EXTENTED RESEARCH PROJECT

SCHOOL OF SOCIAL SCIENCES

Computational phenotyping patients using multimodal EHRs

Student ID: 11356488

Supervisor: Dr. Xian Yang

2024

An extended research project report submitted to the University of Manchester for the degree of Master of Science in Data Science with Business and Management in the Faculty of Humanities.

TABLE OF CONTENTS

Word Count : 6861

LIST OF ABBREVIATIONS

Abbreviation	Definition
EHR	Electronic Health Records
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
UMLS	Unified Medical Language System
MeSH	Medical Subject Headings
NER	Named Entity Recognition
VAE	Variational Autoencoder
BERT	Bidirectional Encoder Representations from Transformers
MIMIC-III	Medical Information Mart for Intensive Care III
ICD	International Classification of Diseases
TF-IDF	Term Frequency-Inverse Document Frequency
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
UMAP	Uniform Manifold Approximation and Projection

LIST OF FIGURES

LIST OF TABLES

Abstract

The rise in digitization of Electronic Health Records (EHRs) has created a unique opportunity to extract clinically relevant insights from multimodal data sources. This study proposes a computational phenotyping framework utilizing the MIMIC-III dataset to merge structured data disease codes(ICD-9) and unstructured clinical notes for meaningful disease topic and phenotypes identification. Our methodology integrates the traditional machine learning techniques, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) with K-means clustering and advanced neural topic modelling techniques like BERTopic with highly contextual ClinicalBERT embeddings to identify distinct disease topics and cluster phenotypes based on topic distributions of patients. The findings demonstrate that the multimodal data integration significantly improves phenotyping accuracy and comprehensiveness. Overall, NMF with K-means effectively identifies broader categories of disease phenotypes, whereas BERTopic leveraging fine-tuned ClinicalBERT embeddings, identifies subtle context-derived patterns within clinical data. This holistic approach facilitates the identification of intricate disease phenotypes, aiding the healthcare providers by tailoring treatments plans, early detection, and prevention of diseases and development of precision medicine. This study underscores the transformative potential of merging contemporary modelling techniques with varied data sources to enhance computational phenotyping. By offering a refined perspective on patient health, the framework fosters the creation of more targeted healthcare interventions. Future investigations could focus on incorporating additional data sources like imaging and extending the neural-based models across different healthcare environments to enhance their efficacy and relevance in clinical practice.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Xian Yang for her support and very useful constructive feedback after every presentation and result discussions which helped me refine my ideas and improve the quality of my research and code.

A special note of thanks goes to my friends Ayush and Tanya, whose discussions and collaborations enriched my understanding and contributed to the progress of this research. I am particularly grateful for how willing they were to share their knowledge and insights.

A special word of recognition to my friend and classmate, Abhishek , who was a constant source of motivation and support throughout my ups and downs of this challenging year, and for all his help during this research.

I am very thankful to my parents, without whom I could not have taken on this challenge.

Without the assistance of all these people and university, this work would not have been feasible. Although the above acknowledgments only list a handful by name, I would like to express my gratitude to everyone who contributed to this project.

Declaration

No portion of the work referred to in this extended research project report has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Intellectual Property Statement

i. The author of this extended research project report (including any appendices and/or schedules to this report) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this report, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks, and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the report, for example graphs and tables ("Reproductions"), which may be described in this report, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this report, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see https://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant dissertation restriction declarations deposited in the University Library, The University Library's regulations (see https://www.library.manchester.ac.uk/about/regulations/) and in The University's Guidance for the Presentation of dissertations.

CHAPTER 1

INTRODUCTION

1.1 Background

The widespread adoption of electronic health records (EHRs) has significantly revolutionized the healthcare sector by offering a detailed, continuous record of the patient care (Abdul-Husn and Kenny, 2019). EHRs encompass diverse modalities such as structured patient demographics, diagnosis codes and lab results, and unstructured discharge summaries, medical history and radiology images. This extensive repository of data has paved the way for advancements in personalized medicine, enabling optimized treatment plans and disease risk prediction for individual patients according to their specific health profiles (Mohsen et al., 2022). As of 2021, nearly 96% of hospitals in developed countries like United States had integrated EHRs. However, the size and complexity of EHR data pose significant challenges in multimodal data fusion, interpretation, and analysis. A promising strategy to address these challenges (Liu et al., 2021) is computational phenotyping, which entails the identification and categorization of patient subgroups based on common clinical disease traits extracted from their EHRs.

Traditional computational phenotyping techniques rely on rule-based systems, which are time-consuming, prone to human error, and limited in scalability. Recent advances in AI and machine learning (Che & Liu, 2017) (Banda et al., 2018), particularly in neural topic modelling, tensor factorization, and clustering techniques, offer automated, scalable, and more accurate alternatives. These methods have been successfully utilized by EHR data, enabling discovery of latent disease topics and classification of patients into clinically meaningful phenotypic groups (Zeng et al., 2018). However, conventional methods like Latent Dirichlet Allocation (LDA) often face limitations related to predictive capability and scalability. To overcome these challenges, researchers are increasingly adopting deep leaning-based neural topic models, advanced tensor factorization techniques (Kim et al., 2017) for phenotype identification.

1.2 Scope of the Study

The study seeks to investigate and develop computational phenotyping techniques for analysing multimodal EHR data to identify disease phenotypes that can enhance precision medicine. By integrating both structured data like diagnosis codes, and unstructured data, such as clinical notes, the study intends to uncover meaningful disease phenotypes. A key component of this study is a comparative analysis of advanced topic modelling methodologies to analyse EHRs and produce clinically relevant phenotypes. The study will evaluate a range of unsupervised methods, from traditional techniques like LDA, Nonnegative Matrix Factorization (NMF) (Becker et al., 2022) with K-Means clustering, to more advanced neural topic modelling strategies such as transformer-based BERTopic (Grootendorst, 2022).

The comparative analysis will address key questions such as how advanced computational topic modelling methods can improve accuracy and scalability of phenotyping using multimodal EHR and how the integration of structured and unstructured EHR data can enhance disease phenotypes (Zhang et al., 2023). The study's findings aim to make significant contributions to the healthcare sector by refining methods for identification of disease phenotypes, improving patient stratification and disease risk prediction.

1.3 Structure

The report systematically outlines the research methodology from data preparation to conclusions and future research directions. It starts with a Literature Review of past studies on computational phenotyping, including topic modelling, tensor factorization, and neural topic models. The Data Preparation section addresses the collection and preprocessing EHR datasets, highlighting the use of natural language processing (NLP) with medical knowledgebase for unstructured text (Zheng et al., 2018). The Data Fusion section focuses on integrating multimodal data into a unified dataset through feature engineering. The Methodology section details the computational phenotyping pipeline and advanced topic modelling techniques like neural network-based models, tensor factorization or clustering applied on vectorized text like neural embeddings. The Results section identifies phenotypes and evaluates their clinical relevance. The Discussion critically examines the implications for precision medicine and the study’s limitations. The report concludes with recommendations for future work and includes appendices with supplementary materials.

CHAPTER 2

LITERATURE REVIEW

2.1 Computational Phenotyping

Computational Phenotyping involves using data-driven methods to define and identify disease phenotypes (categories) derived from large-scale EHR data, with the ultimate objective of grouping patients based on shared clinical characteristics to enhance disease understanding, predict treatment outcomes, and customize treatments (Richesson et al., 2016).This concept has significantly transformed the domain, as prior methodologies of phenotyping were predominantly dependent on manual rule-based frameworks, which lacked scalability for voluminous datasets and were susceptible to bias.

