LLM-Assisted Ontology Mapping for Semantic Interoperability in Structured Biomedical Data

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF TECHNOLOGY

BY

PRASANNA KUMAR S
(MT23234)

COMPUTATIONAL BIOLOGY
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY
DELHI

NEW DELHI- 110020

June 2025

THESIS CERTIFICATE

This is to certify that the thesis titled “LLM-Assisted Ontology Mapping for Semantic Interoperability in Structured Biomedical Data”, submitted by Prasanna Kumar S, to the Indraprastha Institute Information of Technology, Delhi, for the award of the degree of Master of Technology, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Tavpritesh Sethi
Thesis Supervisor
Professor
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 23rd June 2025

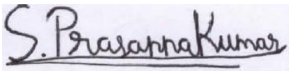
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Supervisor, Dr. Tavpritesh Sethi, for their unwavering guidance, support, and encouragement throughout this journey. Their insights and expertise were invaluable to this work.

I am also profoundly thankful to Akshaya Devadiga and Pradeep Singh for their mentorship and thoughtful advice. Their contributions and constant support played a crucial role in the completion of this thesis.

Thank you all for making this achievement possible.

Sincerely,



(Prasanna Kumar S, MT23234)

ABSTRACT

KEYWORDS: Biomedical Data ; Semantic Interoperability ; FAIR Data Principles ; Data Standardization ; SNOMED CT ; Ontology Mapping ; Structured Data ; Large Language Models

The exponential growth of biomedical data promises new insights, but semantic heterogeneity and inconsistent metadata limit reuse. In practice, many publicly available datasets (e.g., tabular datasets from Figshare or Zenodo) are annotated with non-standardized field names, violating Findable Accessible Interoperable Reusable criteria (FAIR). To bridge this gap, we propose a framework for FAIR Assessment using Ontology Mapping and large language models (LLMs), that assesses and enhances interoperability of such “not-so-FAIR” datasets. First, we quantify dataset FAIRness by mapping variables to standard clinical terms - Systematized Medical Nomenclature for Medicine Clinical Terms (SNOMED CT) – a comprehensive ontology widely used for semantic interoperability. Then we explore the use of large language models – specifically Mistral and LLaMA – to improve SNOMED CT term mapping coverage and disambiguation for dataset fields. We prompt these large language models with field context and compare their predicted SNOMED terms to ground-truth concepts (baseline: Medical Concept Annotation Tool). Our experiments on diverse clinical datasets show that large language models can significantly augment automated ontology mapping and reduce semantic mismatches. Taken together, this work presents a principled approach that integrates ontology-based FAIR assessment with LLM-driven harmonization to close the semantic gap in biomedical data integration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
BD2K	Big Data to Knowledge
BERT	Bidirectional Encoder Representations from Transformers
caBIG	Cancer Biomedical Informatics Grid
CDM	Common Data Model
COVID-19	Coronavirus Disease 2019
cTAKES	clinical Text Analysis and Knowledge Extraction System
DCAT-AP	Data Catalog Vocabulary - Application Profile
EHR	Electronic Health Record
EOSC	European Open Science Cloud
ETL	Extract, Transform, Load
FAIR	Findable Accessible Interoperable Reusable
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
ICD	International Classification of Diseases
i2b2	Informatics for Integrating Biology and the Bedside
LLM	Large Language Model
LOINC	Logical Observation Identifiers Names and Codes
MedCAT	Medical Concept Annotation Tool
NCI	National Cancer Institute
NIH	National Institutes of Health
OAEI	Ontology Alignment Evaluation Initiative
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OSF	Open Science Framework
OWL	Web Ontology Language
RAG	Retrieval-Augmented Generation

SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
UMLS	Unified Medical Language System
U.S.	United States

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Statement

Biomedical research increasingly relies on integrating diverse datasets, but semantic interoperability remains a major challenge [1][2]. The FAIR Guiding Principles explicitly call for metadata that enable machines to find, access, and integrate data [1], yet in practice many datasets lack standardized semantics. For example, in clinical and genomic studies different teams may label the same concept with inconsistent terms or units. One study noted that while researchers routinely share data, these datasets were often "not necessarily findable, accessible, interoperable, or reusable (FAIR)" [3]. In large-scale collaborative settings such as the Observational Health Data Sciences and Informatics (OHDSI) network, billions of patient records exist in distributed silos, and these can only be meaningfully leveraged if mapped to a common data model and ontology [4]. A recent analysis of FAIR principles further highlights that *"semantic interoperability necessitates comprehensive sets of entity mappings and schema crosswalks"* to ensure that meaning is preserved across data sources [2]. In healthcare, achieving this requires linking local data elements to a controlled vocabulary: SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [5] has become a de facto standard for clinical terminology, and using SNOMED mappings is a key strategy for ensuring semantic interoperability [6].

In this work we focus on clinical datasets drawn from open repositories such as Figshare and Zenodo [7][8]. These structured datasets contain valuable clinical information, but they are often "not-so-FAIR" – field names are often inconsistent, descriptions are brief, and key variables are unlabeled with ontologies [3][9]. As a result, fusing multiple such datasets (for meta-analysis or evidence synthesis) is nontrivial. To address this, we propose a Large Language Model (LLM) assisted ontology-driven approach that standardizes field-level semantics.

1.2 Challenges of Semantic Heterogeneity in Biomedical Data

Biomedical datasets exhibit semantic heterogeneity at multiple levels. Different datasets may use different schemas (variable names and structures) for the same concepts. For instance, one diabetes dataset might record patient weight as "Weight_kg", another as "Wt_kg", and a third might code it under a column named "B1". Without manual curation or mapping, an analyst cannot tell these refer to the same feature. Even within the same schema, value coding

can vary: diagnoses might be coded by free text ("Type 2 Diabetes"), by ICD-10 codes [10], or by SNOMED CT concept IDs [5]. Crucially, terminology ambiguity is common: the word "cold" could mean "common cold" or simply the temperature condition, and acronyms (e.g. "BP" for blood pressure) may be interpreted differently in different data sources.

These issues lead to semantic mismatch when merging data. A downstream analysis may wrongly treat two semantically identical fields as unrelated, or wrongly conflate two different concepts. Furthermore, unsynchronized use of ontologies means identical underlying conditions could be annotated by different codes. For example, one dataset may label a cardiovascular condition with SNOMED [5] code 413838009 ("Chronic ischemic heart disease") while another uses ICD-10 [10] code I25.1 ("Atherosclerotic heart disease of native coronary artery"). Without a prior map or automated reasoning, these would not be recognized as referring to the same disease. This heterogeneity is exacerbated by missing or inconsistent metadata: often, publicly shared datasets lack detailed descriptions, making automated interpretation difficult. As one review notes, many shared datasets "are not harmonised at the participant level" and there are few restrictions on design or coding, making cross-dataset analyses "resource-intensive" [9].

In short, the semantic interoperability problem in biomedical data requires reconciling diverse schemas, normalizing values to common ontologies, and disambiguating terms. Traditional solutions rely on expert-driven mappings (e.g. manually mapping ICD [10] to SNOMED [5]) or constrained common data models. However, these are laborious and may not scale to the flood of open data. This motivates exploring automated LLM-driven approaches to assist in disambiguating and aligning heterogeneous biomedical data with standardized vocabularies.

1.3 FAIR Principles and Their Gaps in Open Repositories

The FAIR data principles (Findable, Accessible, Interoperable, Reusable) were formulated to improve data management and sharing. FAIR aims to ensure that scientific data are machine-actionable and richly documented [3]. For example, data should have persistent identifiers (Findable), use open standards and be easily retrievable (Accessible), employ shared vocabularies or ontologies for consistency (Interoperable), and include licensing and provenance metadata (Reusable). In biomedical research, FAIRness is particularly critical for enabling cross-study analyses and reproducibility.

