM agents to leverage external search engines and domain-specific knowledge bases to expand concise mentions into more detailed, semantically consistent, and informative descriptions [20]. They set the agents as a clinical doctor, an outpatient doctor, an internet doctor, and a terminology expert. The prompt each focuses on different biomedical perspectives, making reasoning more comprehensive.

Synthetic data needs data quality evaluation. Dobbins found out LLMs alone may hurt MCN performance, but when combined with other normalization systems, they can help achieve higher precision and recall by synonym and alternative phrasing generation [49]. For generating methods, Meng et al. explored how to combine few-shot training samples with the generated novel data [23]. Chen et al. conducted experiments on how the size of the generated samples would affect the model performance in MCN [1]. They tested zero-shot and few-shot prompting to generate augmentation sizes ranging from 5 to 100. Increasing the size of the generated samples in the combined training set generally improves model performance on the combined testing set. While generating data can be beneficial, the above research also points out that synthetic data needs data quality evaluation, as overly large augmentation sets may introduce noise [1,23,45]. This again emphasizes the importance of data quality and a data-centric workflow.

2.4. Multi-source data fusion

Multi-source data is data collected from different sources. Fusing data from multiple sources refers to integrating them into one merged dataset [13]. It can mitigate data scarcity in health informatics and also utilize complementary information from multiple datasets. As an example in the medical domain, Yang et al. fused 10 mental-health-related datasets from social media platforms into a unified, larger-scale corpus [15].

Source and content similarity facilitate cross-dataset generalization, pointing to the potential of data fusion approaches in MCN tasks. While single-source corpora alone can be limited, fusing datasets enables models to capture more diverse yet compatible patterns, thus improving performance [12–14]. Dai et al. found that the model trained on CADEC [50] also performs well on PsyTAR [51] because both are sampled from online patient forums, despite their different focuses [52]. Peng et al. also found that multi-source medical datasets present higher comprehensiveness over single-source ones [53].

Fusing heterogeneous sources of knowledge, such as context, medical ontologies, and task-specific supervision, can improve concept alignment [33,54]. Integrating textual representations from discharge summaries and other external medical knowledge bases, such as medical dictionaries, can mitigate the scarce annotated data [55]. Xu et al. found that integrating controlled vocabulary is more beneficial than literature sources, with incorporating all corpus yielding the highest performance [56]. Similarly, integrating synonyms generated by both SNOMED-CT and UMLS can improve the performance while training on CADEC [50] and PsyTAR [34,51].

Although fusing training data from multiple sources can make the model more generalizable, the absolute performance may be lower when comparing it to individual source-trained models due to the overfitting issue of a single-source dataset or the consistency issue across datasets [57]. This further requires data quality evaluation across multi-source datasets and a fair comparison of the model performance. Therefore, multi-source fusion also needs to be quality-aware. Zhang et al. put forward four metrics to evaluate multi-source DQ: accuracy, completeness, consistency, and instantaneity [58]. Sun et al. evaluated the trustworthiness of source data from the perspective of data provenance in healthcare [59]. However, to the best of our knowledge, there still lacks a DQ-driven multi-source fusion framework for MCN tasks.

3. Methodology

3.1. Research design

Fig. 2 shows the research design. As mentioned, previous MCN tasks not only rely on a single-source dataset but also overlook the data quality issues. This research design includes five data quality evaluation metrics and corresponding improvement methods. The metrics were proven critical for MCN task [1,4] and in related data quality research [21,27,38–40]. Specifically, to ensure semantic variety and address long-tail problems before fusion, we employ an LLM to conduct data augmentation. By fusing the augmented datasets, we achieve better performance, thus offering a quality-aware solution to MCN tasks. Section 3.3 gives a detailed explanation of the metrics.

3.2. Datasets

AskAPatient, CADEC, COMETA, PsyTAR, TwADR-S, and TwiMed were selected based on shared characteristics that make them both comparable and complementary for multi-source MCN research. The six datasets were chosen because they are all widely used MCN benchmarks derived from user-generated social media text and mapped to SNOMED-CT. Moreover, exhibit known data quality issues such as low semantic variety [1], and long-tail concept imbalance [27,38], making

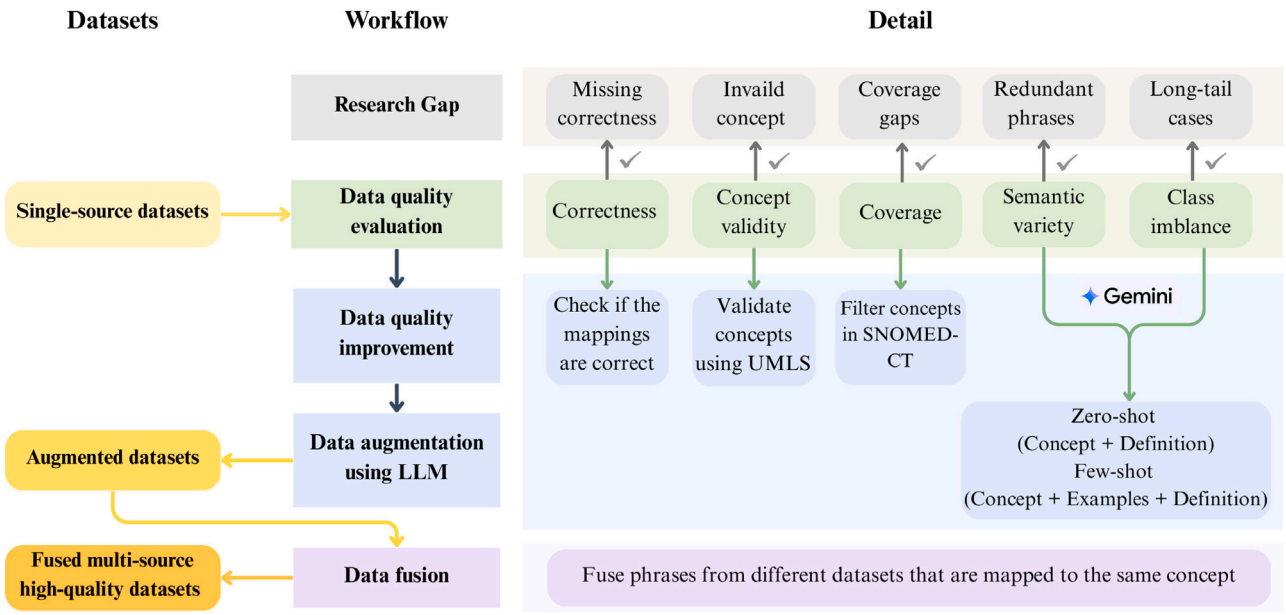


Fig. 2. Research design for quality-aware multi-source data fusion and enhancement. The data quality evaluation metrics in the second row indicate how this approach addresses the research gap in this field. Fused datasets are enriched across other datasets with higher quality compared to the original ones.

Table 2
Counts of phrases and concepts of original datasets.

Dataset	Count of phrases	Count of concepts
AskAPatient	4,496	1,036
CADEC	17,842	1,052
COMETA	20,015	3,901
PsyTAR	6,159	690
TwADR-S	201	58
TwiMed	791	185
Total	**49,504**	**6,922**

them suitable for evaluating quality-aware augmentation. Each dataset is introduced below, and summarized in Table 2.

AskAPatient [25] is derived from blog posts on the AskaPatient website, which contains patient-reported experiences with prescription medications. The drug ratings and reviews mainly discuss health issues, symptoms, drugs, adverse drug reactions (ADR), and others. As the original 10-fold training and testing datasets present significant duplication data quality issues, which cause overfitting [1,7], here we adopt the de-duplicated version.

CADEC [50] is a corpus of user discussions from AskaPatient posts annotated for adverse drug event mentions, drug names, and others.

COMETA [9], Corpus of Online Medical Entities, is a dataset for medical entity linking drawn from Reddit health-related subreddits, mapping to SNOMED-CT concepts. It covers diverse biomedical concept types such as symptoms, anatomy, diseases, and procedures.

PsyTAR [51], 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 AskaPatient, each with demographic information, treatment duration, and detailed patient assessments.

TwADR-S [25] is a dataset of short tweets on Twitter mapped to SNOMED-CT concepts. The "S" stands for "short" tweets.

TwiMed [60], short for Twitter–Medline, is a biomedical text corpus designed to support social media-based health and pharmacovigilance research, particularly focusing on user-generated health-related tweets.

While the six datasets together cover a diverse range of biomedical and social media sources for MCN, they also present unique features. AskAPatient and CADEC are derived from patient forums, containing informal user-generated posts that describe personal medication experiences and adverse effects. COMETA, TwADR-S, and TwiMed data are from Twitter. COMETA is the largest of the three, containing tens of thousands of automatically mapped health-related mentions. In contrast, TwADR-S is a small, manually annotated dataset focused specifically on ADR mentions, offering high annotation quality but limited data volume. Domain-wise, PsyTAR focuses specifically on psychiatric medication reviews, offering manually annotated data that captures mental health-related expressions.

3.3. Data quality evaluation

Fig. 3 summarizes and compares the data quality evaluation metrics adopted in this research. Each of them is part of the essential DQ evaluation and has a different focus, as illustrated in Fig. 2. These metrics together form a comprehensive DQ framework that covers the relationships between phrases, concepts, and the mapping of each phrase-concept pair.

3.3.1. Correctness

Correctness evaluates whether each phrase is mapped to a concept without error [1]. Our previous research [1] has demonstrated LLMs' effectiveness in evaluating MCN mapping's correctness. We generated 5 phrases for each concept in AskAPatient and TwADR-L, yielding 5180 and 11,000 evaluated mappings, respectively. Two graduate students with health informatics backgrounds independently judged whether each mapping was correct. The agreement scores were 0.9342 and 0.9526 for AskAPatient and TwADR-L under zero-shot prompting, and 0.9465 and 0.9375 under few-shot. These results support the reliability of structured correctness evaluation for MCN mappings. In this study, we also manually checked Gemini's evaluation results.