As elucidated by Ho et al. (2014), the emergence of machine learning and deep learning techniques has facilitated scalable, automated, and impartial methods for the identification of phenotypes, thereby enabling the uncovering of latent patterns with EHR data and the recognition of novel phenotypes. The efficacy of these methods was demonstrated by the findings of Kim et al. (2017) and Henderson et al. (2017). However, Hripcsak et al. (2015) emphasized significant challenges related to the integration of diverse data types and development of robust models that generalize across varied patient cohorts. This research aims to develop a robust phenotyping pipeline that combines established methods with novel strategies to enhance the efficacy and scalability of phenotyping.

2.2 Multimodal Data Fusion

Multimodal data fusion is pivotal for integrating heterogeneous data types, including structured EHR data and unstructured information (e.g., clinical notes). This integration is vital for creating comprehensive patient phenotypes. Liu et al. (2019) noted that traditional fusion methods often rely on simple feature concatenation, which can yield suboptimal results due to varying scales and distributions.

Advancements have led to more sophisticated methods like feature-level and decision-level fusion. Feature-level fusion combines multiple modalities into a single vector for better modal learning, while decision-level fusion merges predictions from different models to enhance robustness against missing or noisy data as explained by Wu et al. (2014).

Recent deep learning innovations have enabled powerful fusion techniques. Ngiam et al. (2011) illustrates that multimodal autoencoders and attention-based models have improved the identification of inter-modal relationships, effectively capturing dependencies between structured and unstructured data that leads to better performance in phenotyping tasks (Gao et al., 2020).

2.3 Medical Ontology Mapping

Medical ontology matching is essential for the integration and standardization of diverse EHR data sources. Structured vocabularies like Unified Medical Language System (UMLS), SNOMED CT or Medical Subject Heading (MeSH) facilitate semantic interoperability within the healthcare domain. Bodenreider (2004) articulated the role of UMLS in integrating biomedical terminologies and standards, thereby interlinking disparate health information systems. Lipscomb (2000) characterized MeSH as a controlled vocabulary employed for the indexing of biomedical literature, thus promoting consistent and precise information retrieval. In the context of computational phenotyping, Zeng et al. (2019) noted that medical ontology alignment is instrumental in correlating unstructured text data with standardized terminologies, thereby improving clinical relevance. Examining these knowledge bases reveals broader applications for phenotyping.

2.4 Topic Modelling

Topic modelling is extensively employed in the field of computational phenotyping to derive latent themes from extensive textual corpora, exemplified by clinical documentation (Bodenreider, 2008). Blei et al. (2003) introduced Latent Dirichlet Allocation (LDA), a seminal approach within this sphere, conceptualizing each document as a composite of various topics and each topic as a probabilistic distribution over vocabulary. While effective in revealing disease-related themes, LDA struggles with scalability and capturing complex features.

2.5 Tensor Factorization for Phenotyping

Tensor Factorization has become a valuable method for phenotyping due to its ability to handle multi-dimensional EHR data. Henderson et al. (2017) delineated methodologies such as Granite, which employs diversified and sparse nonnegative tensor factorization, and demonstrated its effectiveness in extracting phenotypes from high-dimensional EHR data. The tensor factorisation methods have shown promise in generating relevant phenotypes, but there is a need for more interpretable models to handle complexity and heterogeneity of data.

2.6 Clustering Methods for Phenotyping

Clustering methods are extensively used in computational phenotyping to group patients based on clinical data similarities. One of the most effective methods due to simplicity and efficiency is K-Means clustering (Ikotun et al., 2022). While effective, Xu and Tian (2015) note that K-means requires careful feature selection and determining the number of clusters, which can be challenging in high-dimensional EHR data.

More advanced clustering techniques investigated by Joelson et al. (2021), such as hierarchical clustering and density-based clustering, have been applied in phenotyping to capture more complex relationships within data, as noted by Xu and Tian. These methods often require more computational resources and may struggle with scalability, but they can provide more nuanced insights into patient phenotypes. Recent studies underscore the necessity for advanced models to address the complexity and heterogeneity inherent in EHR data when employing clustering methods.

2.7 Neural Topic Modelling

Neural topic modelling has advanced the extraction of latent topics from extensive textual data, particularly within EHRs. In contrast to traditional models like LDA, neural topic models utilize deep learning to capture complex semantic and contextual relationships. Devlin et al. (2018) highlighted the role of pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers), emphasizing their ability to produce high-quality embeddings that capture the context of words within clinical notes. BERTopic, developed by Grootendorst (2020), Top2Vec by Angelov (2020), utilize these embeddings to cluster semantically related test and generate coherent topics, enhancing the interpretability of clinical data.

2.8 Embeddings in Medical NLP

Embeddings play a critical role in conversion of textual information into numerical vectors that encapsulate the semantic interrelations among word meanings, which is vital for the execution of NLP tasks. The conventional Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, as discussed by Salton and Buckley (1988), is widely used for document classification and information retrieval, owing to its straightforwardness and interpretive clarity. However, TF-IDF fails to encapsulate contextual nuances, limiting its applicability in intricate NLP tasks.

To address these limitations, Mikolov et al. (2013) introduces Word2Vec model that generates dense word vectors by analysing their co-occurrence within a designated context window. This approach has proven efficacious in the identification of medical concepts and enhancement of NLP models with the integration of domain-specific embeddings, as highlighted by Choi et al. (2016).

The emergence of transformer-based architectures, notably BERT introduced by Vaswani et al. (2017), has fundamentally transformed the landscape of NLP by producing contextual embeddings that reveal complex relationships among words. Tailored adaptations for specific domains , such as BioBERT for biomedical field and ClinicalBERT for medical field, have been conceived to address the distinctive challenges presented by biomedical or clinical texts.

2.9 Research Gaps and Future Directions

While considerable progress has been made in utilizing multimodal EHR data for computational phenotyping, numerous challenges persist. A primary obstacle pertains to the effective integration of both structured and unstructured data, which constraints the capacity to fully exploit the potential of EHRs. Moreover, the scaling of these methodologies to accommodate large datasets presents considerable difficulties, particularly for deep learning-based approaches that necessitate substantial computational resources, as highlighted by Shickel et al. (2017). Future endeavours should focus on the development of more robust and scalable methodologies for the fusion of multimodal data, alongside enhancements in the interpretability of the resulting phenotypes.

CHAPTER 3

METHODOLOGY

3.1 Conceptual Framework

The primary aim of this research is to develop a robust computational phenotyping framework that leverages multimodal EHRs to identify clinically significant disease phenotypes that can enhance clinical decision-making. By leveraging both structured and unstructured data, this research seeks to uncover significant patterns within patient information that can improve clinical decision-making processes. The pipeline prioritizes data integration, feature engineering, and sophisticated topic modelling techniques to address the complexities of clinical data and enhance the interpretability of phenotypes (Carrell et al., 2024). This section elaborates on each phase, elucidating the rationale for the chosen methods and their alignment with the research.

3.2 Data

Data Overview and Source

The data for this research was sourced from MIMIC-III (Medical Information Mart for Intensive Care III) Clinical Database, which contains de-identified patient information and is openly available for researchers. The information compiled in this database is pertaining to more than forty thousand patients who were admitted to care units at Beth Israel Deaconess Medical Center in Boston from 2001 and 2012 (Johnson et al. 2016).An overview of the MIMIC-III is shown in Figure 3.1.