Despite wide endorsement of FAIR, many published datasets still fall short. Multiple analyses report that in practice, most datasets on repositories lack interoperable annotation. For instance, a 2023 survey of immune and infectious disease data found that researchers shared data widely, but these datasets were often "not necessarily findable, accessible, interoperable, or reusable (FAIR)" [3]. Similarly, a scoping review of COVID-19 data platforms noted an explosion of shared clinical and omics data, but most resources had only minimal semantic integration: data were often siloed by disease or data type and did not follow common standards [9]. Many found that increased data volume did not automatically translate into better FAIRness; semantic and technical interoperability remained the key bottleneck.

Open repositories like Figshare [7] or Zenodo [8] provide access to datasets, but they do not enforce harmonization. As Maxwell et al. observed, researchers frequently upload datasets to

such “data lake”-style platforms, but these stores “house datasets that are not harmonised at the participant level” [9]. There is little restriction on how data are coded or described, so two “FAIR” datasets may still speak different languages. Even when metadata fields exist, they are often incomplete or inconsistent (e.g. missing standardized vocabularies). Therefore, open science efforts still leave a semantic gap: data may be technically accessible, but require substantial post-hoc processing to become truly interoperable.

This thesis focuses on closing that gap in practice. We accept that many biomedical datasets (especially those from open repositories) are not fully FAIR and aim to develop methods to effectively bridge their inconsistencies. In particular, we investigate whether LLMs and advanced semantic techniques can automatically infer connections and standardize terminology, thereby enhancing interoperability despite the data’s initial shortcomings.

1.4 Thesis Scope, Objectives, and Contributions

The scope of this thesis is the semantic harmonization of structured datasets from open repositories – Figshare & Zenodo. These datasets vary in schema and coding, representing a prototypical “not-so-FAIR” scenario. We impose minimal preconditions: the data are freely available (FAIR Findable/Accessible), but not curated to common standards (lacking Interoperable/Reusable metadata). The main objectives are: (1) to assess the current state of interoperability of structured datasets from open repositories and develop a workflow for linking heterogeneous dataset fields to a common ontology (we choose SNOMED CT as the target vocabulary); and (2) to evaluate the role of large language models (Mistral & LLaMA) in this process versus a traditional tool (MedCAT). Specifically, we propose using LLMs for ontology matching tasks: given a column name, prompt the LLM to suggest the best matching SNOMED concept.

The contributions of this thesis are:

1. **Ontology-driven FAIR assessment:** We introduce a method to evaluate dataset FAIRness by mapping fields to SNOMED CT, quantifying semantic coverage and interoperability with a standardized medical vocabulary.

2. **LLM-assisted ontology mapping:** We leverage recent foundation models to improve term mapping. By prompting LLMs with field context, we achieve higher accuracy in assigning SNOMED CT terms compared to MedCAT baseline.

3. **Empirical validation on clinical data:** We apply our framework to multiple real-world clinical datasets, demonstrating that our approach significantly reduces semantic mismatches and makes the data more interoperable for downstream analysis.

By uniting traditional interoperability techniques with modern language models, this thesis provides a scalable path toward closing the semantic gap in healthcare data. We show that even “messy” open datasets can be brought closer to FAIR compliance through systematic harmonization, enabling more reliable multi-dataset studies and evidence generation.

1.5 Thesis Organization

The remainder of the thesis is organized as follows. Chapter 2 reviews relevant background and related work, including FAIR data principles in biomedicine, the role of ontologies and controlled vocabularies (with emphasis on SNOMED CT), common data harmonization frameworks (OMOP, i2b2, etc.), similarity metrics for schema alignment, and prior efforts on concept extraction (MedCAT) and LLMs for ontology matching.

Chapter 3 details the datasets used and the preprocessing steps undertaken. It describes the acquisition of publicly available biomedical datasets and outlines a rigorous quality assessment and cleaning pipeline, including description and data availability checks, handling of compressed and multi-sheet files, structured format filtering, and normalization of both dataset descriptions and field names.

Chapter 4 presents the core methodology for quantifying the FAIR-ness of structured biomedical data through ontology mapping. It begins with mapping field names to SNOMED CT concepts, followed by leveraging large language models to assist ontology mapping through prompt-based matching, disambiguation of multiple candidate terms, and evaluation of the semantic coverage.

Chapter 5 discusses key findings, limitations, and implications of the proposed approach in the context of biomedical data interoperability.

Chapter 6 concludes the thesis with a summary of contributions and outlines potential directions for future research.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 FAIR Guiding Principles in Biomedical Research

The FAIR Guiding Principles laid out requirements for scientific data to be machine-actionable [1]. In summary, data should be Findable (rich metadata and persistent identifiers), Accessible (standardized access protocols), Interoperable (using shared vocabularies and ontologies), and Reusable (clear usage licenses and provenance). These principles have been widely adopted in biomedical research to promote data sharing and reuse. NIH and other funders now mandate data sharing policies that implicitly aim at FAIR compliance [3].

In the biomedical domain, numerous initiatives apply FAIR guidelines. Large-scale projects like the European Open Science Cloud (EOSC) aim to federate data along FAIR lines [11][12] (see Section 2.3.3). Academic repositories increasingly require FAIR metadata; for example, Dryad [13] and Open Science Framework (OSF) [14] encourage use of controlled ontologies. However, surveys show persistent gaps. Hughes et al. (2023) found that even in NIH-funded data, researchers often provide incomplete metadata, citing lack of incentives, which limits findability and reusability [3][9]. Similarly, a Lancet Digital Health study of COVID-19 platforms observed that many resources only partially implement FAIR: clinical data registries were less likely to use standardized ontologies compared to omics or imaging platforms [9]. The review concluded that "more data sharing does not equate to better data sharing," highlighting that adherence to FAIR (especially the Interoperable aspect) remains inadequate [9].

In practice, achieving FAIRness in biomedical research faces challenges: diverse data types, legacy systems, and privacy concerns. Interoperability is the hardest to fulfill – it requires not just open access but semantic standardization. This gap motivates research into methods for automatically improving interoperability. If tools (such as LLMs) can help annotate data with ontology terms and align vocabularies, they could substantially advance FAIR objectives by making heterogeneous data more Interoperable and thus more Reusable. This thesis explores such an approach in the context of open clinical datasets.

2.2 Semantic Interoperability and Ontologies

Semantic interoperability is the ability of systems to exchange data with unambiguous, shared meaning. In healthcare, achieving this requires ontologies and controlled vocabularies: predefined sets of terms and relationships that standardize how concepts are represented. Controlled vocabularies ensure that when different datasets use the same concept, they refer to the same underlying idea. For instance, by mapping various expressions of "type 2 diabetes" to a single code, datasets can be joined meaningfully. Value sets and coding systems

(e.g. ICD [10], LOINC [15]) are central: as one study notes, value set codes are often derived from commonly used clinical terminology standards such as SNOMED CT [5], ICD-9/10 [10], LOINC [15] and RxNorm [16][17]. Using controlled terms in data fields thus underpins interoperability and enables federated queries across systems.

2.2.1 Role of Controlled Vocabularies

A controlled vocabulary (or terminology) lists standardized terms and sometimes definitions for entities in a domain. It often enforces rules like single-meaning (uniqueness of concept per term) and may provide synonyms for human usability. In clinical research, these vocabularies take the form of medical ontologies. A key property is that vocabularies can be linked to or drawn from reference ontologies: if two datasets both label a variable using vocabulary terms (e.g. a SNOMED [5] or ICD code [10]), automated tools can recognize equivalence.