Correctness serves as the first step of the evaluation process, as the results determine whether the dataset is suitable for the intended task. A very low correctness rate suggests the dataset may not be reliable for further analysis. In this study, if the correctness rate exceeds a predefined threshold, we consider the dataset to be of sufficient quality for subsequent experiments.

$$Correctness = \frac{\text{Number of correct records}}{\text{Total number of records in the dataset}} \quad (1)$$

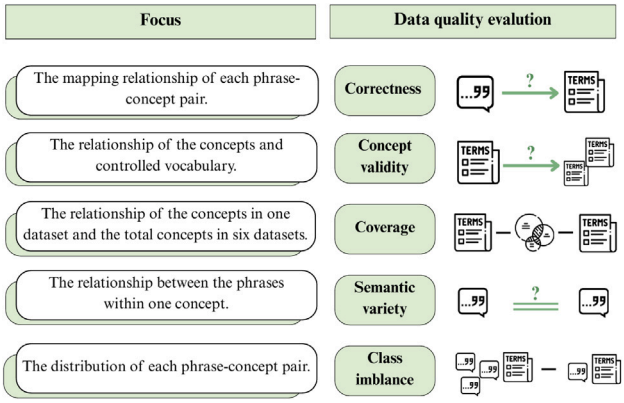


Fig. 3. Data quality (DQ) metrics and their focuses. The "TERMS" icon represents a medical concept, and the quote icon represents the phrases mapped to this concept. These two combined refer to the MCN datasets that contain many mappings.

Defined in Eq. (1), this metric can be evaluated either by an LLM judge or by human annotators. We employ Gemini-2.5-flash as the evaluation model. It needs to determine if the phrase is a reasonable description, synonym, or common expression of the concept, and only reject mappings that are clearly unrelated. When rejecting, it must suggest the most precise alternative SNOMED-CT term.

3.3.2. Concept validity

The concepts in datasets could be updated in controlled vocabulary over time, becoming invalid. Checking concept validity helps diagnose potential DQ issues, such as annotation and labeling errors. A concept c in the dataset is valid if it is in SNOMED-CT, the controlled vocabulary for all six datasets. Concept Validity evaluates how many valid concepts are in the dataset, as shown in Eq. (2).

$$CV = \frac{\text{Number of valid } c \text{ in the dataset}}{\text{Total number of } c \text{ in the dataset}} \tag{2}$$

For invalid concepts, we use the UMLS API to retrieve the top five closest matches. We adopt the UMLS API because our target vocabulary, SNOMED-CT, is fully integrated within the UMLS, allowing reliable validation. Its cross-terminology mappings and updated concept information also enable automated recovery from annotation errors or outdated identifiers at scale. As shown in Fig. 4, the UMLS API returns semantically similar candidates for the invalid concepts.

We determine the best replacement for invalid concepts using semantic similarity. The candidate with the highest similarity with the searched concept is selected. If the UMLS API returns no candidates, the invalid concept is removed. For example, the original term 'on examination - tremor outstretched hands' yields no exact matching results in UMLS, thus becoming invalid. To replace it, the candidate 'Bilateral outstretched hands tremor' is chosen because of its high similarity with the original term.

3.3.3. Coverage

Coverage evaluates what percentage of concepts from each dataset represent the total mapped concepts, as shown in Eq. (3). A higher coverage means the dataset contains more valid concepts than others.

$$Coverage = \frac{\text{Number of } c \text{ in one dataset}}{\text{Total number of } c \text{ across all datasets}} \tag{3}$$

3.3.4. Semantic Variety

Semantic Variety evaluates the extent to which one concept appears through different linguistic expressions, beyond simple reorderings or synonym replacement. Previous study has shown that using the BERT

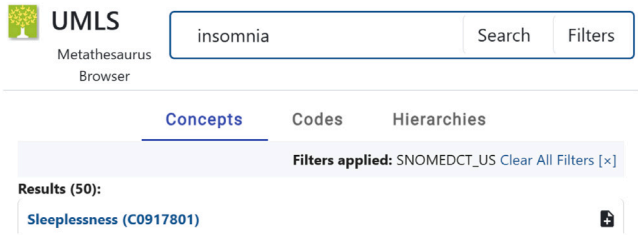


Fig. 4. Retrieving result of 'insomnia' in SNOMED-CT.

(bidirectional encoder representations from transformers) model [61] to calculate the cosine similarities between the embeddings of informal phrases for medical concepts can reflect the semantic variety of the datasets [1]. High mean cosine similarity means low semantic variety, implying the phrases are nearly redundant. Conversely, a lower mean similarity reflects greater diversity in expressing the medical concept.

Compute it by embedding all phrases p for the concept c and summarizing pairwise cosine similarities using Eq. (4), where n_p means the number of phrases for a concept c.

$$SV = 1 - \frac{2}{n_p(n_p-1)} \sum_{i<j} \cos(\mathbf{e}_i, \mathbf{e}_j) \tag{4}$$

3.3.5. Class imbalance

Class imbalance measures how the phrase-concept pairs are distributed. The outcome of this metric is usually a distribution figure. One dataset may have many mentions of common or general terms, and few mentions of rare ones. This causes an imbalanced distribution, where a small number of concepts with a large number of phrases dominate the data [1,62]. High class imbalance makes the model biased towards the common classes and perform poorly on the rare classes [63].

3.4. Data augmentation using LLM

To improve semantic variety and address class imbalance issues, we performed data augmentation using the Gemini-2.5-flash model [64]. The process involved normalization, definition enrichment, prompt design, and human review.

To filter the rare concepts that require augmentation, we set a threshold of 20, meaning concepts with more than 20 phrases are considered sufficiently diverse and excluded from augmentation. It refers to previous research [1], to balance model performance improvement and augmentation cost. As shown in Algorithm 1, zero-shot refers to prompting the LLM without examples, whereas few-shot uses examples for guidance. Zero-shot candidates are phrase–concept pairs whose normalized forms were identical, indicating no meaningful example phrases for that concept [16]. Few-shot candidates are concepts with fewer than 20 distinct phrases.

For the zero-shot and few-shot sets, each concept was then enriched with a definition to provide context. Definitions were retrieved from the UMLS API when available; otherwise, Gemini-2.5-flash generated correct substitutes with sources, as in Algorithm 2. Prompt design is as follows, and the augmented data were manually reviewed to ensure quality. Algorithm 3 provides the process of data augmentation.

3.5. Data fusion

Six individual datasets were fused based on shared medical concepts identified by their concept names and IDs. Since many concepts appear across multiple datasets, this fusion process aggregates related phrases and attributes under the same concept identifier. For instance, if the concept 'Agitation' and its ID in AskAPatient also appear in other datasets, all associated phrases from those datasets will be fused into

Algorithm 1 Split Subsets to Zero Shot and Few Shot

Require: Table T with columns `concept`, `phrase`; FEW_SHOT_Threshold = 19
Ensure: `To_Zero_Shot`, `To_Few_Shot`, `Remainder`

function NORMALIZE(s)
  if s = null **then**
    return ""
  end if
  $s \leftarrow$ lower(s)
  remove substrings matching `( . . . )` from s
  remove punctuation; collapse spaces; trim
  return s
end function
for all $r \in T$ **do**
  $r.n\_concept \leftarrow$ NORMALIZE($r.concept$)
  $r.n\_phrase \leftarrow$ NORMALIZE($r.phrase$)
end for
for all $c \in T$.`concept` **do**
  $P(c) \leftarrow$ distinct n_phrase where $concept = c$
  $exact(c) \leftarrow (\exists r : r.concept = c \wedge r.n\_phrase = r.n\_concept)$
  if $exact(c)$ **and** $|P(c)| = 1$ **then**
    mark c as ZERO
  end if
end for
`To_Zero_Shot` $\leftarrow \{r \in T \mid r.concept$ is ZERO$\}$
$R \leftarrow T \setminus$ `To_Zero_Shot`
$R\_d \leftarrow$ unique rows of R keyed by $(concept, n\_phrase)$
for all $c \in R\_d$.`concept` **do**
  $P'(c) \leftarrow$ distinct n_phrase in R_d where $concept = c$
  if $|P'(c)| \leq$ FEW_SHOT_Threshold **then**
    add rows with $concept = c$ to `To_Few_Shot`
  end if
end for
`Remainder` $\leftarrow R\_d \setminus$ `To_Few_Shot`
return (`To_Zero_Shot`, `To_Few_Shot`, `Remainder`)

the new fused AskAPatient dataset, as illustrated in Fig. 1. This ensures that each concept in the fused dataset contains a comprehensive and enriched set of linguistic expressions representing the same concept across all sources.

3.6. Experiment

To systematically evaluate the impact of data enrichment strategies on model performance, three experimental settings were designed on original, augmented, and fused datasets, respectively. The Original experiment establishes a baseline by training and testing SapBERT [24]. We also added KNN-BioEL [65] and KRISSBERT [66] models for cross-model comparison. KNN-BioEL is a retrieval-enhanced model that combines encoder-based similarity with evidence from the k nearest training instances, allowing the model to leverage similar examples during inference. KRISSBERT is a knowledge-rich self-supervised entity linking model initialized from PubMedBERT and trained with UMLS-based supervision automatically generated from unlabeled biomedical text. The Augmentation experiment extends the training and testing set with a combination of original and augmented data. The purpose is to examine whether improving semantic variety can improve model performance. Similarly, the fusion experiment incorporates a combination of original, augmented, and fused data in both training and testing phases. The rationale is that training on augmented or fused data but testing solely on the original dataset can penalize models unfairly, as augmented data and fused data may shift the feature distribution away from the original data [1].

Algorithm 2 Add Definitions via UMLS API → LLM

Require: Splits (`To_Zero_Shot`, `To_Few_Shot`, `Remainder`); UMLS key; LLM
Ensure: `Definitions` as tuples (c, def, src)

$C \leftarrow$ unique concepts in `To_Zero_Shot` ∪ `To_Few_Shot` ∪ `Remainder`
load definition cache; $ticket \leftarrow$ UMLSAUTH(key)
for all $c \in C$ **do** ▷ Resolve definition for each concept
  if c in cache **then**
    $(def, src) \leftarrow$ cache[c]
  else
    $cui \leftarrow$ UMLSSEARCHCONCEPT($ticket, c$)
    if $cui \neq \emptyset$ **then**
      $D \leftarrow$ UMLSGETDEFINITIONS($ticket, cui$)
      if $D \neq \emptyset$ **then**
        $(def, src) \leftarrow$ CHOOSEBEST(D) ▷ prefer SNOMEDCT_US, then MeSH, NCI, ...
      else
        $(def, src) \leftarrow$ LLMDEFINE(c) ▷ clear
      end if
    else
      $(def, src) \leftarrow$ LLMDEFINE(c)
    end if
    cache[c] $\leftarrow (def, src)$
  end if
  add (c, def, src) to `Definitions`
end for
return `Definitions`

Few-shot prompt template for data augmentation

System Prompt
You are an expert in biomedical concepts and Medical Concept Normalization tasks. Follow the schema and guidelines strictly to ensure the output is valid JSON only.