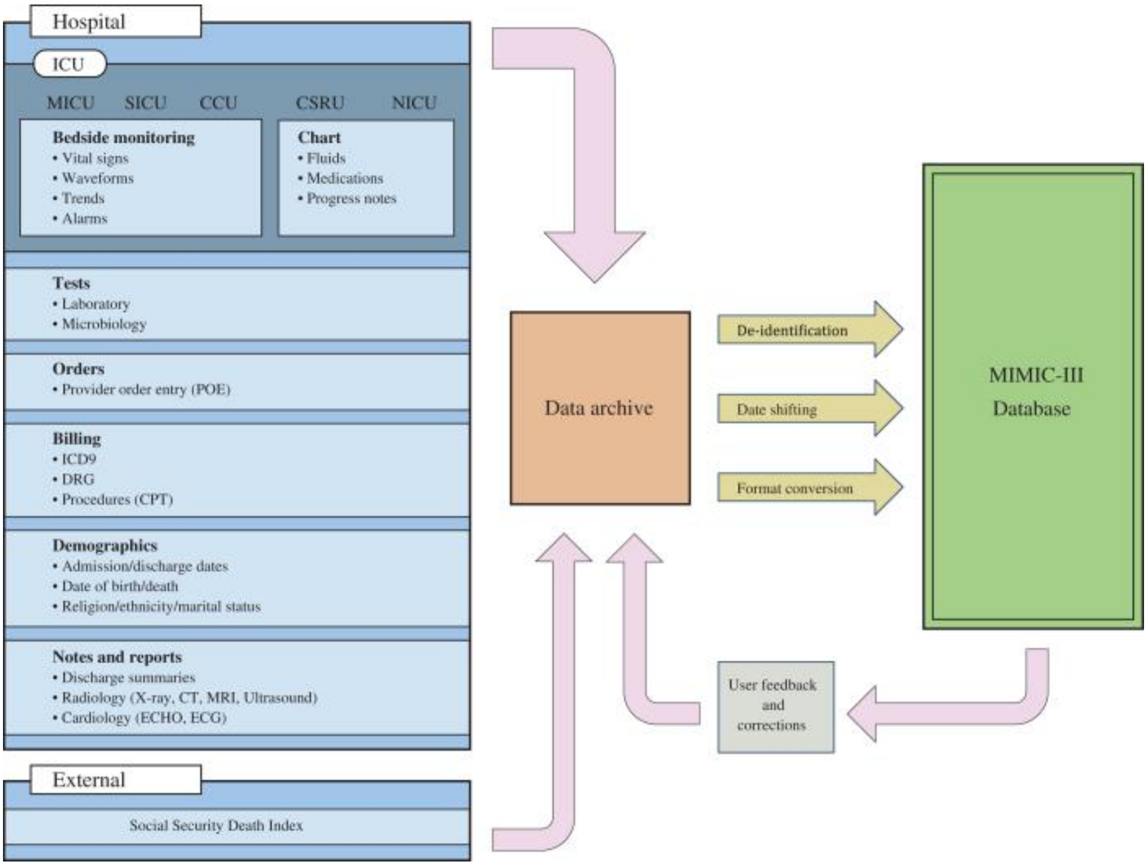


Figure 1 : Overview of the MIMIC-III critical care database (Johnson et al., 2016)

Dataset Composition and Structure

The MIMIC-III dataset constitutes of 26 tables (Johnson et al., 2016), each consisting of distinct categories of clinical data. This dataset captures multimodal information like :

- **Structured Data Tables** : These include demographic information, laboratory test results, medical procedural codes, medication codes and diagnostic codes. The data features are numerical or categorical.
- **Unstructured Data Tables**: Comprising of clinical physician notes, discharge summaries, imaging reports and caregiver annotations.
- **Dictionary Tables** : Dictionary Descriptions of ICD codes.

Data De-Identification and Privacy

To ensure patient privacy, entire MIMIC-III dataset have undergone a rigorous de-identification process adhering to the Health Insurance Portability and Accountability Act (HIPAA). The procedure involved systematic data cleaning to eliminate or anonymise Personal Health Identifiers (PHIs) and date adjustments to obscure temporal data while retaining the chronological order of clinical occurrences (Dernoncourt et al., 2016).

Data Generation Process

The MIMIC-III dataset is derived from critical care systems like Philips CareVue and iMDsoft MetaVision (Johnson et al., 2016), includes diverse data types such as physiological measurements, medication logs, caregiver observations and structured EHR like demographics, ICD codes and mortality statistics.

3.3 Exploratory Data Analysis

This exploratory analysis highlights the key patterns in structured and unstructured modality of EHR data, highlighting patient demographics, diagnosis, symptoms and conditions.

3.5.1 Overview of Datasets

An overview of the MIMIC-III datasets used in this study is highlighted in Table 3.1, detailing their original sources, key columns and data types.

Dataset Name	Original Table	Number of Records	Key Columns	Data Type
Patient Data	PATIENTS	46,520	SUBJECT_ID, GENDER, DOB	Structured (Numerical, Categorical)
Admissions Data	ADMISSIONS	58,976	HADM_ID, ADMITTIME, DISCHTIME	Structured (Datetime, Categorical)

ICU Stay Records	ICUSTAYS	61,532	ICUSTAY_ID, INTIME, OUTTIME	Structured (Datetime)
Medical Codes	DIAGNOSES_ICD	65,000	HADM_ID, ICD9_CODE, SEQ_NUM	Structured (Categorical)
Clinical Notes	NOTEEVENTS	20,83,180	HADM_ID, TEXT, CATEGORY	Unstructured (Text)

1.Table 3.1: Overview of MIMIC-III Datasets

3.5.2 Analysis of Structured Data Modality

Demographic Features Analysis

Correlation Analysis

1. **Gender and Mortality**

   The dataset has a nearly balanced gender distribution, with a slightly higher number of males (Figure 3.3). Hospital mortality is relatively low, Figure 3.4 shows that around 50,000 patients are alive, and 5,000 are deceased. Survival of most patients is important for building robust models across genders.