Ontologies go further than flat vocabularies by encoding hierarchical and associative relationships among terms. For example, an ontology might state that "Insulin is a treatment for Diabetes". Such semantic networks support more advanced reasoning and query capabilities. In practice, controlled terminologies are embedded in many Health Information Technology systems (EHRs, registries, analytics platforms) to standardize recording of diagnoses, procedures, findings, etc. As noted earlier [17], clinical quality measures in the U.S. rely on value sets drawn from multiple terminologies to ensure that data from different EHR systems are comparable. Similarly, data warehouses like i2b2 (Informatics for Integrating Biology & the Bedside) [18] use ontology-driven approaches to index patient data, allowing users to query across diverse sources by concept.

Thus, controlled vocabularies and ontologies are essential tools for overcoming semantic heterogeneity. They provide the "common language" that enables data from different origins to be interpreted consistently. In this thesis, we leverage this principle by mapping dataset fields to SNOMED CT concepts [5], effectively imposing a controlled vocabulary on the heterogeneous data.

2.2.2 SNOMED CT: Structure and Use Cases

SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [5] is one of the most comprehensive clinical ontologies. It contains hundreds of thousands of concepts covering diseases, findings, procedures, body structures, organisms, and more. SNOMED CT currently has on the order of 300,000–350,000 concepts, making it far larger than simpler classification systems [5]. As a clinical reference terminology, it is designed for detailed patient record encoding, enabling high granularity. SNOMED CT [5] is structured hierarchically: every concept can have one or more parent concepts (via "is-a" relationships), forming trees such as "Cardiovascular disease" → "Ischemic heart disease" → "Myocardial infarction". It also includes multiple other relationships (e.g. causative agent, morphology) and fully searchable descriptions (synonyms, definitions) for each concept [5].

In practice, SNOMED CT is used globally to standardize clinical documentation. It is mandated in many countries and is often embedded in EHR systems as the core reference terminology [5]. Clinicians record problem lists, diagnoses, procedures and findings using

SNOMED CT codes, ensuring a common framework across hospitals. This enables interoperability: two systems both using SNOMED CT can exchange coded data that unambiguously refer to the same clinical concepts. In research, SNOMED CT is frequently used to annotate datasets or to define cohorts (e.g. “all patients with SNOMED code 44054006 for Diabetes mellitus type 2”) [5].

Because of its size and expressiveness, SNOMED CT also supports advanced analytics. Its semantic network (viewable as an OWL ontology) allows reasoning—for instance, inferring that “bacterial pneumonia” is a kind of “infectious disease” via the hierarchy. Efforts have been made to represent SNOMED CT formally as an ontology (e.g. the SNOMED CT ontology based on Ontology for General Medical Science [19]), enabling consistency checking and integration with other ontologies. For our purposes, SNOMED CT serves as the target ontology for harmonizing data. By mapping dataset attributes and values to SNOMED concepts, we impose a shared semantic layer [5]. However, mapping is non-trivial because terms in raw data may not match SNOMED labels exactly (they may be free-text or use alternative phrasing). This is where techniques like string matching, word embeddings, or LLM prompting can help bridge the gap from raw terms to SNOMED concepts.

2.3 Data Harmonization Frameworks

Over the past decade, several data integration frameworks have been developed to address heterogeneity in clinical research. These include common data models (CDMs) and platforms that prescribe schemas and vocabularies for harmonized data sharing.

2.3.1 OMOP Common Data Model (OHDSI)

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is at the heart of the OHDSI (Observational Health Data Sciences and Informatics) network. It defines a standardized schema for patient-level observational data, encompassing tables for demographics, diagnoses, procedures, medications, etc. Crucially, OMOP relies on standardized vocabularies: health concepts are coded using a unified reference set (the OHDSI Standardized Vocabularies [4]) which integrates terminologies like SNOMED CT [5], RxNorm [16], LOINC [15], and ICD [10]. Reich et al. (2024) report that OHDSI’s vocabulary repository includes over 10 million concepts from 136 source vocabularies, and is mandatory for all sites in the network [4]. In practice, each data partner maps their local codes to OMOP’s standard concepts during ETL (Extract, Transform, Load), thus harmonizing semantics. This enables network-wide analyses on 2+ billion patient records across dozens of countries [4].

OMOP also leverages SNOMED CT [5] for conditions (diagnoses) and procedures, and RxNorm [16] for drugs. Its design abstracts away source data differences: once data conform to the CDM, research studies can be written once and executed at scale. OHDSI provides tools (like ATLAS) for cohort definition using these standard concepts [4]. However, converting existing datasets to OMOP can be labor-intensive. It usually involves building concept maps and handling data quality issues. Our work is similar in spirit: we seek to harmonize data via standard vocabularies, though on a smaller scale and using new automated

methods.

2.3.2 i2b2 and caBIG Platforms

i2b2 (Informatics for Integrating Biology and the Bedside) [18] is another widely-used platform for clinical data integration. It offers an ontology-driven query interface: patient data are stored in a star schema, and an associated ontology (typically derived from SNOMED or other terminologies) lets researchers select concepts to form cohorts. i2b2's ontology is hierarchical (folders of concepts) and can be customized. It has been deployed in many hospitals to support cohort discovery [18]. Like OMOP, i2b2 requires local data to be mapped to the ontology terms for interoperability. One advantage of i2b2 is that it can handle diverse data (clinical, genomic, etc.) within the same framework, using a consistent metadata layer [18].

caBIG (Cancer Biomedical Informatics Grid) [20] was an earlier NIH initiative (now retired) aimed at creating a federated network of tools and data for cancer research. It promoted standards and ontologies (e.g. National Cancer Institute (NCI) Thesaurus) to enable sharing. For example, the caBIG "Integration Hub" allowed disparate databases to link via service-oriented architecture [20]. CaBIG highlighted the importance of shared metadata standards and common data elements long before FAIR. While caBIG itself is less active today, its philosophy lives on in modern efforts: using enterprise service buses, APIs, and common ontologies to integrate clinical and genomic data across institutions. It demonstrated that large-scale integration requires both technical infrastructure and community-agreed semantics.

2.3.3 EOSC and Global FAIR Initiatives

The European Open Science Cloud (EOSC) is a pan-European initiative that aims to provide a federated environment for research data and tools [11][12]. Its vision is explicitly a "Web of FAIR Data and Services". EOSC promotes interoperability by connecting national research data infrastructures and ensuring they adopt common standards and services. For example, EOSC supports machine-actionable metadata standards (like DCAT-AP, DataCite) and encourages linking data with digital objects (software, workflows) to maximize reusability [11][12].

Globally, initiatives like GO (Global Open) FAIR [21] advocate bottom-up, practical FAIR implementation steps in health and biomedical contexts. The U.S. NIH's BD2K (Big Data to Knowledge) [22] and Data Commons [23] projects also align with FAIR. A recent workshop report notes that federated data spaces (akin to EOSC) can boost reproducibility by making data "accessible to researchers and machines" [11][12]. Despite these efforts, large-scale semantic interoperability remains work in progress. Many global FAIR efforts emphasize metadata and infrastructure, but the task of mapping actual clinical terminologies and ontologies is still largely manual. Our thesis contributes to this landscape by experimenting with LLMs as a scalable tool to semantically harmonize real-world biomedical datasets, potentially advancing FAIRness in practice.

2.4 Similarity Metrics for Schema and Data Alignment

To quantify how well two datasets or schemas align semantically, researchers use various similarity metrics. These metrics provide scores that indicate closeness of concepts, which is valuable for automated mapping.