Output Format:	Return ONLY valid JSON matching the provided schema.
Schema:	`{json.dumps(schema_hint, indent=2)}`
Language:	US English
Task:	Generate informal phrases that patients could use to describe this concept in everyday language on social media.
Quantity:	Generate exactly `{num_new}` distinct phrases per concept.
Restrictions:	No emojis, hashtags, @mentions, or URLs. Do not copy example phrases verbatim.
Diversity:	Try to cover a variety of ways people might express this condition.
Return Format:	JSON object of the form: `{"results": {"<concept>": {"phrases": ["...", "..."]}, ...}}`

User Prompt
Concepts, definitions, and few-shot examples in JSON

Across all experiments, a 10-fold cross-validation scheme was employed to ensure statistical reliability. For each fold, the training set was explicitly de-duplicated against the corresponding testing fold

Algorithm 3 Augment Using LLM and Combine

Require: Splits from Alg. 1; `Definitions` from Alg. 2; K phrases per concept
Ensure: `Aug_Zero`, `Aug_Few`, `Final_Improved`

function BuildPrompt(c, def, E, K)
  return instruction with: concept c, definition def, target K, example phrases E (if any), and:
  "Generate K informal phrases patients might use on social media to describe this concept; cover varied ways people might express this condition. Output JSON `{ "phrases": [ . . . ] }` only."
end function
$S \leftarrow$ set of $(concept, n\_phrase)$ from `To_Zero_Shot` ∪ `To_Few_Shot` ∪ `Remainder`
`Aug_Zero` ← ∅, `Aug_Few` ← ∅
for all $c \in$ `To_Zero_Shot.concept` **do** ▷ no example phrases
  $def \leftarrow$ definition of c from `Definitions`
  $E \leftarrow \emptyset$
  $prompt \leftarrow$ BuildPrompt(c, def, E, K)
  $J \leftarrow$ LLMJSON($prompt$) ▷ strict JSON
  $P \leftarrow J$.`phrases`
  for all $p \in P$ **do**
    $np \leftarrow$ Normalize(p)
    if $np \neq$ "" $\wedge\ np \neq$ Normalize(c) $\wedge\ (c, np) \notin S$ **then**
      add $(c, p,$ "LLM") to `Aug_Zero`; $S \leftarrow S \cup \{(c, np)\}$
    end if
  end for
end for
for all $c \in$ `To_Few_Shot.concept` **do**
  $def \leftarrow$ definition of c from `Definitions`
  $E \leftarrow$ distinct original phrases for c from `To_Few_Shot`
  $prompt \leftarrow$ BuildPrompt(c, def, E, K)
  $J \leftarrow$ LLMJSON($prompt$); $P \leftarrow J$.`phrases`
  for all $p \in P$ **do**
    $np \leftarrow$ Normalize(p)
    if $np \neq$ "" $\wedge\ np \neq$ Normalize(c) $\wedge\ (c, np) \notin S$ **then**
      add $(c, p,$ "LLM") to `Aug_Few`; $S \leftarrow S \cup \{(c, np)\}$
    end if
  end for
end for
`Final_Improved` ← `Remainder` ∪ `Aug_Zero` ∪ `Aug_Few`
return (`Aug_Zero`, `Aug_Few`, `Final_Improved`)

to prevent data leakage, ensuring that no instance in the training data overlapped with the current test fold. To assess the statistical significance of the observed gains, we apply a paired t-test to the per-dataset Macro-F1 scores across the six datasets, treating $p < 0.05$ as the threshold for significance.

4. Results

4.1. Data quality evaluation

Table 3 summarizes data quality results on the first three metrics. All the datasets present above 80% **correctness** rate, and we consider them to be correct enough for the following tasks.

Concept validation evaluates whether each concept corresponds to a valid SNOMED-CT entry using the UMLS API. As shown in Table 3, PsyTAR has a relatively low concept validity rate compared to the other datasets. To further investigate this, Table 4 lists the top ten validated concept corrections in PsyTAR. Three major types of validation adjustments are observed.

1. Resolving 'no code' mappings by assigning appropriate valid SNOMED-CT concepts.
2. Replacing non-SNOMED terms with their valid ones. For example, 'insomnia' is not a recognized SNOMED-CT term and should be replaced by 'sleeplessness,' as shown in Fig. 4.
3. Adding semantic tags such as '(finding)' to specify the SNOMED-CT concept type.

These types of validation adjustments apply to all six datasets, though they appear more frequently in PsyTAR due to its higher proportion of informal or non-standard terms. Overall, the validation process not only ensures that each concept belongs to the same controlled vocabulary but also harmonizes terminology across datasets, enabling concept fusion in subsequent analyses.

For **coverage**, COMETA contains the most SNOMED-CT concepts among all, with TwADR-S contributing the least. This corresponds to the datasets' total concept counts. This step also filters out the concepts that are specifically in SNOMED-CT, since the validation step may return concepts in other vocabularies.

Fig. 5 shows that the AskAPatient and CADEC datasets show high embedding similarity between phrases for each concept, indicating that expressions are semantically very close. In AskAPatient, this result is particularly notable because the version used here has already had duplicates removed, yet the similarity remains high. This highlights a serious issue if the original 10-fold dataset were used, as data redundancy in the training and testing sets will cause model biases [7]. COMETA and PsyTAR have much lower similarities. However, high variety can also introduce noise, potentially complicating alignment between phrases and standardized concepts. With fewer data, TwADR-S and TwiMed exhibit moderate variety, suggesting limited but present variation in phrasing.

Fig. 6 shows that every dataset has a **class imbalance** issue. The long-tail distribution indicates that only a few concepts are presented in many phrases, whereas most concepts are underrepresented. Hundreds or thousands of concepts appear only once or twice. This makes it hard for the model to learn the rare cases. AskAPatient and CADEC display the steepest declines and longest tails. COMETA also shows substantial imbalance, though its larger scale distributes phrases slightly more evenly across low-frequency concepts. In contrast, PsyTAR presents a relatively milder long-tail shape, while TwADR-S and TwiMed suffer from sparsity due to their small overall sizes.

Overall, data quality evaluation reveals issues with these datasets that have been almost neglected before. For correctness, although LLM judgment could only serve as a reference, it points to potential mapping quality issues. Regardless of whether these mappings are human-annotated or automatically generated, there are incorrect or ambiguous mappings that need further investigation. This could be caused by invalid concepts that are outdated or beyond the scope of the targeted vocabulary. In addition, redundant phrases lower the semantic variety, causing model overfitting. Similarly, if the phrase merely mirrors the concept, then it can be considered non-informative. To enrich the semantic expression and learn from multiple sources, we employ data augmentation and fusion.

4.2. Data augmentation and fusion for quality improvement

Data augmentation primarily addresses intra-dataset class imbalance issues by generating phrases for infrequent cases. It also focuses on semantic variety by requiring different styles of daily expression. As shown in Table 5, augmentation increases the number of phrases per concept, especially for small datasets such as TwADR-S.

The impact of data fusion largely depends on the degree of concept overlap across datasets. As illustrated in Fig. 7, AskAPatient and CADEC exhibit nearly complete concept overlap, reflecting their shared focus on content. Consequently, the fusion of these two datasets yields a harmonization of their phrase sets, resulting in comparable final dataset sizes. TwiMed also increases substantially because of sharing over 70% of concepts with COMETA. Another factor is the original concept count.

Table 3
Data quality evaluation results on correctness, concept validity, and coverage for original datasets.

Dataset	Total concepts (Count)	Correctness (%)	Valid concepts (Count)	Concept validity (%)	SNOMED-CT concepts (Count)	Coverage (%)
AskAPatient	1,036	85.4	950	91.7	864	15.8
CADEC	1,052	85.3	943	89.6	857	15.7
COMETA	3,901	83.2	3,693	94.7	3,197	58.6
PsyTAR	690	81.0	337	48.8	315	5.8
TwADR-S	58	89.1	56	96.6	53	1.0
TwiMed	185	81.8	184	99.5	167	3.1
Total/**Average**	6,922	**84.3**	6,163	**82.9**	5,435	**16.7**

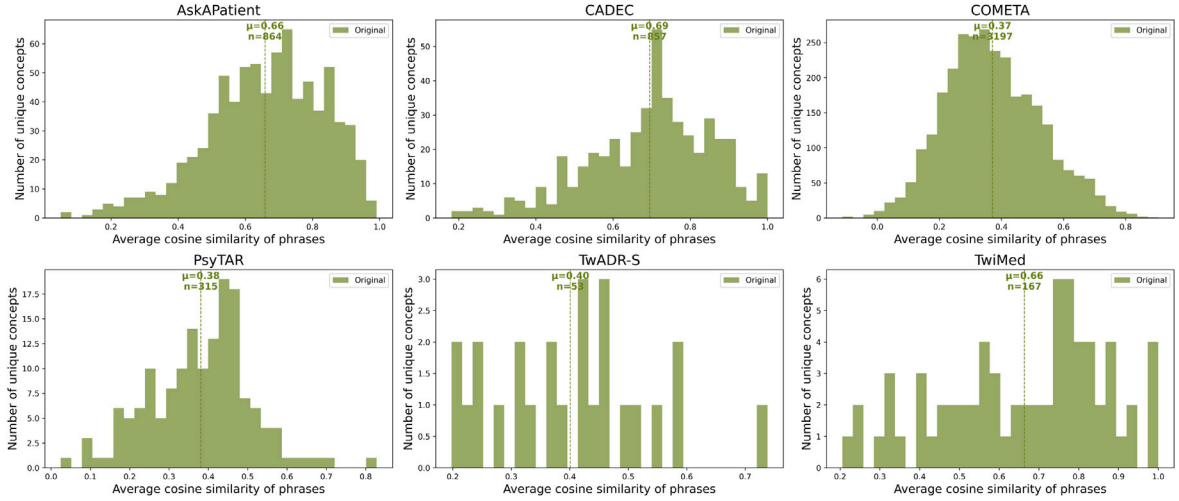


Fig. 5. The embedding similarity of original datasets.