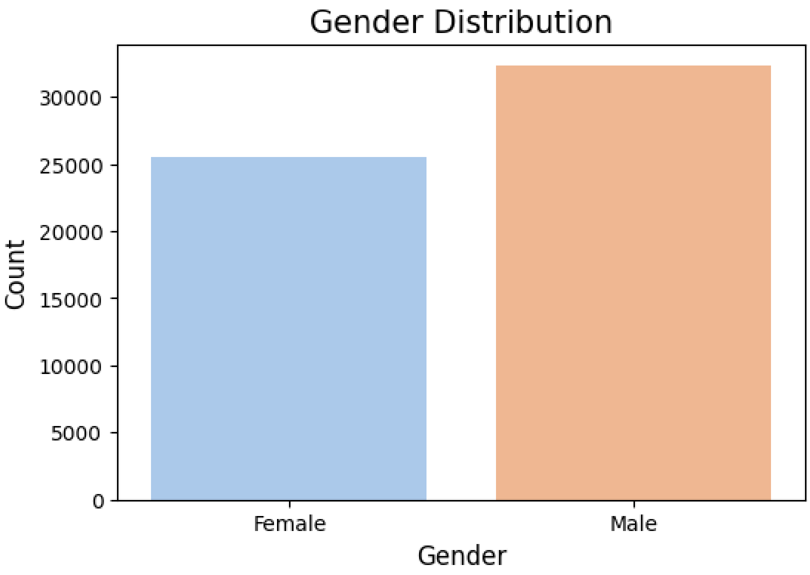


Figure 2:Patient Distribution by Gender

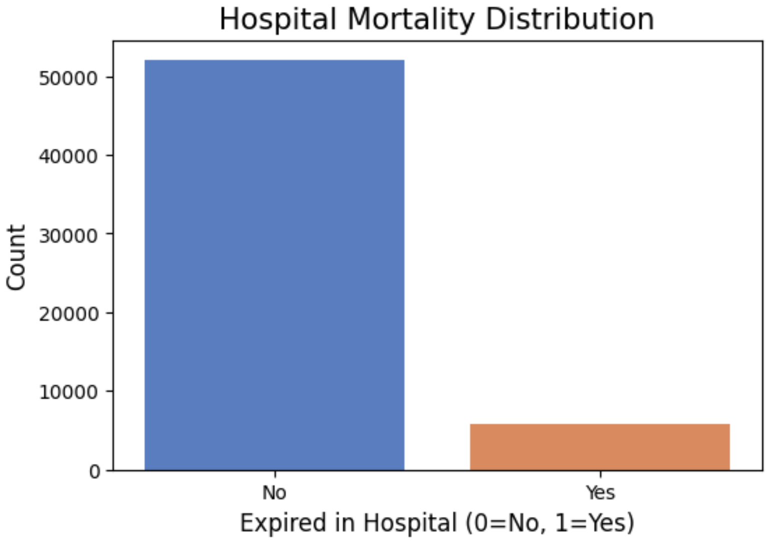


Figure 3:Mortality Count

2. **Age -Related Trends: Death, Mortality Rate and Admission**

   The age at death is similar across genders, with most patients dying in older age as seen in Figure 3.7. Mortality rates increase significantly for patients aged 80 and above (Figure 3.8). The dataset skews towards older adults aged 50-70, reflected in the distribution of age at admission (Figure 3.8). The focus on older adults is key for understanding age-related health conditions.

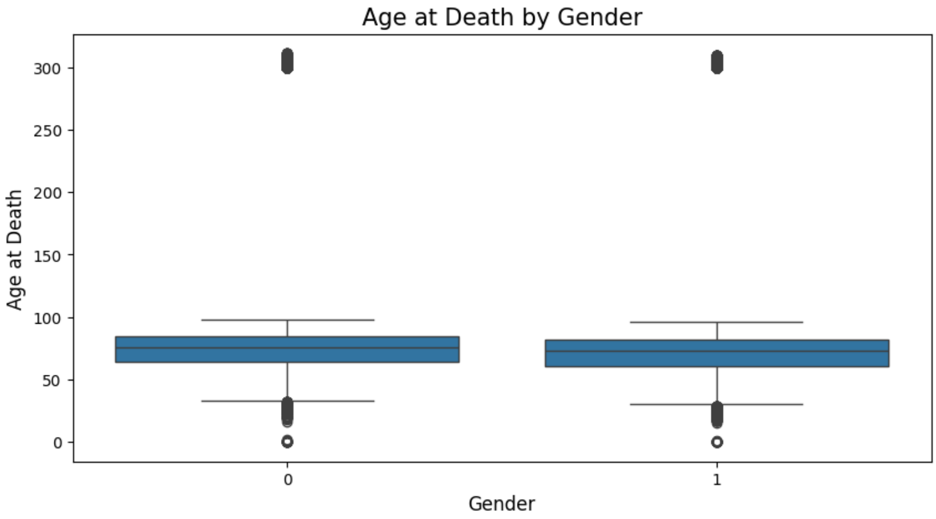


Figure 4: Age at Death by Gender

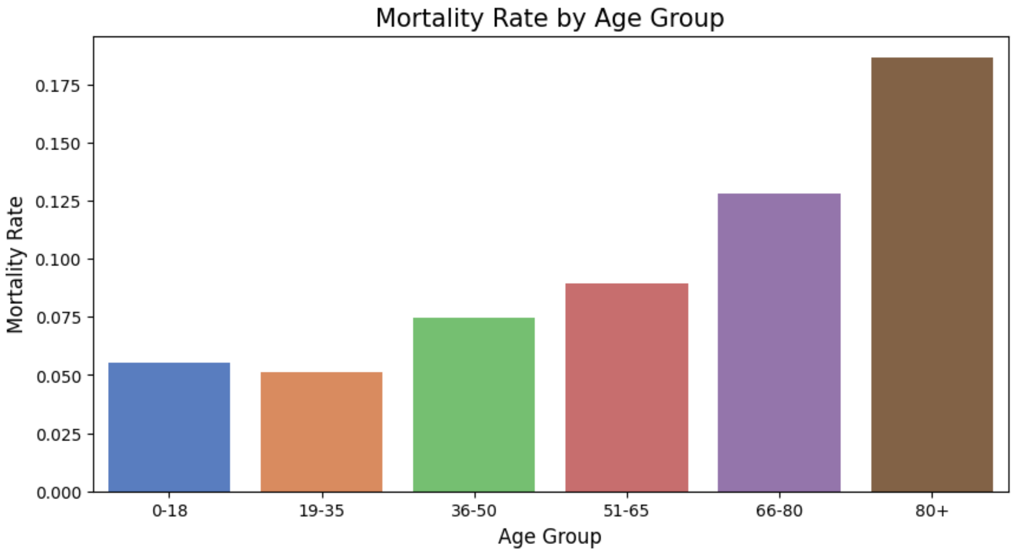


Figure 5:Mortality Rate

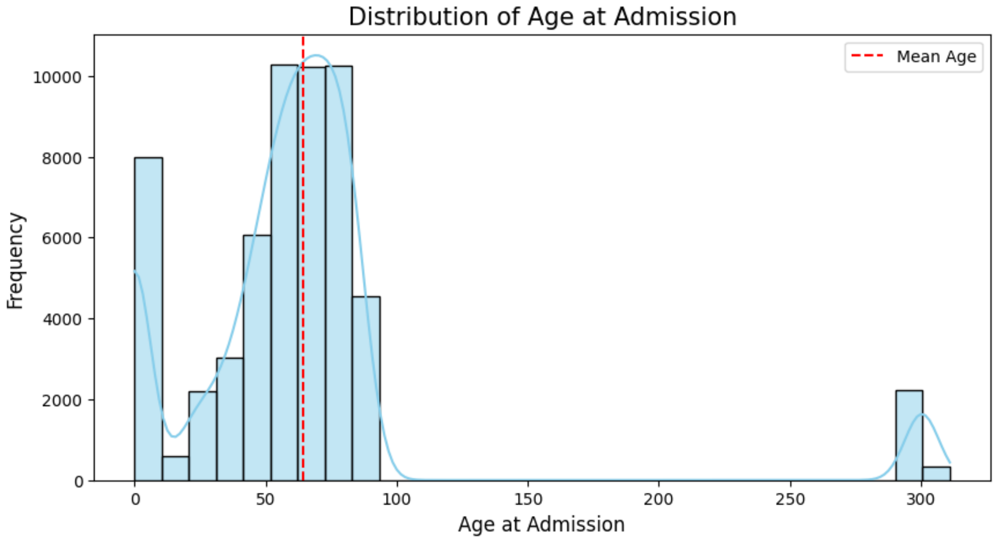


Figure 6: Age at Admission Histogram

Clinical Features Analysis

Analysing patient diagnosis to understand common conditions.

1. **Top Diagnoses**

   The most frequent ICD-9 codes include conditions like hypertension and heart failure (Figure 3.12). This concentration on common, severe conditions will help target the most impactful phenotypes in the patient population.