2.4.1 Embedding-Based Approaches

One modern approach is to use distributed representations (embeddings) of terms or schema elements. For instance, words or phrases (column names, concept descriptions) can be embedded using language models or graph embeddings that capture semantic meaning. Similarity between two elements is then measured by vector distance, most commonly cosine similarity. If v_1 and v_2 are embedding vectors of two terms, their cosine similarity (the cosine of the angle between them) reflects semantic relatedness. Embedding approaches have been widely used in schema and ontology alignment: one can embed SNOMED concept descriptions and query terms to find best matches [24]. In practice, embeddings from clinical BERT [25] or word2vec [26] applied to biomedical corpora can reveal semantic similarity even when words differ lexically. For example, “myocardial infarction” and “heart attack” yield close cosine distance in a medical embedding space, aiding normalization.

2.4.2 Set-Based Measures

Another class of metrics operate on sets. The Jaccard similarity measures overlap between two sets of tokens, defined as size(intersection) / size(union), ranging from 0 (no overlap) to 1 (identical sets). In prior work, such as the use of inter-terminology mappings for value set maintenance, Jaccard similarity was applied to quantify the degree of code overlap between mapped and original terminologies - highlighting where mappings preserved semantic intent and where they diverged, especially across vocabularies with varying granularity [17]. For instance, when mapping SNOMED CT codes to ICD-9-CM and ICD-10-CM, the study found substantial overlap, indicating that the mappings effectively captured relevant codes from the target terminologies [17]. However, mappings between SNOMED CT and ICD-10-PCS exhibited lower Jaccard Similarity scores, highlighting discrepancies due to differences in granularity between the terminologies [17]. By utilizing Jaccard Similarity, the researchers could objectively evaluate the effectiveness of inter-terminology mappings, identify areas where mappings were strong, and pinpoint where improvements were needed to enhance semantic interoperability in healthcare data systems. Other set-based measures include overlap coefficients or Dice similarity; string-based measures like Levenshtein distance also exist for name matching, though they ignore meaning beyond text.

By combining embedding and set-based metrics, one can triangulate semantic similarity. For instance, a candidate mapping between two schema attributes might be confirmed if their term embeddings are close (high cosine), their label token sets have high Jaccard overlap. In our work, we leverage some of these metrics to identify semantically similar candidate SNOMED CT terms, and subsequently leverage large language models to select the most appropriate term from these candidates.

2.5 Large Language Models for Ontology Mapping

Recent advances in natural language processing have seen Large Language Models (LLMs) applied to tasks in biomedical informatics. Their ability to interpret and generate text suggests they might aid ontology and schema matching. We consider two categories of tools: traditional concept extraction systems (MedCAT) and emerging LLM-based matching.

2.5.1 MedCAT and Traditional Concept Extraction

MedCAT (Medical Concept Annotation Tool) is a state-of-the-art unsupervised system for named entity recognition and linking in clinical text [27]. It builds embeddings for biomedical entities and uses context to disambiguate mentions against SNOMED/UMLS concepts. In its original evaluation, MedCAT outperformed previous tools on tasks like disease entity detection and general concept linking [27]. Although MedCAT was designed for free-text (clinical notes), its underlying approach – matching terms to SNOMED CT concepts – makes it a useful baseline for ontology mapping. We adapt MedCAT (or similar rule/dictionary-based methods) to structured data by feeding column labels and value terms as input, to see how well it can suggest SNOMED mappings. Traditional methods like MedCAT are reliable when vocabulary matches closely, but they may struggle with ambiguous or uncommon phrasing.

Besides MedCAT, other older tools exist (MetaMap [28], cTAKES [29]) that map text to UMLS concepts. They rely on language lexicons and statistical scoring. These have been extensively used for clinical NLP tasks. However, they often require supervised training or tuning and may not leverage large-scale world knowledge. In contrast, LLMs offer a new paradigm: zero-shot or few-shot mapping by leveraging broad language understanding.

2.5.2 Emerging LLM-Based Schema Matching

Several recent studies have begun using LLMs for ontology and schema matching. Parciak et al. (2024) evaluated an off-the-shelf general LLM on schema matching in a biomedical benchmark [30]. They used GPT-style models with carefully constructed prompts and found that LLMs can identify true semantic correspondences from schema element names and descriptions [30]. In their experiments, newer model versions improved performance and "have potential in bootstrapping the schema matching process" [30]. In an ontology matching context, Taboada et al. (2024) proposed MILA, a system that uses an LLM in a retrieve-and-prompt pipeline with a prioritized search heuristic [31]. MILA achieved top performance on the 2023–2024 Ontology Alignment Evaluation Initiative (OAEI) biomedical matching tasks, outperforming existing systems [31]. Likewise, Babaei et al. (2024) introduced the LLMs4OM framework, showing through evaluations on 20 ontology matching datasets that zero-shot LLMs can "match and even surpass the performance of traditional OM systems" [32]. Collectively, these works demonstrate that LLMs, when properly applied, can handle semantic alignment tasks competitively, often without task-specific training.

These LLM-based methods typically involve two steps: retrieving relevant candidate concepts (e.g. using lexical search or embeddings) and then prompting the LLM to choose or rank the best matches. Our approach similarly uses LLMs to interpret schema elements and map them to SNOMED CT. For example, given a column name from a dataset, an LLM can be asked to return the most appropriate SNOMED concept. We then compare its output against MedCAT's suggestions. The emerging consensus is that while LLMs may hallucinate or require careful prompting, they bring strong general language understanding to the table. By applying them to ontology mapping, we build on the idea that semantic knowledge encoded in LLMs can be harnessed to improve interoperability of biomedical data.

CHAPTER 3

DATASETS AND PREPROCESSING

This chapter describes the systematic collection, extraction, and preprocessing pipeline implemented to curate a well organised corpus of diverse clinical tabular datasets from public repositories. Specifically, datasets were retrieved from Figshare [7] and Zenodo [8], two widely used open-access platforms for hosting scientific datasets. The preprocessing pipeline was designed to handle a wide range of file formats, ensure structural consistency, and improve the usability of the data for downstream tasks such as SNOMED CT term mapping.

3.1 Data Acquisition

The initial step in this study involved identifying publicly available datasets relevant to COVID-19, breast cancer and diabetes research. To facilitate the systematic acquisition and local organization of publicly available datasets, a Python script was developed for automated querying and downloading of disease-specific records from the Figshare & Zenodo repositories. The script leverages the Figshare & Zenodo API to identify relevant datasets, download associated metadata and files, and structure them into a coherent directory format for downstream processing and analysis.

The requests library was employed to send HTTP GET requests to Figshare & Zenodo's search API endpoint. To accommodate large result sets, the API queries were executed in a paginated manner, with configurable parameters for the disease-specific search keyword (e.g., “COVID-19”, “breast cancer” and “diabetes”), page number, and number of records per page. The script iteratively traversed the pages of results, processing each record sequentially.

For each identified dataset, a redundancy check was performed by verifying the existence of a local directory named after the record’s unique ID. This ensured that previously downloaded datasets were not re-downloaded, thereby optimizing both storage and network bandwidth. For new records, the script created a dedicated folder containing:

- A ‘metadata.json’ file storing the complete metadata of the dataset in JSON format,
- A ‘description.txt’ file capturing the textual summary or description of the dataset, and
- A ‘Data’ subdirectory housing all downloadable files associated with the record.

Files were retrieved using direct download links extracted from the API response, and directory structures were automatically generated to mirror the organization of the source dataset. This automated pipeline enabled efficient, reproducible construction of a local dataset repository spanning COVID-19, breast cancer and diabetes research, supporting subsequent stages of data analysis and integration.

Due to the heterogeneity in data organization and formatting, a robust preprocessing pipeline was essential to isolate and retain only structured, machine-readable data suitable for computational analysis.

3.2 Data Preprocessing and Quality Assessment

To ensure reliable downstream analysis, the acquired datasets underwent systematic preprocessing and quality checks. The preprocessing pipeline was designed to streamline heterogeneous inputs into a consistent, analysis-ready form while preserving data integrity. Figure 3.1 illustrates the key preprocessing steps involved in transforming raw datasets from open repositories into structured, machine-readable formats suitable for downstream analysis. The following steps outline the core components of this workflow.