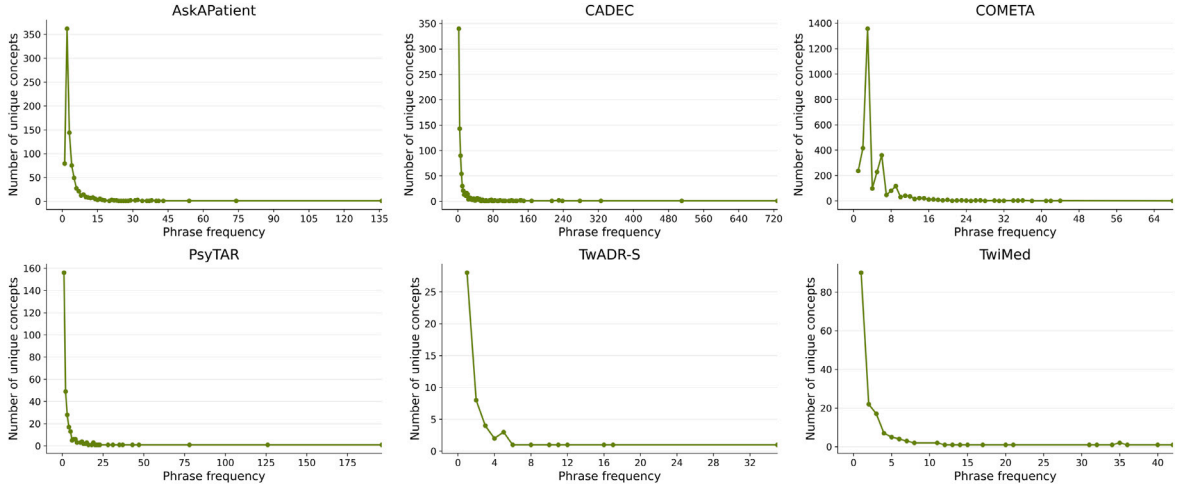


Fig. 6. Data quality evaluation on class imbalance for original datasets.

COMETA is the largest dataset, thus it also increases significantly after fusion.

Fig. 8 reveals the effects of data augmentation and fusion on **semantic variety**. For highly similar datasets such as AskAPatient, CADEC, and TwiMed, the similarity distributions shift leftward, indicating that augmented phrases introduce greater semantic variety and reduce redundancy. COMETA and PsyTAR exhibit moderate rightward shifts, suggesting that augmentation mitigates excessive variability for the datasets that have very low phrase similarity.

For **class imbalance**, data augmentation expands low-frequency coverage, while fusion integrates diverse expressions across datasets. Across all datasets, data augmentation reduces the number of rare cases where concepts are represented by only a few phrases. As shown in Fig. 9, the red distributions shift rightward, starting from frequencies of about 20 and higher. The fused blue lines extend across a wide frequency range but display fewer extreme spikes than either the original green or augmented red lines. Fusion not only effectively enlarges the scale but also smooths the distribution. AskAPatient and CADEC show similar distribution patterns because these datasets share many concepts and contain similar types of content. The other four datasets show more similarity between the red and blue lines, indicating that fusion closely follows the augmentation pattern while maintaining improved class balance.

After identifying the data quality issues, data augmentation and fusion together improve semantic variety and class imbalance in our quality-aware multi-source fusion framework.

Table 4
Top 10 validated concepts in the PsyTAR dataset.

Original concept	Examples of validated concept	Count
no code	blurred vision, feeling tired, sleeplessness, etc.	194
insomnia	sleeplessness	67
loss of appetite	loss of appetite (finding)	52
brain shivers	shivering	50
excessive weight gain	excessive body weight (finding)	43
tired	feeling tired	39
feeling suicidal	feeling suicidal (finding)	33
excessive sweating	increased sweating	30
reduced libido	decreased libido	29
indifference	apathy	26

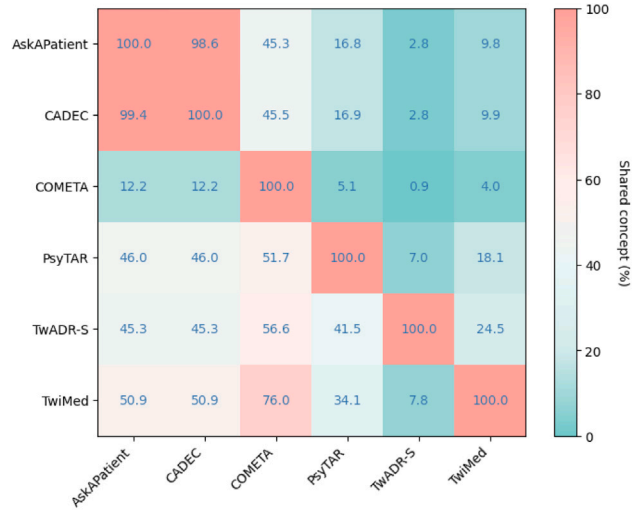


Fig. 7. Each cell (i, j) shows the percentage of concepts in dataset i (row) that also appear in dataset j (column). For example, the first row presents the percentage of concepts in AskAPatient shares with the other datasets. 98.6% of AskAPatient concepts also appear in the CADEC dataset. In the second row, 99.4% of CADEC's concepts are found in AskAPatient.

4.3. Model performance

Table 6 summarizes model performance across the original, augmented, and fused versions of the six MCN datasets. Under SapBERT phrase-level metrics, each phrase is treated independently. Because frequent concepts contribute more mentions, they dominate the metric. As a result, a model may appear to perform well overall even if it consistently fails on rare concepts. For this reason, we introduce concept-level macro metrics, which compute performance for each concept individually and then average across all concepts with equal weight. For concept set $C = \{c_1, \ldots, c_K\}$, macro F1 is defined as in Eq. (5):

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} \frac{2 \cdot P_k \cdot R_k}{P_k + R_k}, \tag{5}$$

where P_k and R_k denote precision and recall for concept c_k. Macro metrics are therefore highly sensitive to rare-concept performance and provide a direct measure of whether augmentation improves underrepresented concepts rather than merely refining predictions for common ones.

Table 9 reports the number of concepts per phrase-frequency group, and the mean accuracy improvement within each group. The results show greater improvement in low-phrase groups (1–3 phrases), confirming that data augmentation reduces sparsity and strengthens normalization performance for long-tail concepts.

Overall, the augmented and fused datasets produce substantial improvements for COMETA, PsyTAR, TwADR-S, and TwiMed. For these four datasets, augmentation increases accuracy, recall, and F1 scores, with gains ranging from moderate to large across different metrics. Fusion further enhances performance for most metrics and provides the highest overall results, especially for recall and F1, indicating more comprehensive coverage of phrase–concept mappings.

Data augmentation and fusion lead to different effects across datasets. AskAPatient and CADEC experience performance decreases after augmentation, compared to their relatively high accuracy on original datasets. In these two datasets, fusion performs slightly better than augmentation but remains below the original accuracy, indicating that neither augmented phrases nor cross-dataset phrases improve performance. Recall, precision, and F1 score show similar patterns.

In contrast, COMETA, PsyTAR, TwADR-S, and TwiMed have clear performance gains with both augmentation and fusion. COMETA shows the largest gains for Top-1 and Top-5 accuracy by 26.34% and 28.65% after augmentation, respectively. PsyTAR and TwADR-S follow a similar pattern. PsyTAR improves significantly on Top-5 accuracy. TwADR-S improves the most across all the metrics, among the 6 datasets. TwiMed shows moderate improvement under augmentation and a further increase under fusion.

Across all datasets, the fused versions consistently show stronger recall and F1 scores compared with single-source baselines. The improvements are most pronounced for PsyTAR, TwADR-S, and COMETA, where fusion yields up to double-digit gains in recall and F1. Even in datasets where augmentation reduces performance, such as AskAPatient and CADEC, the fused versions still outperform the augmented models, suggesting that integrating samples from multiple datasets generally leads to more stable results.

These patterns are consistent across models. Table 7 reports the same original, augmented, and fused experiments for kNN-BioEL and KRISSBERT alongside SapBERT. Across all three models, concept-level Macro-F1 increases under augmentation on every dataset. For example, on COMETA, Macro-F1 increases from 32.83 to 67.60 for SapBERT, from 34.01 to 70.02 for kNN-BioEL, and from 31.06 to 64.27 for KRISSBERT. The dataset-dependent phrase-level behavior is also largely preserved across models: COMETA, PsyTAR, TwADR-S, and TwiMed show clear gains after augmentation and fusion, whereas AskAPatient and CADEC show weaker or negative phrase-level gains, consistent with the redundancy-related effects observed in the SapBERT results. The consistency of these trends across different models demonstrates that the improvements from quality-aware data enhancement are stable. A paired t-test across the six datasets confirms that the gain in concept-level Macro-F1 under augmentation is statistically significant for all three models (Table 8, $p < 0.05$), providing quantitative evidence that quality-aware augmentation yields systematic improvements on rare concepts.

5. Discussion

In contrast to AskAPatient and CADEC, where augmentation reduced performance, the datasets COMETA, PsyTAR, TwADR-S, and TwiMed all show substantial performance gains after augmentation and fusion. This section aims to discuss the reasons for the differences through ablation experiments on data quality.

5.1. Overview of main findings

Across the six MCN datasets, our experiments reveal three major patterns. First, data augmentation leads to substantial performance improvements in four datasets—COMETA, PsyTAR, TwADR-S, and TwiMed, across Top-1 accuracy, Top-5 accuracy, recall, precision, and

Table 5
Phrase statistics before and after data augmentation and fusion. The percentages indicate the relative increase compared to the original phrase count in the second column.