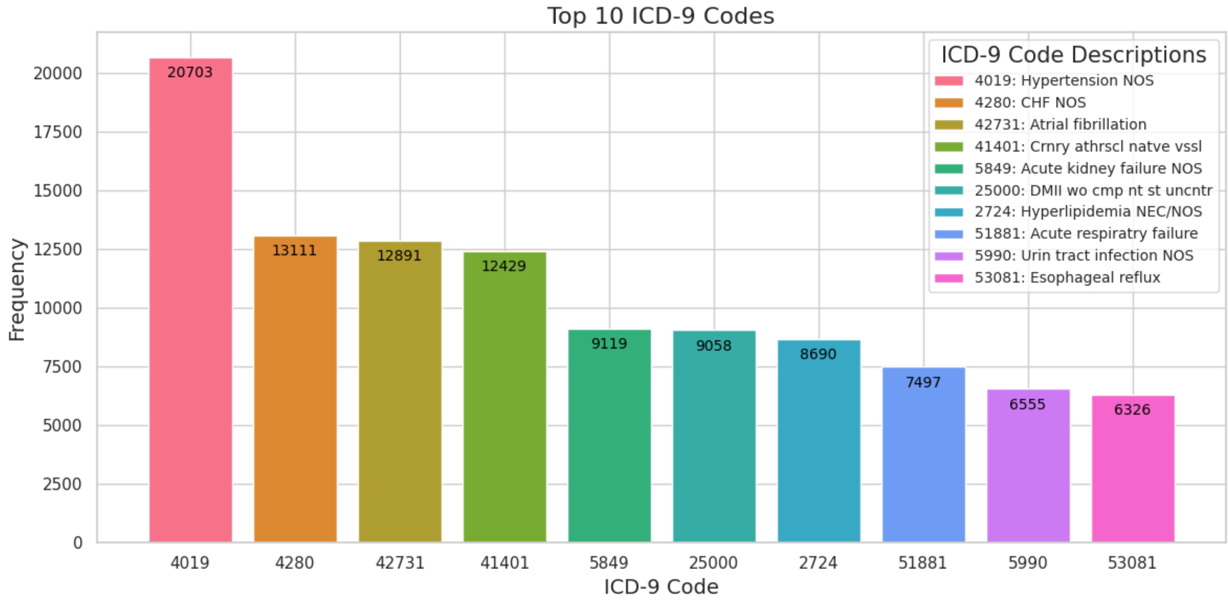


Figure 7:Top ICD-9 Diagnostic Codes

3.5.3 Analysis of Unstructured features

The analysis of unstructured text features from Discharge Notes is essential to group disease related symptoms, medication and treatments for patients.

1. **Distribution of Note Categories**

The distribution of different note categories like Radiology, ECG, Nursing etc. as seen in Figure 3.13, highlighted that ‘Discharge Summary’ constitutes only around 3% of all EHR notes. This indicates the specific and detailed nature of discharge summaries, which contain critical patient information.

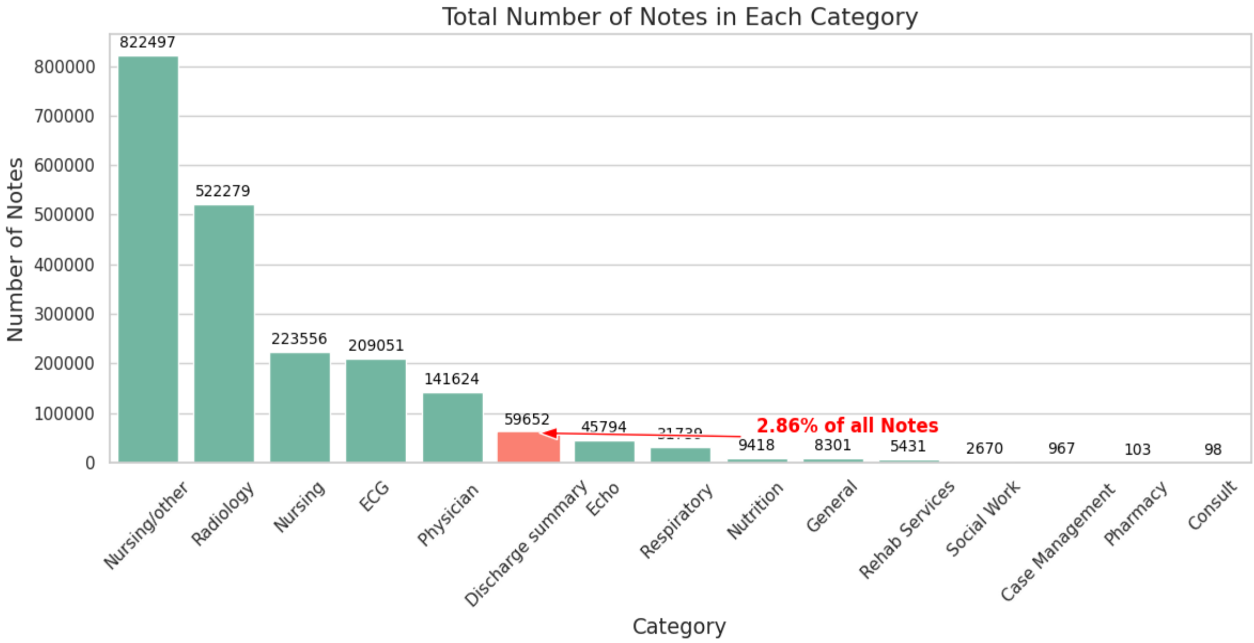
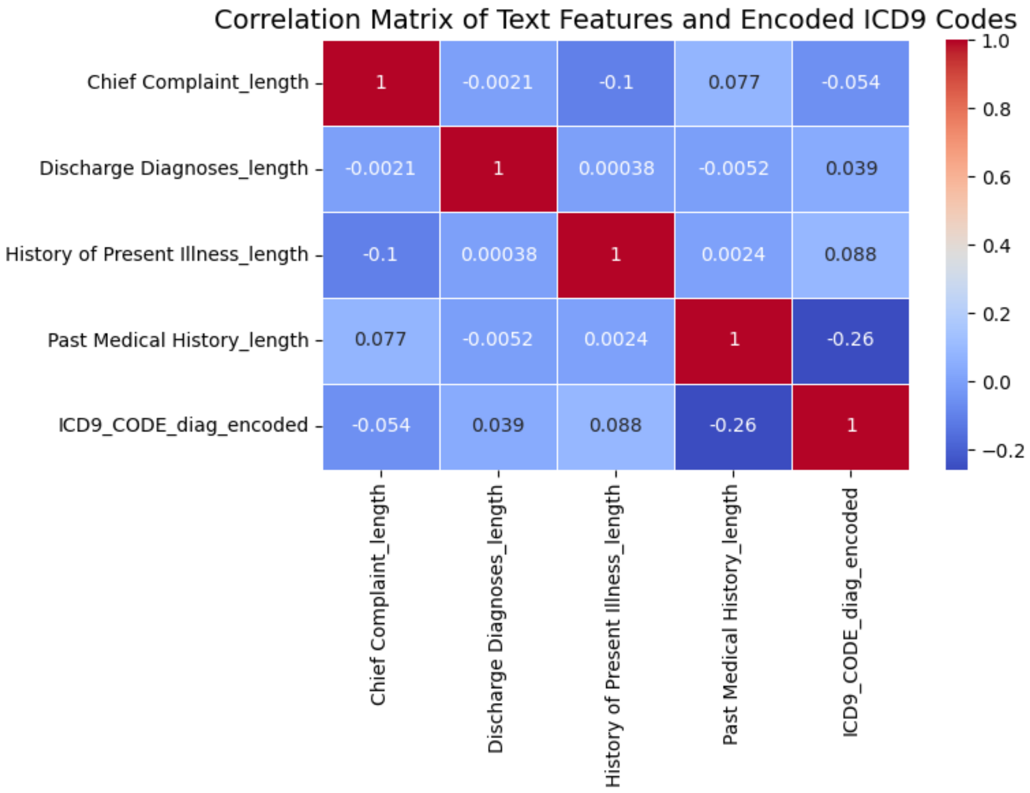


Figure 8: Distribution of EHR Notes Category

2. Correlation of <u>Unstructured text features</u> with <u>Structured ICD9 Codes</u>

The correlation matrix between text features and diagnostic codes reflect that the correlation is generally low, with ‘Discharge Diagnosis’ and ‘History_of_Present_Illness’ showing highest though still weak correlation with ICD codes. These unstructured features will provide contextual information after preprocessing thoroughly (Figure 3.14).