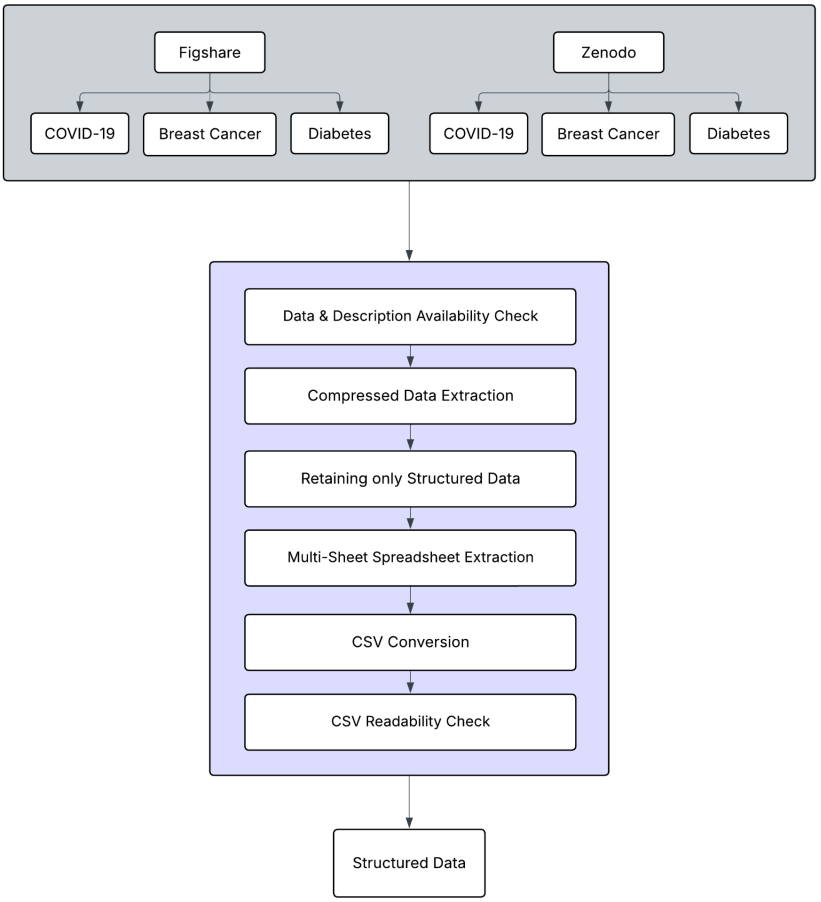


Figure 3.1 Key preprocessing steps involved in obtaining analysis ready structured datasets

3.2.1 Data and Description Availability Check

Each dataset folder was inspected for the presence of:

- Raw data files (typically located in ‘Data’ subdirectory)
- Accompanying description file

Only datasets containing both elements were retained for further processing.

3.2.2 Compressed Data Extraction

Datasets found in compressed formats such as .zip or .tar.gz were extracted recursively using automated scripts to expose nested structured files.

3.2.3 Structured Data Format Filtering

Only files with extensions belonging to a predefined set of allowed structured formats were retained. Allowed formats include: 'csv', 'tsv', 'sav', 'tab', 'mtx', 'sas7bat', 'dta', 'xls', 'xltx', 'xlsx', 'xlsb', 'xlsm', 'ods', 'sqlite', 'db', 'arff', 'json', 'parquet', 'orc', 'avro', 'feather', 'h5', and 'hdf5'. File extensions were normalized using an automated script to ensure uniform lowercase representation. This helped eliminate redundancy and ensured compatibility with format-specific parsers during conversion.

3.2.4 Multi-Sheet Spreadsheet Handling and CSV Conversion

Given the diversity in formats, a specialized parser was developed to convert all structured formats into a unified CSV format. The approach was tailored to each format using the appropriate packages and strategies:

- Excel (xls, xlsx, etc.): pandas.read_excel() using openpyxl or odf engines. All sheets were extracted and saved as individual CSV files.
- Parquet, ORC, Feather: Read using pyarrow and converted to pandas DataFrames.
- SQLite/DB files: Each table was extracted via SQL queries and saved as separate CSVs.
- SPSS/SAS/Stata (sav, sas7bdat, dta): Converted using pandas.read_spss(), read_sas(), or read_stata().
- ARFF: Parsed using scipy.io.arff and converted to DataFrames.
- MTX and TSV: Read with custom delimiters and saved as CSV.

Files were processed in parallel using ThreadPoolExecutor to ensure scalability and improve runtime efficiency.

3.2.5 CSV Readability Check

To ensure data integrity post-conversion, a multi-encoding readability check was

performed using the following strategy:

- Attempted reads using encodings: 'utf-8', 'ISO-8859-1', and 'latin1'.
- Files were flagged as readable if any encoding successfully loaded the file without parser errors.
- Files failing all attempts were logged with error diagnostics.

This check confirmed that a majority of the converted CSVs were structurally intact and ready for downstream processing.

3.2.6 Description Preprocessing

Before making use of data descriptions in downstream analysis, dataset descriptions were standardized by converting text to lowercase and replacing special characters such as '_' (underscores) with a space to ensure consistency in representation.

3.2.7 Field Name Preprocessing

To enable accurate SNOMED CT mapping and harmonization, field (column) names were thoroughly cleaned using the following rule set:

- Removed columns with names like 'Unnamed', 'Column1', or those that were purely numeric and <3 characters long.
- Skipped the first row if >50% of columns were unnamed, assuming potential header shift.
- Transformed column names to lowercase and removed extra spaces.
- Replaced ‘_’ and ‘-’ with spaces; removed special characters.
- Removed parenthetical content if it contained only "yes" or "no".
- Collapsed multiple whitespaces into single spaces and stripped surrounding whitespace.

These cleaning steps ensured uniformity across datasets during term-matching against controlled vocabularies.

3.3 Summary of Preprocessing Outcomes

After preprocessing, the number of datasets and files retained were significantly reduced but of much higher quality and usability. This significant reduction reflects the strict criteria for structured format retention and successful parsing. Table 3.1 summarizes the dataset and file counts before and after preprocessing.

Table 3.1 Repository statistics before and after preprocessing

Repository	**Dataset**	**Dataset IDs (Before)**	**Files (Before)**	**Dataset IDs (After)**	**Files (After)**
Figshare	COVID-19	8,833	55,712	1,291	3,910
Zenodo	COVID-19	5,455	2,52,606	807	32,288
Figshare	Breast Cancer	9,215	9,772	299	1,191
Zenodo	Breast Cancer	5,447	75,18,934	112	5,764
Figshare	Diabetes	9,132	19,074	724	2,245
Zenodo	Diabetes	10,348	4,92,190	414	16,448

CHAPTER 4

QUANTIFYING FAIR-NESS VIA ONTOLOGY MAPPING

To assess the interoperability and reusability of public clinical datasets through a FAIRness lens, we leveraged SNOMED CT [5], a comprehensive clinical terminology system, to map dataset fields to standardized medical concepts. This chapter presents the ontology mapping framework, evaluates the quality of semantic alignment across datasets, and examines how large language models (LLMs) can enhance conventional mapping techniques.

4.1 SNOMED CT Mapping of Structured Data

To promote semantic interoperability, we systematically mapped fields from preprocessed structured datasets to SNOMED CT concepts using string-matching techniques. This baseline mapping approach categorized each field into one of three distinct outcomes:

1. **Exact Match:** A direct and unambiguous correspondence between the dataset field and a SNOMED CT concept.
2. **Multiple Matches:** More than one SNOMED CT term matched the dataset field, indicating ambiguity or broader conceptual coverage requiring disambiguation.
3. **No Match:** No SNOMED CT equivalent was identified, indicating gaps in terminological coverage or presence of highly specific concepts.