Dataset	Count of phrases	After augmentation (count, % increase)	After fusion (count, % increase)
AskAPatient	4496	20,680 (+359.9%)	60,763 (+1,251.6%)
CADEC	17,842	23,580 (+32.1%)	60,575 (+239.5%)
COMETA	20,015	77,868 (+289.1%)	107,990 (+439.7%)
PsyTAR	6,159	7,645 (+24.1%)	23,062 (+274.4%)
TwADR-S	201	1,234 (+513.9%)	4,707 (+224.2%)
TwiMed	791	3,698 (+367.5%)	15,159 (+1,816.7%)
Total	**49,504**	**134,705 (+172.1%)**	**272,256 (+450.0%)**

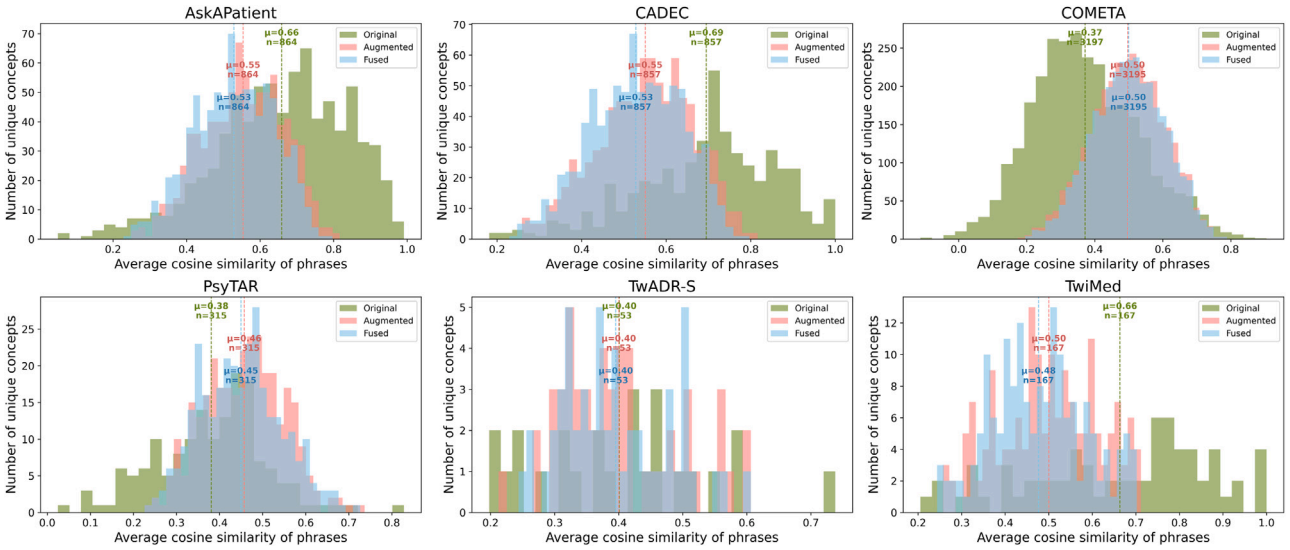


Fig. 8. The embedding similarity before and after data augmentation and fusion.

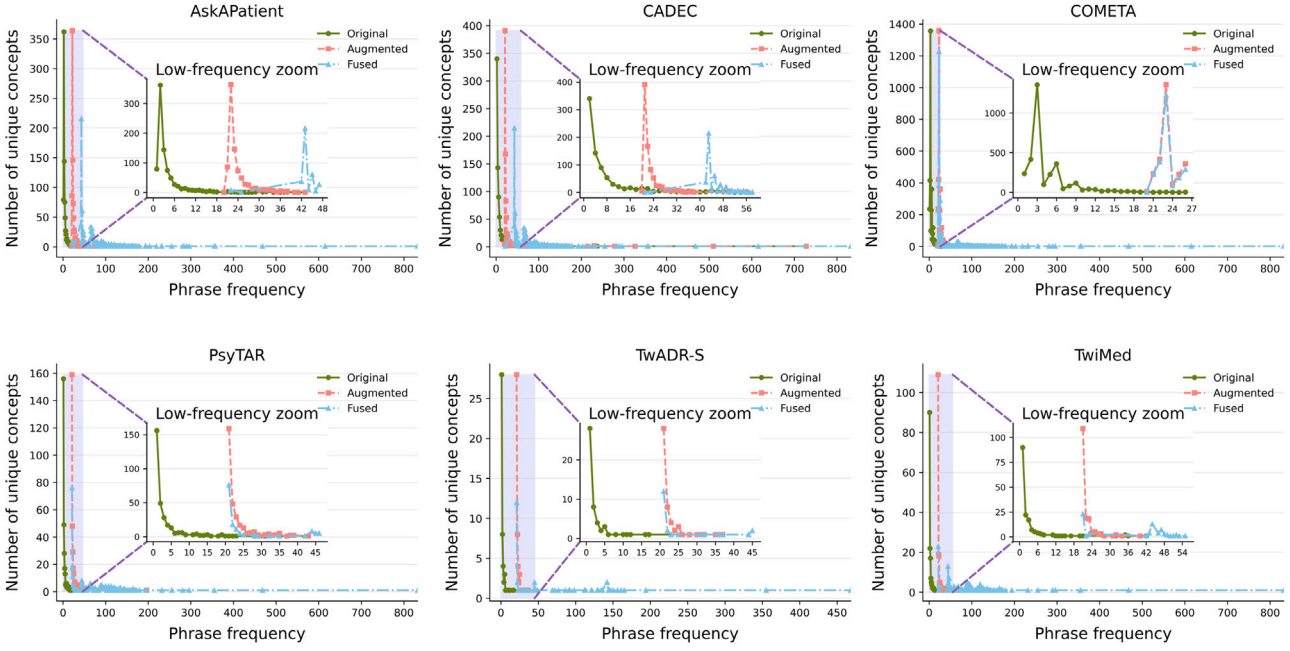


Fig. 9. Data quality evaluation on class imbalance before and after data augmentation and fusion. The shaded rectangle in each main panel highlights the low-frequency region containing 80% of the total concepts. The centered zoom-in plot provides finer inspection of differences.

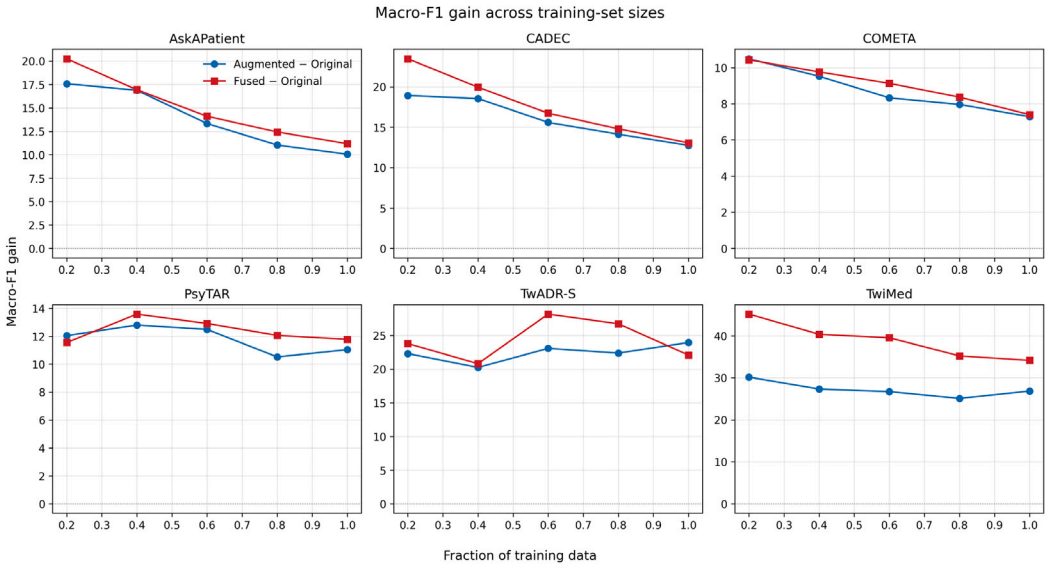
F1 scores. Second, augmentation decreases performance on AskAPatient and CADEC, though multi-source fusion partially recovers these degradations. Third, despite the mixed outcomes of augmentation, fusion consistently produces more stable gains across most datasets, particularly in recall and F1. These observations highlight that augmentation is highly sensitive to dataset-specific characteristics, emphasizing

Table 6
SapBERT phrase-level performance and concept-level macro performance across original (Org), augmented (Aug), and fused datasets. Acc@1 denotes Top-1 accuracy and Acc@5 is Top-5 accuracy. Precision@1, Recall@1, and F1@1 indicate evaluation using only the top-ranked prediction. Phrase-level metrics follow the standard SapBERT evaluation protocol, while concept-level macro metrics evaluate all concepts with equal weights regardless of their frequency. Positive performance gains are highlighted in light green.