Correlation Matrix of Text Features and Encoded ICD9 Codes

	Chief Complaint_length	Discharge Diagnoses_length	History of Present Illness_length	Past Medical History_length	ICD9_CODE_diag_encoded
Chief Complaint_length	1	-0.0021	-0.1	0.077	-0.054
Discharge Diagnoses_length	-0.0021	1	0.00038	-0.0052	0.039
History of Present Illness_length	-0.1	0.00038	1	0.0024	0.088
Past Medical History_length	0.077	-0.0052	0.0024	1	-0.26
ICD9_CODE_diag_encoded	-0.054	0.039	0.088	-0.26	1

Figure 9: Heatmap of Subheadings and ICD Codes

3. Top Words in Discharge Summaries

Top 20 most frequent words in discharge summaries (Figure 3.15), included medical terms like “patient”, “pain “and “tablet”. These are general medical terms; our preprocessing strategy will focus on understanding themes related to disease diagnoses and symptoms.

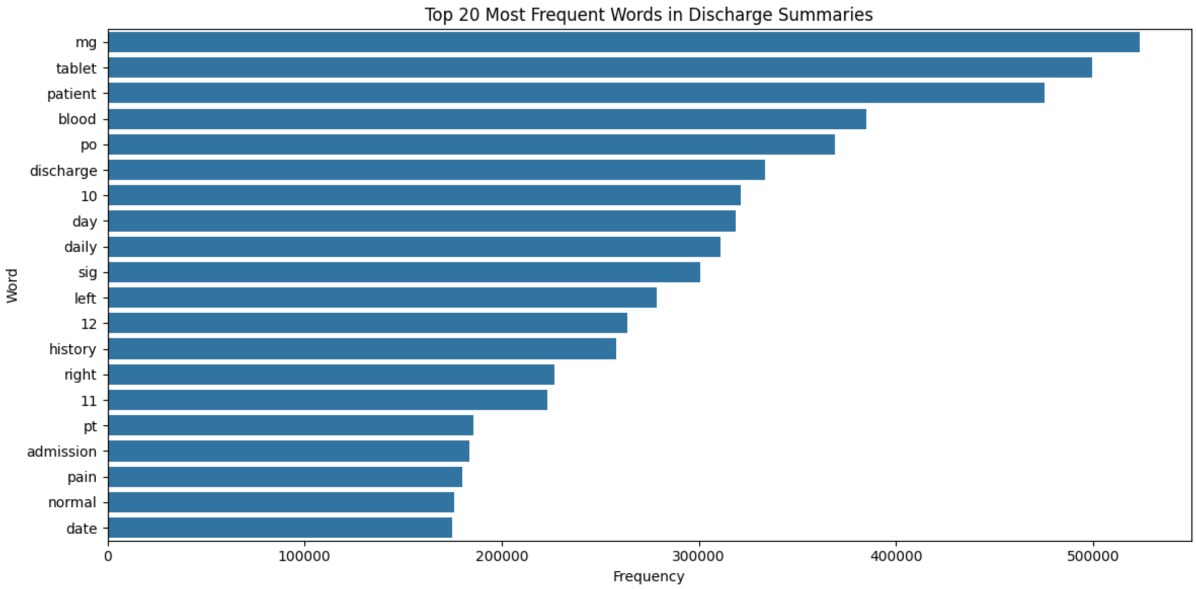


Figure 10:Frequent words in Raw Discharge Summaries

3.6 Data Preparation and Integration

The primary objective of this section is to prepare and merge the datasets that we will be utilised later for computational phenotyping tasks. The following steps were undertaken to prepare the MIMIC-II dataset for further analysis:

3.6.1. Data Extraction

The initial phase involved the extraction of pertinent tables from the MIMIC-III database, which encompasses of structured modality like patient demographic information, admissions data, ICU stay records, diagnosis codes, and unstructured clinical notes data. These six tables were extracted from the entire database, owing to their multimodal nature which will be used for robust analysis.

3.6.2. Data Cleaning and Normalisation

This step focuses on standardizing the format of structured data fields, handling missing values and maintaining consistency across tables. To ensure the integrity of the tables, missing data were handled by dropping the records having incomplete or duplicate data, we chose this approach since computational phenotyping requires accurate data and imputation of missing records could compromise the quality of data owing to the sensitive nature of medical information.

Normalisation of Patients, Admissions and ICUStays tables

- These tables included key patient information such as age, gender, ethnicity, admission and discharge times, length of ICU stay and other demographic details.
- **Conversion of data** : Transformed the columns representing Date of Birth, Date of Death, Admission time, discharge time, death time and ICU intime/outime to datetime format to guarantee uniform chronological data integrity.
- Conversion of categorical fields: Assigned numerical encoding to GENDER column, designating 'M' as 1 and 'F' as 0.
- Determined **Age_at_Death** feature by calculating the difference between the year of death and year of birth.
- **Handled missing values in categorical columns**: Imputed 'Unkown' as the replacement for null values like Ethnicity, Marital status and similar columns.

Normalisation of structured Diagnoses_ICD table

- **Filtering Primary Diagnoses**: Isolated the initial diagnosis (SEQ_NUM=1.0) for each patient admission to concentrate on primary medical conditions and mitigate data redundancy.

Preparation of Unstructured Clinical Notes (NOTEEVENTS) Table

- Concentrated on extracting **'Discharge Summary'** notes (3% of total notes) by filtering over 2 million records of NOTEEVENTS table specifically for this category. This research choses Discharge summary for their detailed account of a patient's hospital stay, providing valuable insights into the treatment, symptoms and outcomes of each patient.
- **Handled missing values** : As seen in Figure 3.16, there were four features that reported 100% missing values, hence adding no value to our dataset.

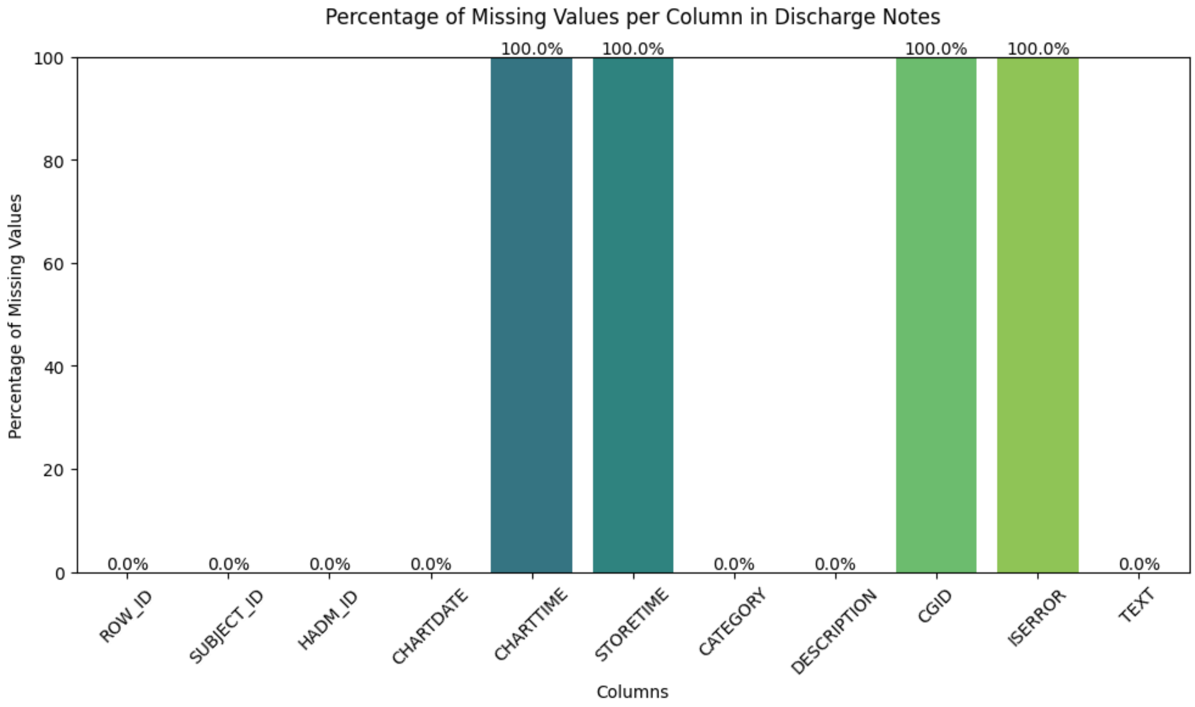


Figure 11:Null Values in Discharge Notes

3.6.3. Data Integration

The process of data merging step was crucial for combining structured codes with rich narrative clinical text from multiple sources to create a cohesive dataset for an in-depth and holistic patient record analysis.

