NLP based Extraction and OMOP Standardization of Breast Cancer Clinical Data from Indian Discharge Summaries

Dr.SwethaKiranmayi Jakkuva (Real-World Data Research Coordinator at GVW Technologies), Khansa Fathima (Assistant Professor at JSSAHER) , M R Sai Dileep (Data Scientist at GVW Technologies), Shreema S Rao (Data Engineer, GVW Technologies), Sanjay R (Healthcare Data Analyst, GVW Technologies), Sai Pattabhiram L (Junior Software Engineer at GVW Technologies)

Background

India's healthcare system is undergoing a digital shift, with an estimated 35–40% of hospitals adopting some form of electronic health records or hospital information systems. In Tier 1 cities, up to 80% of private hospitals have digitized operations, yet much of this data remains **unstructured, siloed, and not analytics ready**. Public hospitals, especially at district and sub-district levels, continue to rely heavily on paper-based records, with fewer than 25% showing significant digital adoption.

This fragmented landscape presents a major barrier to real-world evidence (RWE) generation, longitudinal tracking, and scalable research. National programs like the Ayushman Bharat Digital Mission (ABDM) have laid the groundwork for a unified digital health ecosystem, but full integration and interoperability are still evolving. Recognizing this gap the Breast Cancer Study was initiated to demonstrate how unstructured clinical data starting with discharge summaries can be transformed into standardized, interoperable formats using NLP and OMOP CDM. This effort serves as a foundational step toward building scalable, research-ready data pipelines across India's diverse healthcare settings.

Methods

This study is focused on extracting clinically relevant breast cancer data from PDF discharge summaries (2014–2022) sourced from **JSS AHER, Mysuru**. The pipeline involves multiple key phases:

Deidentification - We used Redact, a healthcare-customized rule-based engine, to irreversibly de-identify sensitive information from PDF discharge summaries. By removing all identifiable elements without retaining mappings, this process ensures full compliance with privacy standards such as HIPAA and GDPR, enabling secure downstream data extraction and analysis.

NLP extraction is central to transforming unstructured breast cancer discharge summaries into structured research-ready data. By leveraging BERT-based Named Entity Recognition and relationship extraction, the pipeline identifies key clinical concepts diagnoses, procedures, lab values, and medications along with their contextual modifiers. Unlike rule-based methods, NLP handles inconsistent formats, enables semantic understanding, and automates concept normalization to vocabularies like UMLS and SNOMED CT. This structured output is mapped to OMOP CDM, laying the foundation for interoperable analytics and scalable real-world evidence research in oncology.

Annotation Strategy

High-quality annotation was foundational to model development. Using tools such as MedTator, five primary entity types and 22 modifiers were manually annotated across discharge summaries. Annotation guidelines were developed to ensure semantic consistency and reproducibility. Inter-Annotator Agreement (IAA) scores exceeded 93% across batches, with Cohen's Kappa and Krippendorff's alpha used for adjudication. This gold-standard corpus enabled robust training and evaluation of downstream NLP models.

Named Entity Recognition (NER)

Clinical entity recognition was performed using a Bidirectional Encoder Representations from Transformers (BERT) model, fine-tuned on domain-specific annotated corpora. The bidirectional transformer architecture enables contextual encoding of tokens by simultaneously incorporating both preceding and succeeding lexical elements, which is essential for resolving semantic ambiguity in clinical narratives (e.g., distinguishing “cold” as a symptom from ambient temperature). The model was trained to identify five primary entity categories: problems, procedures, laboratory tests, medications, and demographic attributes.

Entities were annotated using the BIO tagging scheme (Beginning, Inside, Outside), allowing for accurate delineation of multi-token spans and facilitating structured representation of clinical concepts (e.g., “chest pain” → B-PROBLEM, I-PROBLEM). Sub-token classification further improved recognition of abbreviated and fragmented terminology commonly observed in discharge summaries. This configuration supports high-resolution entity extraction and serves as a foundational step for downstream relationship modeling and ontology-based normalization.