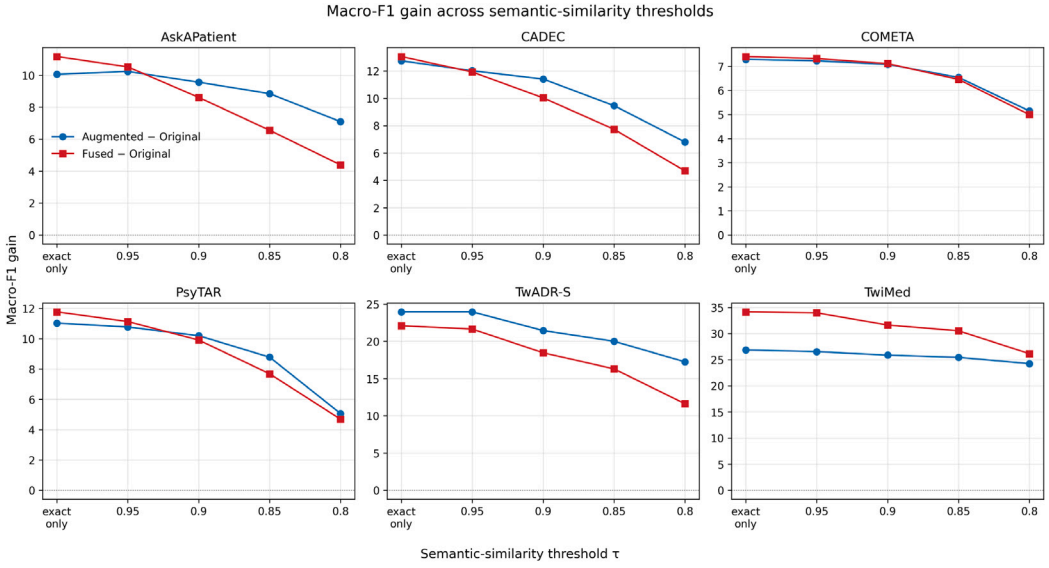
Dataset	Metric	Phrase-level (SapBERT)						Concept-level Macro		
		Org	Aug	Fused	Δ(Aug-Org)	Δ(Fused-Org)	Δ(Fused-Aug)	Org	Aug	Δ(Aug-Org)
AskAPatient	Acc@1 (%)	75.72	65.87	67.14	-9.85	-8.59	1.26	75.72	65.87	-9.85
	Acc@5 (%)	91.08	89.32	89.40	-1.76	-1.68	0.08	91.08	89.32	-1.76
	Recall@1 (%)	80.80	70.30	71.97	-10.50	-8.83	1.67	55.16	64.21	9.06
	Precision@1 (%)	75.70	65.87	67.13	-9.83	-8.57	1.26	54.55	63.86	9.31
	F1@1 (%)	78.17	68.02	69.46	-10.15	-8.71	1.44	53.52	61.63	8.12
CADEC	Acc@1 (%)	73.30	66.66	67.02	-6.63	-6.28	0.36	73.30	66.66	-6.63
	Acc@5 (%)	91.69	90.63	89.30	-1.06	-2.39	-1.33	91.69	90.63	-1.06
	Recall@1 (%)	77.75	70.39	71.96	-7.36	-5.79	1.57	60.10	65.79	5.69
	Precision@1 (%)	73.30	66.66	67.01	-6.64	-6.29	0.35	59.60	65.48	5.87
	F1@1 (%)	75.46	68.47	69.40	-6.99	-6.06	0.93	58.82	63.13	4.31
COMETA	Acc@1 (%)	45.07	71.41	65.73	26.34	20.66	-5.68	45.07	71.41	26.34
	Acc@5 (%)	61.27	89.92	85.93	28.65	24.65	-4.00	61.27	89.92	28.65
	Recall@1 (%)	69.84	76.82	73.41	6.98	3.57	-3.41	34.93	69.94	35.01
	Precision@1 (%)	45.06	71.41	65.72	26.35	20.66	-5.69	32.47	69.94	37.48
	F1@1 (%)	54.78	74.02	69.35	19.24	14.57	-4.67	32.83	67.60	34.78
PsyTAR	Acc@1 (%)	46.77	59.43	62.31	12.66	15.54	2.89	46.77	59.43	12.66
	Acc@5 (%)	59.84	85.61	84.24	25.77	24.40	-1.36	59.84	85.61	25.77
	Recall@1 (%)	73.26	65.30	70.13	-7.96	-3.13	4.83	25.18	58.87	33.70
	Precision@1 (%)	46.52	59.43	62.26	12.91	15.74	2.83	25.41	58.51	33.10
	F1@1 (%)	56.86	62.22	65.96	5.36	9.10	3.74	24.03	56.16	32.13
TwADR-S	Acc@1 (%)	43.19	66.50	74.07	23.31	30.88	7.57	43.19	66.50	23.31
	Acc@5 (%)	71.21	87.04	89.87	15.83	18.66	2.83	71.21	87.04	15.83
	Recall@1 (%)	54.43	70.88	78.96	16.45	24.53	8.08	26.25	65.95	39.69
	Precision@1 (%)	43.19	66.50	74.05	23.31	30.86	7.55	26.37	65.56	39.19
	F1@1 (%)	48.06	68.61	76.42	20.55	28.36	7.81	25.42	63.29	37.87
TwiMed	Acc@1 (%)	70.97	73.53	76.73	2.57	5.76	3.19	70.97	73.53	2.57
	Acc@5 (%)	77.59	91.87	90.49	14.28	12.90	-1.37	77.59	91.87	14.28
	Recall@1 (%)	87.28	77.21	82.44	-10.07	-4.84	5.23	61.88	73.39	11.50
	Precision@1 (%)	70.97	73.53	76.68	2.56	5.71	3.15	61.92	73.30	11.38
	F1@1 (%)	78.13	75.33	79.46	-2.80	1.33	4.13	61.40	71.01	9.62

Table 7
Generalization of the data augmentation and fusion benefit across models. Following the two-panel structure of Table 6, we report phrase-level Top-1 and Top-5 accuracy (Acc@1, Acc@5) on the original (Org), augmented (Aug), and fused datasets, and concept-level Macro-F1 on the original and augmented datasets. SapBERT values are reproduced from Table 6; KNN-BioEL and KRISSBERT are evaluated under the same cross-validation protocol. Cells where Aug or Fused exceed the corresponding Org value are highlighted in light green.

Dataset	Model	Phrase-level						Concept-level Macro	
		Acc@1 (%)			Acc@5 (%)			Macro-F1 (%)	
		Org	Aug	Fused	Org	Aug	Fused	Org	Aug
AskAPatient	SapBERT	75.72	65.87	67.14	91.08	89.32	89.40	53.52	61.63
	KNN-BioEL	76.41	65.04	66.97	92.42	88.14	87.04	54.25	64.32
	KRISSBERT	74.97	65.09	65.86	89.75	87.84	87.98	52.05	60.27
CADEC	SapBERT	73.30	66.66	67.02	91.69	90.63	89.30	58.82	63.13
	KNN-BioEL	73.03	68.94	67.98	90.76	91.47	88.72	62.45	63.97
	KRISSBERT	69.88	65.71	65.27	88.95	89.34	88.02	57.69	62.07
COMETA	SapBERT	45.07	71.41	65.73	61.27	89.92	85.93	32.83	67.60
	KNN-BioEL	45.62	73.02	67.77	62.42	90.34	81.72	34.01	70.02
	KRISSBERT	43.12	70.78	64.50	60.35	87.37	84.50	31.06	64.27
PsyTAR	SapBERT	46.77	59.43	62.31	59.84	85.61	84.24	24.03	56.16
	KNN-BioEL	48.39	59.72	63.63	61.81	87.41	85.23	25.65	56.45
	KRISSBERT	45.63	58.57	61.46	58.12	86.40	84.16	22.69	54.38
TwADR-S	SapBERT	43.19	66.50	74.07	71.21	87.04	89.87	25.42	63.29
	KNN-BioEL	44.63	62.03	75.54	72.72	83.55	91.34	25.81	61.79
	KRISSBERT	41.98	65.33	72.85	69.01	84.85	87.67	24.22	62.07
TwiMed	SapBERT	70.97	73.53	76.73	77.59	91.87	90.49	61.40	71.01
	KNN-BioEL	72.08	75.34	77.45	78.87	93.85	91.54	62.64	71.93
	KRISSBERT	69.30	71.82	75.34	75.79	91.05	87.96	62.25	68.94



(a) Varying training-set size



(b) Varying semantic-similarity threshold

Fig. 10. Concept-level Macro-F1 gain over the original setting (Augmented − Original and Fused − Original) for each dataset. (a) The training set of every setting is subsampled to 20%–100% of its phrases. (b) "Exact only" removes only exact duplicates; at $\tau = 0.95, 0.9, 0.85, 0.8$ we additionally remove every training phrase whose cosine similarity to a same-concept test phrase exceeds τ.

Table 8
Statistical significance of the augmentation gain in concept-level Macro-F1 across the six datasets, assessed with a paired t-test. The improvement of the augmented over the original setting is significant ($p < 0.05$) for all three models.

Model	Mean ΔMacro-F1	p-value
SapBERT	+21.13	0.0197*
kNN-BioEL	+20.61	0.0218*
KRISSBERT	+20.34	0.0232*

the importance of data quality evaluation and enhancement. The results also prove that data fusion offers a robust and reliable mechanism for improving MCN performance.

5.2. Effects of augmentation across datasets

The substantial performance gains observed in COMETA, PsyTAR, TwADR-S, and TwiMed demonstrate the effectiveness of our quality-aware augmentation pipeline in addressing the long-tail characteristics of these datasets. All four datasets suffer from severe class imbalance, where many concepts are represented by only a few, and often noisy phrases. Our quality-aware augmentation mitigates this issue by generating additional, semantically controlled paraphrases for rare concepts. By expanding each concept's representation space, the augmented data improves concept-level coverage and reduces imbalance.

The results in Table 6 validate this approach. In COMETA, PsyTAR, and TwADR-S, concept-level macro F1 increases by up to 38%, while

Table 9
Accuracy improvement on long-tail concepts with less than 20 phrases before and after augmentation.

# Phrase	# Concepts	Mean Δ Acc@1 (%)	Mean Δ Acc@5 (%)
1	337	44.74	59.50
2	1,708	35.19	48.06
3	297	27.97	42.23
4	620	26.74	42.13
5	1,125	26.39	39.42
6	240	32.68	43.73
7	392	22.03	34.85
8	664	21.84	34.33
9	270	19.84	31.25
10	280	23.31	37.03
11	231	22.85	35.13
12	132	26.24	47.29
13	169	25.40	39.52
14	154	16.30	31.13
15	135	20.90	28.50
16	80	16.81	44.01
17	102	15.38	32.92
18	54	32.46	48.25
20	20	9.10	34.74

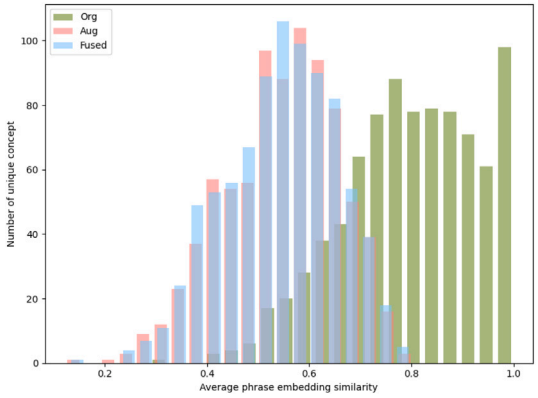


Fig. 11. Phrase embedding similarity comparison for the shared concepts in original (Org), Augmented (Aug), and fused (Fused) AskAPatient and CADEC datasets.