Integrating Structured EHR data:

The integration of structured EHR data from the MIMIC-III dataset involves merging multiple tables to create a cohesive dataset. It starts with linking PATIENTS table containing demographic information, to the ADMISSIONS table using the unique patient ID, resulting in a dataset of 58,967 patients. Next, ICU stay information is added from the ICUSTAYS stable via hospital admission ID, refining the dataset to around 57,000 patients.

Integrating Unstructured Clinical Notes modality:

To augment the dataset, unstructured clinical notes from the NOTEEVNTS table, specifically 'Discharge Summary' clinical notes are integrated. These notes provide comprehensive overviews of patient care during hospital stays and are linked to the existing dataset via admissions ID. This addition offers valuable contextual information for NLP and advanced phenotyping pipeline, capturing nuances of patient care that structured data alone cannot convey.

Comprehensive Dataset:

The integration process results in comprehensive dataset that combines demographic information, admission records, ICU details, primary diagnosis codes, and unstructured clinical notes for each patient. This dataset which contains multimodal data, supports feature extraction, natural language processing, and computational phenotyping in the following chapters, enabling the development of models that accurately represent patient health.

3.6.4. Feature Extraction and Engineering

The primary task here is to extract relevant features and structure the dataset to prepare the foundation for subsequent data processing.

Extracting Relevant Subheadings from Unstructured Notes

- This study applied sophisticated **regular expressions** and text parsing algorithms to detect and **isolate key subheadings as separate features**. Clearly defined sub-sections of unstructured discharge summaries, such as 'Chief Complaint' , 'Discharge Diagnoses', 'History of Present Illness', 'Past Medical History' and 'Discharge Medications'.

- Features were integrated by combining **Discharge Diagnoses** and **Chief Complaint** , to create detailed records that have the potential to improve the effectiveness of topic modelling through richer contextual insights.

- **Clinical-Text Features Analysis (Subheadings)**
  Figure 3.17. illustrates the distribution of text length across different subheadings. Notably, 'Past Medical History' is extensive and Discharge Diagnosis are smaller but they result in faster processing owing to limited computational resources for this project.

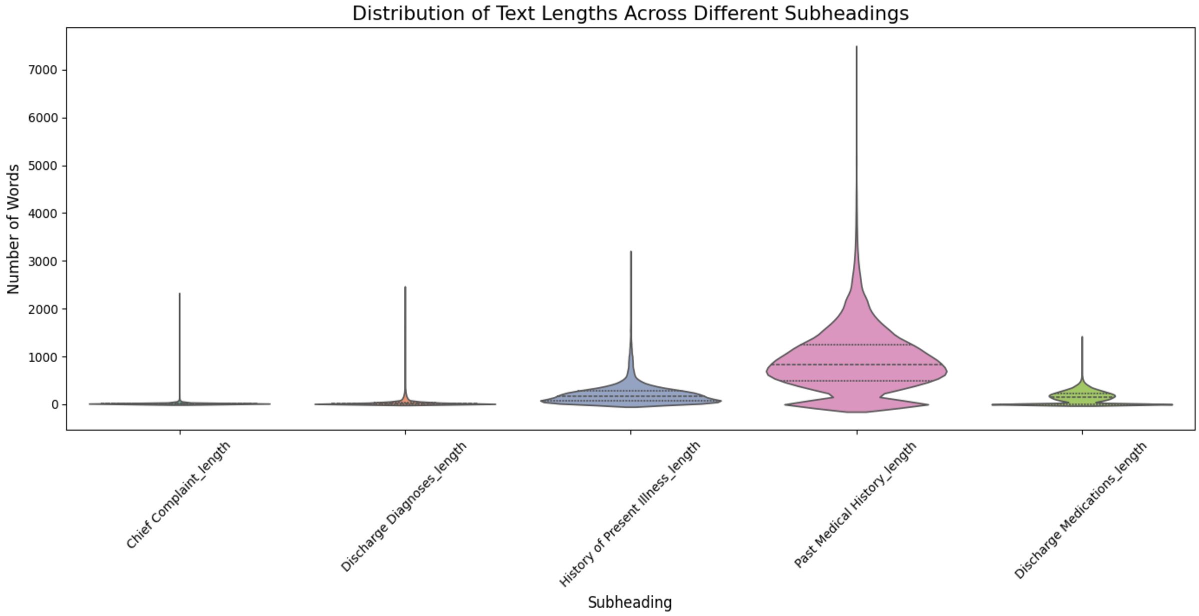


Figure 12:Text-Length of Engineering Features.

Feature Engineering from Structured Data

- Structured data was transformed into usable features by categorizing demographic details, vital signs, and ICD codes, such as grouping age into defined ranges and ICD codes into broader disease categories, improving input for analytical modelling.

3.7 Unstructured Clinical Text Preprocessing

The preprocessing of unstructured clinical data is essential for the effective phenotyping analysis. This research developed a specialized pre-processing pipeline to analyse discharge diagnosis feature from the comprehensive multimodal dataset.

Pipeline Key Stages:

1. **Text Cleaning , Sentence Segmentation** and **Tokenization**: The initial phase involves the elimination of extraneous elements from the text. Subsequently the refined text undergoes sentence boundary detection and tokenization into discrete lexical units. These actions mitigate noise and preserve pertinent clinical data for further analysis.

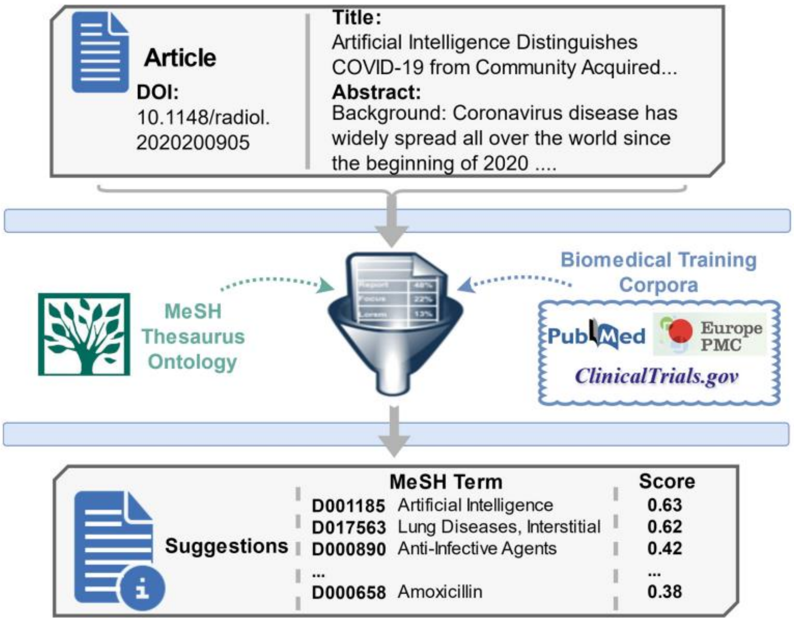
2. **Lemmatization** and **Negation Detection**: Post-tokenization, the lexemes are lemmatized to their fundamental forms. This stage also involves identifying negated terms using negspaCy to eliminate those that signify the absence of conditions or symptoms. Such measures ensure that important, affirmative clinical information is preserved, thereby reducing the risk of misinterpretation.