Relationship Extraction

Following entity recognition, relationship extraction was performed using the same fine-tuned Bidirectional Encoder Representations from Transformers (BERT) architecture. Two modeling strategies were employed: sentence-pair classification and span-based encoding. In the sentence-pair approach, BERT receives a clinical sentence along with a candidate entity pair and predicts the semantic relation type between them (e.g., “drug treats condition,” “test confirms diagnosis”). This method enables contextual disambiguation of relationships within complex clinical narratives.

Alternatively, span-based models encode entity spans using self-attention and pooling mechanisms to capture inter-token dependencies and infer relational semantics. These approaches are particularly effective in discharge summaries, where relationships are often implicit and distributed across fragmented text. Relationship extraction facilitates the construction of structured knowledge graphs from unstructured clinical documentation, enabling downstream applications such as decision support, temporal reasoning, and case-based retrieval. The extracted relations are subsequently aligned with OMOP CDM constructs to ensure semantic interoperability and support real-world evidence generation

Concept Normalization

Concept normalization was performed to map extracted clinical mentions to standardized terminologies, including the Unified Medical Language System (UMLS), SNOMED CT, and RxNorm. This step addresses lexical variability and semantic ambiguity inherent in clinical text, such as synonymous expressions (e.g., “heart attack” vs. “myocardial infarction”). A BERT-based semantic similarity framework was employed to encode both mention and candidate ontology terms into a shared vector space, enabling retrieval of the most contextually appropriate match using cosine similarity or cross-encoder scoring techniques.

Domain-specific embeddings, such as BioBERT and ClinicalBERT, were utilized to enhance semantic representation and improve alignment with medical language. The normalization process ensures interoperability with electronic health record (EHR) systems and supports downstream applications including cohort identification, drug utilization studies, and integration with the OMOP Common Data Model (CDM). Manual validation and random accuracy checks were incorporated to maintain mapping fidelity and analytic reliability.

AI-Supported Vocabulary Standardization

An AI-powered mapping utility was integrated to support terminology normalization. The tool enhances non-uniform drug and clinical terms using Gemini-based query expansion and semantic matching, facilitating alignment with UMLS vocabularies (e.g., RxNorm, SNOMED CT). Batch input support and ontology-linked outputs enabled scalable preprocessing for OMOP CDM conversion.

OMOP CDM Mapping

Normalized entities were mapped to OMOP CDM tables using a modular ETL framework. The registry fields were designed to align with **mandatory OMOP domains**, ensuring structural completeness. Mapping logic was implemented using PostgreSQL and R-based tools, with support for relational databases and REST APIs. Vocabulary mapping was validated using Athena and Usagi, with extensions for Indian terminologies. The structured output supports scalable analytics and real-world evidence generation.

Results

Preliminary model evaluations indicate promising performance across the NLP pipeline. Named Entity Recognition (NER) and relationship extraction components achieved an average accuracy of approximately 80%, based on internal validation against manually annotated gold-standard data. Concept normalization demonstrated consistent alignment with standardized vocabularies, with early mapping success rates exceeding expected benchmarks for Indian clinical terminology.

While final results and registry deployment are pending, full-scale evaluation and publication of outcomes are scheduled for release within the next two months. Ongoing efforts include expanded validation, integration of additional clinical data sources, and further automation of OMOP CDM conversion processes to support scalable real-world evidence generation.

Conclusion

This study presents a reproducible NLP-based pipeline for extracting, annotating, and standardizing clinical data from unstructured breast cancer discharge summaries in India. By leveraging transformer-based models for Named Entity Recognition and relationship extraction, and aligning outputs with OMOP CDM through concept normalization, the system enables scalable, interoperable data generation suitable for real-world evidence research. Preliminary results demonstrate strong model performance, and full registry deployment is underway. The framework offers a foundational step toward structured oncology data infrastructure in Indian healthcare, with potential for adaptation across therapeutic domains and institutional settings.

References

1. AMIA Annu Symp Proc. 2018 Dec 5;2018:770–779.
2. J Biomed Inform. 2021 Apr 28;118:103790. doi: 10.1016/j.jbi.2021.103790
3. Prostate. Author manuscript; available in PMC: 2023 Aug 1.
   Published in final edited form as: Prostate. 2022 May 10;82(11):1107–1116. doi: 10.1002/pros.24363
4. AMIA Symposium 2021:394-403