Table 4.1 provides a summary of these mapping outcomes across all datasets analyzed. Figure 4.1 presents a flowchart of the ontology mapping and disambiguation (see Section 4.2.2) pipeline.

Table 4.1 Summarizes the distribution of exact-match, multiple-matches, and no-match categories across the analyzed datasets

Data Repository	**Dataset**	**No. of Datasets**	**No. of Fields**	**Exact Match (percentage)**	**Remaining Multiple-matches (percentage)**	**No-match (percentage)**
Figshare	COVID-19	1,291	50,692	5,055 (9.98)	5,657 (11.16)	39,980 (78.86)
Zenodo	COVID-19	807	4,24,236	31,225 (7.36)	41,312 (9.74)	3,51,699 (82.90)
Figshare	Breast Cancer	299	12,562	1,603 (12.76)	3,243 (25.82)	7,716 (61.42)
Zenodo	Breast Cancer	112	1,85,843	6,016 (3.24)	6,257 (3.37)	1,73,570 (93.39)
Figshare	Diabetes	724	87,570	4,120 (4.70)	4,902 (5.60)	78,548 (89.70)
Zenodo	Diabetes	414	4,76,920	9,842 (2.06)	23,559 (4.94)	4,43,519 (93.00)

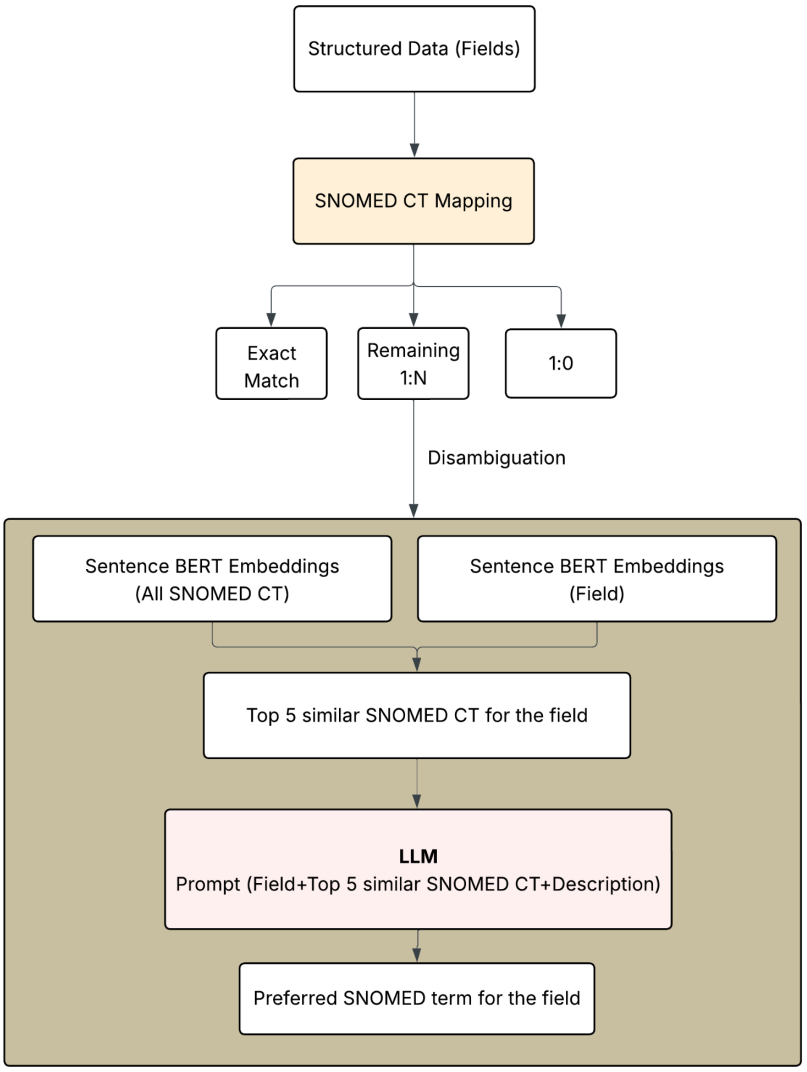


Figure 4.1 Workflow of ontology mapping and LLM-assisted disambiguation for multiple matches

4.2 LLM-Assisted Ontology Mapping

To enhance the quality and coverage of SNOMED CT mappings, we evaluated the performance of large language models (LLMs) in two key areas: expanding matches through prompt-based mapping, and resolving ambiguous multiple-match cases through contextual disambiguation.

4.2.1 Prompt-Based Mapping for Exact Matches

We used exact matches from the baseline string-matching approach to define a ground truth

dataset. These matches were then used to assess recall performance for two LLMs: Mistral-7B [33] and LLaMA 3.1-8B [34]. Both models were provided the following standardized prompt:

"You are tasked with identifying and providing the preferred SNOMED CT term for clinical concepts related to Diabetes. For each provided clinical concept, output the preferred SNOMED CT term within double quotes. If a relevant SNOMED CT term does not exist, return an empty string. Your responses should focus solely on the precision and direct relevance of SNOMED CT terms to the given clinical concept."

Data descriptions were supplied alongside field names to provide contextual grounding. Both the models were run on a NVIDIA L40 GPU optimized for performance with `torch.float16`, and were configured with its standard settings. The baseline model for comparison was MedCAT [27]. Each model's recall was computed, and 95% confidence intervals were estimated using bootstrapped sampling with 10,000 resampling iterations. A detailed comparison of SNOMED CT mapping performance across models and datasets is provided in Table 4.2.

Table 4.2 Performance comparison of MedCAT, Mistral, and Llama3 in SNOMED CT mapping across COVID-19, breast cancer and diabetes datasets for Exact Matches

Model	**COVID-19**		**Breast Cancer**		**Diabetes**	
	Mean Recall	**95% Confidence Interval**	**Mean Recall**	**95% Confidence Interval**	**Mean Recall**	**95% Confidence Interval**
MedCAT	0.33	(0.31, 0.35)	0.39	(0.34, 0.44)	0.27	(0.25, 0.29)
Mistral (Prompt with Data Description)	0.53	(0.51, 0.55)	0.48	(0.43, 0.53)	0.61	(0.59, 0.64)
Llama3 (Prompt with Data Description)	0.53	(0.51, 0.55)	0.47	(0.43, 0.52)	0.52	(0.50, 0.55)

4.2.2 Disambiguation of Multiple Matches

For fields associated with multiple SNOMED CT candidates, we introduced a disambiguation pipeline that combined semantic similarity via embeddings with contextual ranking by LLMs:

- **Embedding Generation:** SNOMED CT concepts and dataset fields were encoded using a pre-trained biomedical Sentence-BERT model [35].
- **Similarity Scoring:** Cosine similarity was calculated between each ambiguous field

and the SNOMED CT vocabulary. The top five most similar SNOMED CT terms were retained as candidates for each field.

- **Contextual Prompting:** The Mistral-7B model [33] was used to select the best candidate term via the following prompt:

"As a clinical terminology expert, you must select the BEST matching term for this dataset column. You MUST select one of these terms. Choose based on:
1. Exact or closest semantic match
2. Clinical relevance
3. Term specificity
Return ONLY the selected term with no additional text or explanation."

Data descriptions and candidate terms were supplied alongside field names to provide contextual grounding. This process enabled informed resolution of ambiguous cases by incorporating both semantic proximity and dataset-level context, as summarized in Table 4.3.