TwiMed shows a substantial 10% improvement. These consistent gains across long-tailed datasets provide strong evidence that augmentation strengthens the model's ability to handle rare concepts by enriching their semantic representations and reducing imbalance. Overall, the improvements in concept-level macro performance confirm that quality-aware augmentation directly alleviates long-tail issues and enhances model robustness in datasets where sparse concept coverage is the primary limiting factor.

However, the performance drops in the AskAPatient and CADEC datasets. We discuss these two datasets together for the following reasons. Firstly, the two datasets present highly similar patterns as observed from Table 3, Fig. 6, Table 5, Fig. 7, and Fig. 8. They have highly overlapping concepts as they are both sourced from the same website. Fig. 11 further shows that for the shared concepts, the phrases from the two original datasets have very high embedding similarity, indicating co-existing data quality issues.

As shown in Fig. 5, the original AskAPatient and CADEC datasets exhibit higher embedding similarity than the other datasets. This indicates that many phrases describing the same concept are semantically repetitive, even if they are not exactly duplicates, such as 'tired,' 'feeling tired,' and 'very tired.' Such low variety produces a narrow feature space, leading to the overclaimed accuracy scores, as reflected in Table 6. These inflated results do not necessarily reflect genuine understanding. Consequently, the performance tends to decrease when more diverse augmented data are introduced. This is consistent with prior findings [1].

To verify this explanation, we further conducted experiments with different semantic variety settings. We curated augmented datasets with average phrase similarity ranging from 0.51 to 0.65. This is achieved by adjusting the initial augmentation prompt from 'try to cover a variety of expressions' to a set of prompts that require the augmented phrases to be similar to the original datasets at different levels. Fig. 12 shows that as the augmented datasets become more similar to the original ones, both Top-1 and Top-5 accuracy rise steadily.However, this increase should not be interpreted as a true performance improvement. It neither reflects superior data quality nor stronger model capability. Rather, it demonstrates the fitting problem caused by data quality issues.

The original AskAPatient and CADEC datasets show higher accuracy in Table 6. However, Fig. 12 demonstrates that augmented datasets with lower or comparable similarity actually outperform the original datasets. This finding reveals how data augmentation enhances DQ in this task. (1) It improves coverage of rare concepts by adding samples for underrepresented classes, reducing the influence of frequent, repetitive phrases that cause overfitting. (2) Within concepts, it adds controlled semantic variance, widening the representation space and helping the model generalize rather than memorize fixed patterns. (3) It can help moderate similarity to align training and test distributions without recreating redundancy.

Together, these results reveal several key insights. (1) The high accuracy of the original AskAPatient and CADEC datasets is largely an artifact of overfitting, as their redundancy allows models to memorize repetitive phrasing instead of learning true semantics. (2) Data augmentation improves quality by enriching diversity and extending long-tail coverage, allowing the model to learn from rare and underrepresented concepts rather than relying on dominant patterns. (3) The design of data quality evaluation metrics should be holistic, considering how these affect model performance. A dataset may appear diverse overall despite the long-tail issue, leaving rare cases poorly represented and resulting in poor model performance. (4) Finally, well-designed prompt-based augmentation can help balance redundancy and diversity. By empirically adjusting the similarity target and generation prompt, our augmentation process can optimize data variety to fulfill our experiment requirements.

5.3. Effects of fusion across datasets

Fusion consistently enhances or stabilizes model performance across all datasets, including those negatively affected by augmentation. This data fusion design increases performance from three key aspects: (1) it enables concept-level complementarity. Each dataset contributes unique phrase variants and concept distributions, and merging these samples expands the overall coverage. Sparse concepts in a single dataset can be better represented in the fused resource. (2) Fusion introduces a noise-averaging effect. Noisy or inconsistent mappings from one dataset are counterbalanced by cleaner or more consistent examples from others. This stabilizing effect mitigates the performance degradations observed in augmented AskAPatient and CADEC. Although their augmented data introduces noise, the fused versions include higher-quality examples from COMETA, PsyTAR, TwADR-S, and TwiMed, allowing the model to learn more reliable representations. (3) Fusion enhances semantic diversity without relying solely on synthetic augmentation. Natural variation across datasets provides authentic user expressions from different subsets of social media health platforms.

To further investigate whether these gains reflect a genuine effect of augmentation and fusion rather than properties of the enriched data, we conduct two additional experiments that vary the amount of training data and the semantic similarity between the training and test phrases. The first experiment (Fig. 10(a)) varies the training-set size: for each dataset, it reports the augmentation and fusion gains over

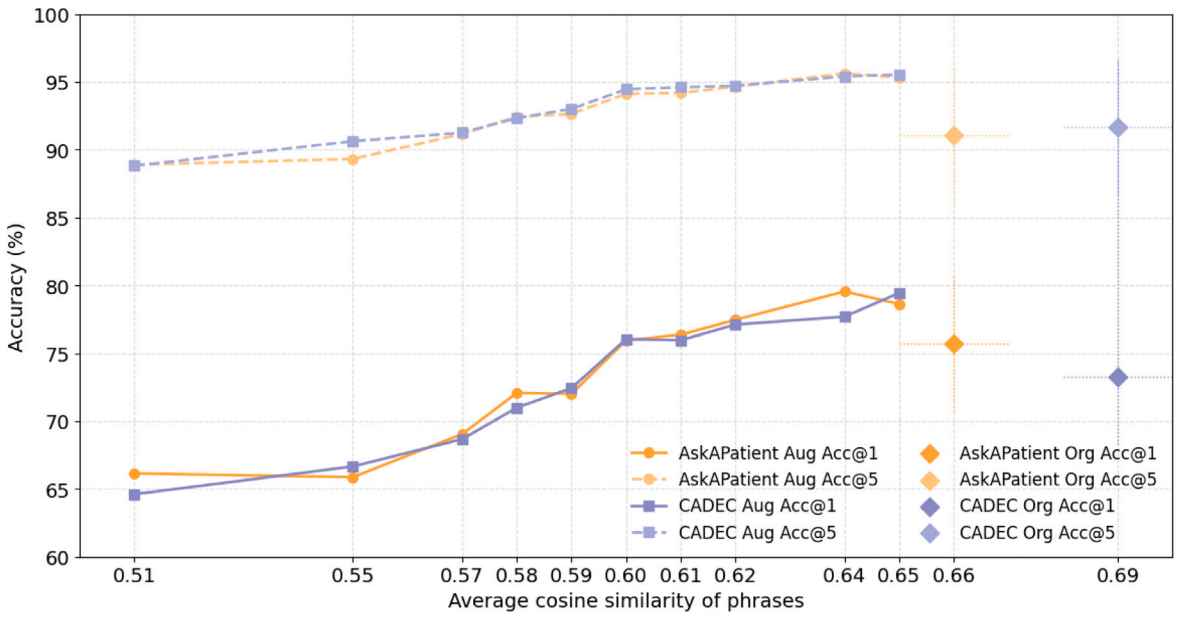


Fig. 12. Comparison of Top-1 (Acc@1) and Top-5 (Acc@5) accuracy tested on original data and augmented data with different semantic variety for AskAPatient (orange set colors) and CADEC (purple set colors). Solid and dashed lines represent the Acc@1 and Acc@5 results on the augmented datasets. Diamond markers at $x = 0.66$ and $x = 0.69$ on the right denote the original dataset performances for AskAPatient and CADEC, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the original setting as the training set grows from 20% to 100% of its phrases. The gain is largest when the training set is small and shrinks as more phrases are added, with the fusion gain on COMETA falling from about 10.5 to 7.4 Macro-F1 points and the augmentation gain on AskAPatient falling from about 17.6 to 10.1. More data therefore reduces the gain rather than increasing it, so the improvement is not an effect of training-set size but comes from the coverage that augmentation and fusion add where the original data are sparse. The second experiment (Fig. 10(b)) varies the semantic similarity between the training and test phrases: from left to right, it removes training phrases that are increasingly similar to the test set, from exact duplicates to every phrase whose cosine similarity to a same-concept test phrase exceeds $\tau = 0.8$. Removing the most similar phrases, the near-duplicates most likely to reflect overlap, barely changes the gain: even after all near-duplicates are removed ($\tau = 0.9$), the augmentation gain drops by at most 2.5 points, for instance from 7.3 to 7.1 on COMETA and from 26.9 to 25.9 on TwiMed. The gain drops noticeably only at the strictest setting ($\tau = 0.8$), and even then it stays clearly positive on every dataset. The phrases most similar to the test set therefore contribute little to the gain, so it does not arise from the similarity between the training and test data. Together, these two experiments show that the augmentation and fusion gains are a genuine effect rather than an artifact of the larger size or the higher similarity of the enriched data.

Lastly and importantly, data fusion should be quality-aware. Fusion is beneficial after addressing data quality issues through augmentation. After data quality enhancement, data fusion contributes the most to smaller or sparser datasets, such as PsyTAR, TwADR-S, and TwiMed, where combining augmented phrases across datasets significantly increases expression coverage. In these cases, fusion builds on the strengthened representations created by augmentation, enabling the model to leverage complementary phrasing and context across datasets.

6. Conclusion

We propose a quality-aware data fusion and enhancement framework for Medical Concept Normalization (MCN). To the best of our knowledge, this work is the first to demonstrate that multi-source dataset fusion can substantially improve MCN performance, and we validated this through extensive experiments on six social media datasets. Our data quality evaluation revealed several issues in existing datasets, including outdated or inconsistent terminology, low semantic variety, and severe class imbalance. Although certain validity inconsistencies may not directly harm model accuracy due to the robustness of embedding-based normalization, they nonetheless create ambiguity in concept definitions and limit dataset reliability. Additionally, we observed that MCN datasets share overlapping concepts, further supporting the feasibility and necessity of fusion as a means to expand conceptual coverage.