Table 4.3 Summarizes the distribution of matches post-disambiguation across the analyzed datasets

Data Repository	**Dataset**	**Remaining Multiple-matches (percentage)**	**Matched Post-disambiguation (percentage)**	**Total Mapping (percentage)**	**Remaining Multiple-matches (percentage)**	**No-match (percentage)**
Figshare	COVID-19	5,657 (11.16)	4,004 (7.89)	9,059 (17.87)	1,653 (3.27)	39,980 (78.86)
Zenodo	COVID-19	41,312 (9.74)	37,954 (8.95)	69,179 (16.31)	3,358 (0.79)	3,51,699 (82.90)
Figshare	Breast Cancer	3,243 (25.82)	1,256 (9.99)	2,859 (22.75)	1,987 (15.83)	7,716 (61.42)
Zenodo	Breast Cancer	6,257 (3.37)	5,355 (2.88)	11,371 (6.12)	902 (0.49)	1,73,570 (93.39)
Figshare	Diabetes	4,902 (5.60)	3,175 (3.63)	7,295 (8.33)	1,727 (1.97)	78,548 (89.70)
Zenodo	Diabetes	23,559 (4.94)	16,490 (3.46)	26,332 (5.52)	7,069 (1.48)	4,43,519 (93.00)

CHAPTER 5

DISCUSSION

This study demonstrates that augmenting traditional string-based matching with LLM-based techniques substantially improves concept mapping in clinical datasets lacking interoperable annotation. By prompting open-source LLMs (Mistral, LLaMA) with structured field labels and contexts, we achieved much higher coverage of SNOMED CT mappings than exact-string matches alone, and the LLMs could resolve ambiguous mappings by leveraging semantic context. For example, generative LLM methods have been shown to improve normalization accuracy over naive string matching [36]; our results align with this, as the LLMs correctly suggested synonyms and context-driven terms that exact matching missed. However, the LLM outputs sometimes included irrelevant or overly general concepts, underscoring the need to validate model suggestions. In comparison with MedCAT (a supervised concept annotation tool), our LLM-assisted pipeline often covered a broader range of terms but occasionally traded off precision, reflecting known challenges of generative approaches. Importantly, our carefully designed prompts and retrieval steps helped steer the LLMs: consistent with recent work on in-context prompting, we found that framing the mapping task with concise, list-style prompts improved LLM performance, whereas more verbose "definition-style" prompts did not offer additional benefit. This echoes findings by Flaharty et al., who also observed that LLM accuracy on biomedical queries can be improved by appropriate prompt structure [37].

Our findings fit into a growing literature on LLMs in biomedical knowledge tasks. For instance, Zhou et al. (2023) emphasize that large language models often rely on prior parametric knowledge rather than on-the-fly context, and that carefully designed prompts (e.g. counterfactual or narrator-based prompts) can improve "contextual faithfulness" [38]. In our mapping task, we took steps to ground the LLMs with the exact field text and candidate synonyms to avoid hallucinations. Similarly, Huang et al. (2024) showed that retrieval-augmented generation (RAG) can dramatically boost normalization by first filtering to relevant SNOMED candidates [36], a strategy we implicitly adopted by constraining outputs to SNOMED concepts. However, Hager et al. (2024) raise a cautionary note: even state-of-the-art LLMs "do not accurately diagnose" in realistic clinical scenarios and fail to follow instructions reliably [39]. In line with their analysis, we observed that our LLMs could be sensitive to the phrasing and ordering of information, and sometimes produced different mappings when prompts were re-worded (consistent with known LLM sensitivity [39]). These limitations suggest that while LLMs can enhance mapping recall, human oversight remains essential.

Our work also expands the scope of LLM-assisted methods to less-curated, FAIR clinical datasets. Unlike many ontology-mapping studies that use rich, clinician-generated text, we applied LLMs to open, heterogeneous datasets on diabetes, cancer, and COVID-19 from repositories like Figshare and Zenodo. Such data are often not "AI-ready" – they may lack standardized vocabularies, contain inconsistent naming, and miss metadata. By mapping field labels to SNOMED CT, our approach effectively imposed an ontology-driven standardization layer on messy FAIR data, making it semantically interoperable. This addresses a critical barrier to AI adoption posed by siloed and non-standardized health data. In other words, we

show that LLMs can serve as “AI-ready data” transformers, converting disparate inputs into a unified SNOMED reference frame.

This LLM-assisted mapping has important implications for ontology-driven standardization and semantic interoperability. By anchoring dataset variables to SNOMED CT, downstream analyses across datasets and institutions become more meaningful. This echoes recent discussions on LLMs for data harmonization: for example, one review notes that LLMs can “standardize disparate EHRs, align ontologies, and mitigate discrepancies in medical coding” [40]. Our results substantiate this potential by showing practical gains on real data. Once fields are linked to SNOMED, they can be merged or compared across studies, queried by ontology-based tools, or automatically aligned with other standardized vocabularies (e.g. future mapping to LOINC or UMLS). In essence, LLM-driven mapping enables FAIR datasets to interoperate at the semantic level, which is a prerequisite for many AI-driven healthcare applications. As semantic interoperability improves, it will support more reliable analytics, cohort discovery, and decision support, since systems will “interpret, understand, and utilize” data with shared meaning rather than just raw values.

However, it is also clear that integrating LLMs into this process must be done cautiously. The broad literature on SNOMED-LLM integration reports mixed outcomes: while many studies find that injecting concept descriptions into LLM inputs boosts concept normalization, others note inconsistent improvements and occasional performance declines [41]. Likewise, our experience indicates that model answers should be cross-checked (for instance, flagged for manual review when confidence is low). Overall, our work suggests a hybrid approach: using LLMs to expand coverage and suggest mappings, while relying on curated tools like MedCAT or rule-based checks for precision. Such a synergy could leverage the strengths of both worlds.

5.1 Limitations

5.1.1 Dataset Scope

We evaluated only three types of public datasets (diabetes, breast cancer, COVID-19). Although these span multiple domains, they may not represent all biomedical fields. Our findings might differ for rare diseases, genomics, or image-derived data. Future work should validate the approach on broader data sources.

5.1.2 SNOMED CT Coverage

SNOMED is a comprehensive clinical ontology, but not exhaustive. Some dataset fields may lack exact SNOMED matches. This can force LLMs to pick the closest available term, which could be imprecise. For some concepts, other terminologies (e.g. LOINC for lab tests) might be more appropriate.

5.1.3 Model Sensitivity and Stability

The LLMs showed sensitivity to prompt phrasing and context order. Small changes in wording could lead to different mappings. This instability is a known LLM limitation [39]. Additionally, closed versus open models can differ in knowledge; our use of open models (Mistral, LLaMA) might miss recent or niche terms that a proprietary model (ChatGPT-4 [42]) would know. In short, the LLM assistance introduces variability, so outputs must be interpreted with caution.

CHAPTER 6

CONCLUSION

This thesis presents a unified framework for enhancing semantic interoperability in structured biomedical data through ontology-driven FAIR assessment powered by large language models. By mapping heterogeneous dataset fields to standardized SNOMED CT concepts, the approach quantifies interoperability gaps while significantly improving term coverage and disambiguation compared to string matching and MedCAT baselines. Through empirical validation on diverse, real-world clinical datasets, it demonstrates that LLMs - when guided by contextual prompts - can effectively harmonize loosely structured public health data, enabling scalable, automated curation and transforming “not-so-FAIR” datasets into more interoperable research assets.

The practical utility of this approach is clear: any research group or healthcare organization with accessible structured data could use our pipeline (with appropriate prompt templates) to accelerate ontology-driven standardization. Because we used open-source models and public data, the method is readily generalizable: the same techniques should work for other conditions and domains, not just the examples studied. Moreover, one can integrate an LLM-based suggestion engine into EHR data entry systems or metadata repositories which could help non-expert users tag variables with standard terms, thus improving future data interoperability.