However, naïvely fusing datasets without addressing data quality would amplify existing biases, such as overfitting caused by low semantic variety in AskAPatient and CADEC, and long-tail cases in COMETA, PsyTAR, TwADR-S, and TwiMed. Therefore, we introduced a quality enhancement stage prior to fusion. LLM-based controlled-variety data augmentation is applied to increase coverage for rare concepts. Our experiments show that this approach not only improves the representation of long-tail concepts but also enables more effective and balanced data fusion, especially for datasets with limited phrases.

This study provides both a practical, data-driven, and quality-aware framework for MCN, and a new, larger-scale fused dataset resource for the community. Based on this, our future work aims to: (1) extend the framework to MCN datasets beyond social media, such as clinical notes or biomedical articles; (2) explore multi-modal fusion by incorporating medical images to enrich information; (3) develop targeted strategies for difficult cases, such as ambiguous or multi-symptom expressions; and (4) study prompt design in LLM-based augmentation, as our findings indicate that prompt-controlled semantic variation plays a critical role in achieving balanced data quality.

CRediT authorship contribution statement

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

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve the language and grammar of the manuscript. After using this tool, the authors reviewed and edited the content as needed. The authors take full responsibility for the content of the published article.

Declaration of competing interest

Junhua Ding reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available at GitHub Repository.

References

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

[2] N. Pattisapu, S. Patil, G. Palshikar, V. Varma, Medical concept normalization by encoding target knowledge, in: Machine Learning for Health Workshop, PMLR, 2020, pp. 246–259.

[3] F.A. Bernardi, D. Alves, N. Crepaldi, D.B. Yamada, V.C. Lima, R. Rijo, Data quality in health research: integrative literature review, J. Med. Internet Res. 25 (2023) e41446.

[4] H. Chen, J. Chen, J. Ding, Data evaluation and enhancement for quality improvement of machine learning, IEEE Trans. Reliab. 70 (2) (2021) 831–847, http://dx.doi.org/10.1109/TR.2021.3070863.

[5] A. Wang, C. Liu, J. Yang, C. Weng, Fine-tuning large language models for rare disease concept normalization, J. Am. Med. Informatics Assoc. 31 (9) (2024) 2076–2083.

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

[7] 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, 2017, pp. 462–469, http://dx.doi.org/10.1109/ICHI.2017.59.

[8] 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.

[9] M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A corpus for medical entity linking in the social media, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 3122–3137, http://dx.doi.org/10.18653/v1/2020.emnlp-main.253, URL: https://aclanthology.org/2020.emnlp-main.253/.

[10] 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.

[11] 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.

[12] Y.F. Luo, W. Sun, A. Rumshisky, MCN: A comprehensive corpus for medical concept normalization, J. Biomed. Informatics 92 (2019) 103132, http://dx.doi.org/10.1016/j.jbi.2019.103132, URL: https://www.sciencedirect.com/science/article/pii/S1532046419300504.

[13] S.R. Tuly, S. Ranjbari, E.A. Murat, S. Arslanturk, From silos to synthesis: A comprehensive review of domain adaptation strategies for multi-source data integration in healthcare, Comput. Biol. Med. 191 (2025) 110108.

[14] P. Yang, H. Wang, Y. Huang, S. Yang, Y. Zhang, L. Huang, Y. Zhang, G. Wang, S. Yang, L. He, et al., LMKG: A large-scale and multi-source medical knowledge graph for intelligent medicine applications, Knowl.-Based Syst. 284 (2024) 111323.

[15] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, S. Ananiadou, MentaLLaMA: interpretable mental health analysis on social media with large language models, in: Proceedings of the ACM on Web Conference 2024, 2024, pp. 4489–4500.

[16] F. Gallego, P. Ruas, F.M. Couto, F.J. Veredas, Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero-and few-shot scenarios, Knowl.-Based Syst. 314 (2025) 113211.

[17] A. Singh, S. Krishnamoorthy, J.E. Ortega, NeighBERT: Medical entity linking using relation-induced dense retrieval, J. Heal. Informatics Res. 8 (2) (2024) 353–369.

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

[19] G. Luo, N. Shi, G. Wang, B. Tang, Contextual information contributes to biomedical named entity normalization, J. Biomed. Informatics 165 (2025) 104806.

[20] Y. Fan, K. Xue, Z. Li, X. Zhang, T. Ruan, An LLM-based framework for biomedical terminology normalization in social media via multi-agent collaboration, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 10712–10726.

[21] J. Xiong, L. Yu, X. Niu, Y. Leng, XRR: Extreme multi-label text classification with candidate retrieving and deep ranking, Inform. Sci. 622 (2023) 115–132.

[22] 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.

[23] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, J. Han, Tuning language models as training data generators for augmentation-enhanced few-shot learning, in: International Conference on Machine Learning, PMLR, 2023, pp. 24457–24477.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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. Informatics Assoc. 25 (10) (2018) 1274–1283.

[29] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, et al., Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 173–198.

[30] 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.

[31] R.B. Correia, I.B. Wood, J. Bollen, L.M. Rocha, Mining social media data for biomedical signals and health-related behavior, Annu. Rev. Biomed. Data Sci. 3 (1) (2020) 433–458.

[32] 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.

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

[34] K.S. Kalyan, S. Sangeetha, Social media medical concept normalization using RoBERTa in ontology enriched text similarity framework, in: Proceedings of Knowledgeable NLP: The First Workshop on Integrating Structured Knowledge and Neural Networks for NLP, Association for Computational Linguistics, Suzhou, China, 2020, pp. 21–26, http://dx.doi.org/10.18653/v1/2020.knlp-1.3.

[35] 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. Informatics Assoc. 28 (3) (2021) 516–532.

[36] 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.

[37] M. Drosou, H.V. Jagadish, E. Pitoura, J. Stoyanovich, Diversity in big data: A review, Big Data 5 (2) (2017) 73–84.

[38] Y. Zhang, J.K. Lee, J.C. Han, R.T.H. Tsai, Task reformulation and data-centric approach for Twitter medication name extraction, Database 2022 (2022) baac067.
[39] 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.
[40] 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.
[41] S. Lin, Y. Wang, L. Zhang, Y. Chu, Y. Liu, Y. Fang, M. Jiang, Q. Wang, B. Zhao, Y. Xiong, et al., MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism, Brief. Bioinform. 23 (1) (2022) bbab421.
[42] Y. Zhou, H. Chen, K. Sha, A novel multi-layer task-centric and data quality framework for autonomous driving, 2025, arXiv preprint arXiv:2506.17346.
[43] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, T. Schaeffter, The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review, NPJ Digit. Med. 7 (1) (2024) 203.
[44] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.T. Lu, C.C. Aggarwal, J. Pei, Y. Zhou, A comprehensive survey on data augmentation, IEEE Trans. Knowl. Data Eng. (2025) 1–20, http://dx.doi.org/10.1109/TKDE.2025.3622600.
[45] J. Xie, L. Sun, Y.F. Zhao, On the data quality and imbalance in machine learning-based design and manufacturing—A systematic review, Engineering 45 (2) (2025) 105–131.
[46] J. Clusmann, F.R. Kolbinger, H.S. Muti, Z.I. Carrero, J.N. Eckardt, N.G. Laleh, C.M.L. Löffler, S.C. Schwarzkopf, M. Unger, G.P. Veldhuizen, et al., The future landscape of large language models in medicine, Commun. Med. 3 (1) (2023) 141.
[47] C.P. Phan, B. Phan, J.H. Chiang, Optimized biomedical entity relation extraction method with data augmentation and classification using GPT-4 and Gemini, Database 2024 (2024) baae104.
[48] I. Jahan, M.T.R. Laskar, C. Peng, J.X. Huang, A comprehensive evaluation of large language models on benchmark biomedical text processing tasks, Comput. Biol. Med. 171 (2024) 108189.
[49] N.J. Dobbins, Generalizable and scalable multistage biomedical concept normalization leveraging large language models, Res. Synth. Methods (2024) 1–12.
[50] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A corpus of adverse drug event annotations, J. Biomed. Informatics 55 (2015) 73–81.
[51] 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.
[52] X. Dai, S. Karimi, A. Sarker, B. Hachey, C. Paris, MultiADE: a multi-domain benchmark for adverse drug event extraction, J. Biomed. Informatics 160 (2024) 104744.
[53] Q. Peng, Y. Cai, J. Liu, Q. Zou, X. Chen, Z. Zhong, Z. Wang, J. Xie, Q. Li, Integration of multi-source medical data for medical diagnosis question answering, IEEE Trans. Med. Imaging 44 (3) (2025) 1373–1385, http://dx.doi.org/10.1109/TMI.2024.3496862.
[54] M. Torii, K. Wagholikar, H. Liu, Using machine learning for concept extraction on clinical documents from multiple data sources, J. Am. Med. Informatics Assoc. 18 (5) (2011) 580–587.
[55] 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.
[56] R. Xu, W. Shi, Y. Yu, Y. Zhuang, B. Jin, M.D. Wang, J. Ho, C. Yang, Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2024, pp. 754–765.
[57] D. Galea, I. Laponogov, K. Veselkov, Exploiting and assessing multi-source data for supervised biomedical named entity recognition, Bioinformatics 34 (14) (2018) 2474–2482.
[58] Z. Zhang, W. Wu, D. Wu, A multi-mode learning behavior real-time data acquisition method based on data quality, in: 2021 2nd International Symposium on Computer Engineering and Intelligent Communications, ISCEIC, IEEE, 2021, pp. 64–69.
[59] Y. Sun, T. Lu, N. Gu, A method of electronic health data quality assessment: Enabling data provenance, in: 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design, CSCWD, IEEE, 2017, pp. 233–238.
[60] 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.
[61] 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.
[62] M. Rithani, R.P. Kumar, A. Ali, A dynamic ensemble learning based data mining framework for medical imbalanced big data, Knowl.-Based Syst. 310 (2025) 112947.
[63] W.H. Weng, J. Deaton, V. Natarajan, G.F. Elsayed, Y. Liu, Addressing the real-world class imbalance problem in dermatology, in: Machine Learning for Health, PMLR, 2020, pp. 415–429.
[64] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025, arXiv preprint arXiv:2507.06261.
[65] 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.
[66] 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.