Future work should broaden and deepen this approach. One direction is to apply the method to other ontologies beyond SNOMED CT. For example, LOINC (for lab tests) or UMLS (for cross-domain concepts) could be targeted with similar prompts; though prompt tuning might be needed. We also plan to refine the prompting strategy itself: techniques like contextual disambiguation (inspired by Zhou et al. [38]) or few-shot examples could further reduce errors. Fine-tuning small open models on biomedical concept mapping tasks might enhance consistency. Another avenue is integration into EHR or database systems: embedding an LLM mapping layer in the data ingestion pipeline could automate semantic annotation of incoming records. Additionally, expanding evaluation to multilingual datasets would test the generalizability of language-based mapping across locales.

In summary, this work highlights the transformative potential of LLMs for semantic data integration. By bridging the gap between messy real-world data and formal biomedical ontologies, we pave the way for richer, more interoperable datasets. Ultimately, improving semantic interoperability accelerates research and supports AI-driven healthcare, because analyses become more reliable when they operate on “clean” common vocabularies. In a healthcare landscape increasingly driven by data-sharing and AI, methods like ours – which combine modern LLM capabilities with traditional ontology standards – are essential for unlocking the full value of disparate biomedical data sources.

REFERENCES

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

2. Vogt, L., Strömert, P., Matentzoglu, N. et al. Suggestions for extending the FAIR Principles based on a linguistic perspective on semantic interoperability. Sci Data 12, 688 (2025). https://doi.org/10.1038/s41597-025-05011-x

3. Hughes, L.D., Tsueng, G., DiGiovanna, J. et al. Addressing barriers in FAIR data practices for biomedical data. Sci Data 10, 98 (2023). https://doi.org/10.1038/s41597-023-01969-8

4. Reich, C. et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. J. Am. Med. Inform. Assoc. 31, 1181–1192 (2024). https://doi.org/10.1093/jamia/ocad247

5. Vuokko R, Vakkuri A, Palojoki S. Systematized nomenclature of medicine–clinical terminology (SNOMED CT) clinical use cases in the context of electronic health record systems: systematic literature review. JMIR Med Inform. 2023;11:e43750. doi: 10.2196/43750.

6. Meredith, J., Whitehead, N. & Dacey, M. Aligning semantic interoperability frameworks with the FOXS stack for FAIR health data. Methods Inf. Med. 62, e39–e46 (2023). https://doi.org/10.1055/a-1993-8036

7. Thelwall M, Kousha K. Figshare: a universal repository for academic resource sharing? Online Information Review. 2016;40(3):333-346. doi: 10.1108/OIR-06-2015-0190.

8. Nowak K, Nielsen LH, Ioannidis Pantopikos AT. Zenodo, a free and open platform for preserving and sharing research output. Zenodo. 2016. doi: 10.5281/zenodo.51902.

9. Rao, A. et al. FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries. Lancet Digit. Health 5, e712–e736 (2023). https://doi.org/10.1016/S2589-7500(23)00129-2

10. World Health Organization (WHO). International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). World Health Organization (2019). https://icd.who.int/browse10/2019/en

11. European Commission. European Open Science Cloud (EOSC) Strategic Implementation Plan. Publications Office of the European Union (2019). https://doi.org/10.2777/202370

12. European Commission. Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC). Publications Office of the European Union (2022). https://doi.org/10.2777/935288

13. Dryad. Dryad Digital Repository. Dryad (2025). https://datadryad.org

14. Center for Open Science. Open Science Framework (OSF). Center for Open Science (2025). https://osf.io

15. McDonald, C. J. et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 49, 624–633 (2003). https://doi.org/10.1373/49.4.624

16. Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T. & Moore, R. RxNorm: a normalized naming system for generic and branded drugs. J Am Med Inform Assoc 18, 441–448 (2011). https://doi.org/10.1136/amiajnl-2011-000116

17. Fung KW, Xu J, Gold S. The Use of Inter-terminology Maps for the Creation and Maintenance of Value Sets. AMIA Annu Symp Proc. 2020 Mar 4;2019:438-447. PMID: 32308837; PMCID: PMC7153132.

18. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-30. doi: 10.1136/jamia.2009.000893. PMID: 20190053; PMCID: PMC3000779.

19. El-Sappagh, S., Franda, F., Ali, F. et al. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak 18, 76 (2018). https://doi.org/10.1186/s12911-018-0651-5

20. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109-12. doi: 10.1056/NEJMp1607591. PMID: 27653561; PMCID: PMC6309165.

21. Mons, B. et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37, 49–56 (2017). https://doi.org/10.3233/ISU-170824

22. Margolis, R. et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc 21, 957–958 (2014). https://doi.org/10.1136/amiajnl-2014-002974

23. Bonazzi, V. et al. The NIH Common Fund Data Ecosystem. Cell Genomics 1, 100004 (2021). https://doi.org/10.1016/j.xgen.2021.100004

24. Karadeniz, İ., Özgür, A. Linking entities through an ontology using word embeddings and syntactic re-ranking. BMC Bioinformatics 20, 156 (2019). https://doi.org/10.1186/s12859-019-2678-8

25. Huang, K. et al. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019). https://arxiv.org/abs/1904.05342

26. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013). https://arxiv.org/abs/1301.3781

27. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. Artif Intell Med. 2021;117:102083. doi: 10.1016/j.artmed.2021.102083.

28. Aronson, A. R. & Lang, F. M. An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. 17, 229–236 (2010). https://doi.org/10.1136/jamia.2009.002733

29. Savova, G. K. et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. 17, 507–513 (2010). https://doi.org/10.1136/jamia.2009.001560

30. Parciak M, Vandevoort B, Neven F, Peeters LM, Vansummeren S. Schema Matching with Large Language Models: an Experimental Study. arXiv:2407.11852 [Preprint]. 2024.

31. Taboada M, Martinez D, Arideh M, Mosquera R. Ontology Matching with Large Language Models and Prioritized Depth-First Search. arXiv:2501.11441 [Preprint]. 2024.

32. Babaei H, D'Souza J, Auer S. LLMs4OM: Matching Ontologies with Large Language Models. arXiv:2404.10317 [Preprint]. 2024.

33. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B [Internet]. arXiv [Preprint]. 2023. Available from: https://arxiv.org/abs/[arXiv identifier]

34. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models [Internet]. arXiv preprint. Available from: https://arxiv.org/abs/[arXiv identifier]

35. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3982–3992 (2019). https://doi.org/10.48550/arXiv.1908.10084

36. Berkowitz, J. S., Srinivasan, A., Acitores Cortina, J. M., Fatapour, Y., & Tatonetti, N. P. Biomedical text normalization through generative modeling. J. Biomed. Inform. 145, 104850 (2025). https://doi.org/10.1016/j.jbi.2025.104850

37. Flaharty KA, Hu P, Hanchard SL, et al. Evaluating large language models on medical, lay-language, and self-reported descriptions of genetic conditions. Am J Hum Genet.

2024;111(9):1819-1833. doi: 10.1016/j.ajhg.2024.07.011. Epub 2024 Aug 14. PMID: 39146935; PMCID: PMC11393706.

38. Zhou W, Zhang S, Poon H, Chen M. Context-faithful prompting for large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023; Singapore. Association for Computational Linguistics; 2023. p. 14544-14556.

39. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat. Med. 30, 2613–2622 (2024). https://doi.org/10.1038/s41591-024-03097-1

40. Kokash, N., Wang, L., Gillespie, T.H. & Belloum, A. Ontology- and LLM-based data harmonization for federated learning in healthcare. arXiv preprint arXiv:2505.20020 (2025).

41. Chang, E. & Sung, S. Use of SNOMED CT in large language models: scoping review. JMIR Med. Inform. 12, e62924 (2024). https://doi.org/10.2196/62924

42. OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023).