3. **Stopword removal:** By eliminating stopwords at this point, the remaining

vocabulary retains integrity becomes more relevant for analysis, further reducing noise.

4. **Medical Ontology Mapping:** A vital aspect of the pipeline involves linking the text to the medical vocabulary - Medical Subject Headings(MeSH) terms using an n-gram approach to generate relevant medical concepts. This framework (Figure 3.18) aligns unstructured text with standardised medical concepts to capture both single and multi-word medical entities , thus enhancing the interpretability and clinical relevance of extracted words.

5. **Named Entity Recognition (NER):** Following the MeSH mapping, NER was utilized to identify and categorize key medical entities within the processed text, such as diseases, medications and symptoms. These techniques enrich the dataset by converting unstructured text into structured information.



DOI: 10.1148/radiol.2020200905

Title: Artificial Intelligence Distinguishes COVID-19 from Community Acquired...

Abstract: Background: Coronavirus disease has widely spread all over the world since the beginning of 2020 ....

MeSH Term		Score
D001185	Artificial Intelligence	0.63
D017563	Lung Diseases, Interstitial	0.62
D000890	Anti-Infective Agents	0.42
...		...
D000658	Amoxicillin	0.38

Figure 13: MeSH annotation mapping (Koutsomitropoulos & Andriopoulos, 2020)

Batch Preprocessing

Given the **high computational requirements** of MeSH mapping and NER, batch processing was adopted to handle these tasks more efficiently by reducing the load.

Discharge Diagnosis Analysis Post-Processing

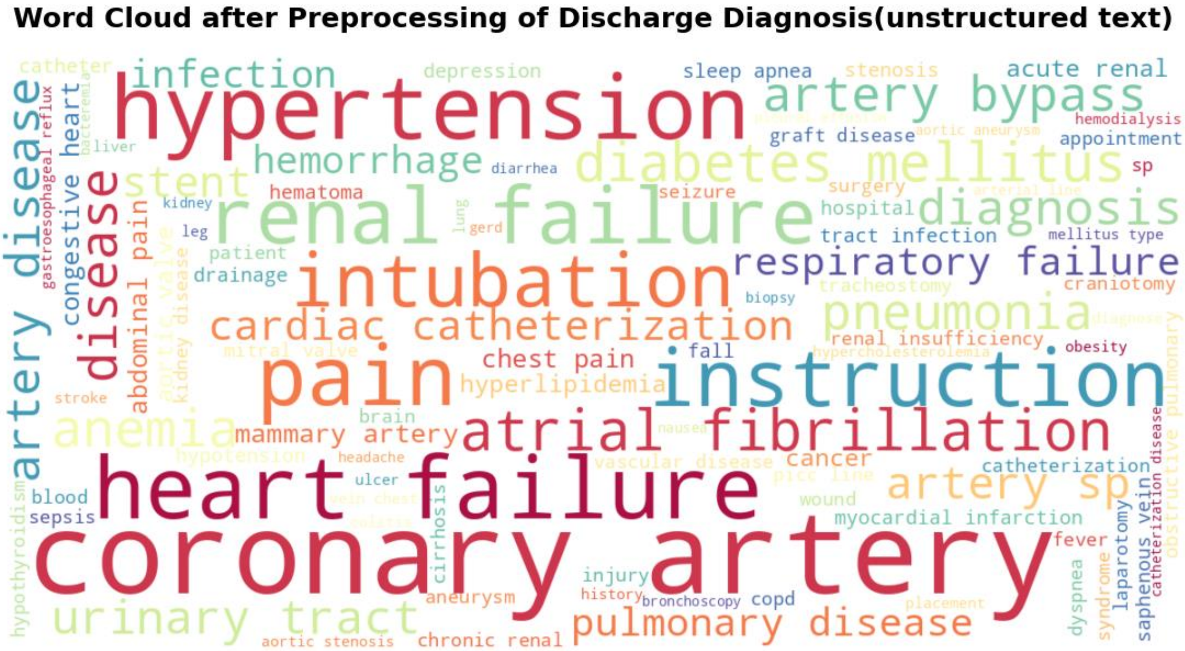

Word Cloud after Preprocessing of Discharge Diagnosis(unstructured text)

Figure 14:Word Cloud after Pre-processing

Overall, the pre-processed text column **'Discharge Diagnosis'** as depicted by the word cloud in Figure 3.18, is now well-prepared for computational phenotyping focusing on medical terms

3.8 Multimodal Data Fusion

Multimodal fusion is a vital component of this research methodology, focusing on the effective vectorization and integration of structured and unstructured data to prepare the dataset format for phenotyping models.

Applying **Late Fusion** Strategies:

1. **For TF-IDF based Modelling Methods**: Clinical notes, and ICD descriptions were combined into a single text block, using Text Frequency – Inverse Document Frequency (TF-IDF) to create a cohesive feature space, leveraging the strengths of both structured codes and clinical narrative text.

2. **For ClinicalBERT Embedding-Based Modelling Methods**: Separate Embeddings for clinical notes and ICD descriptions were concatenated before modelling, preserving the unique insights and semantic context of each data modality.

These late fusion techniques are applied on structured and unstructured data right before topic modelling and clustering, ensuring accuracy and robustness of computational phenotyping by capturing the complexity of clinical data.

3.9 Modelling and Evaluation

In this research methodology, we incorporate both unsupervised machine learning as well as deep learning algorithms to implement computational phenotyping models that identify clinically meaningful disease topics from multimodal EHR data. We integrate multimodality into these models and evaluate the impact on the overall coherence and quality of the disease phenotypes.

The modelling procedure involved two key steps: **Topic Modelling** and **Clustering for Phenotype Identification.**

2.9.1. Topic Modelling to Identify Disease Topics

Traditional ML methodologies utilised **TF-IDF** for multimodal data to construct a document-term matrix. This text representation allows models to effectively utilize both **structured ICD descriptions** and **unstructured diagnosis notes,** revealing underlying themes in EHRs.

1. **Latent Dirichlet Allocation (LDA) with TF-IDF features:**

   LDA represents a generative probabilistic model utilized for the purpose of highlighting latent topics that are embedded within textual documents. For this research, LDA assumes that each document(patient record) is a mixture of multiple topics. LDA was chosen for its effectiveness in uncovering patterns within clinical texts, crucial for interpreting medical datasets.

   Mathematically, LDA is defined as:

$$p(w_d|\alpha,\beta) = \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

Equation 1: LDA Equation

   where *p(wd|α,β)* represents the probability of words in document d, given the Dirichlet prior parameters *α* and *β*.

2. **Non-negative Matrix Factorization (NMF) with TF-IDF features:**

   NMF being a dimensionality reduction technique was employed to decompose a non-negative matrix into two lower-rank matrices as depicted by Devarajan. (2008), thereby revealing latent structures(topics) (Figure 3.19). When applied to the clinical TF-IDF, NMF facilitates the identification of discrete clinical phenotypes by breaking down the data into additive components. NMF was selected for its ability to generate clear , interpretable topic distributions. Equation 2 explains the topic calculation.

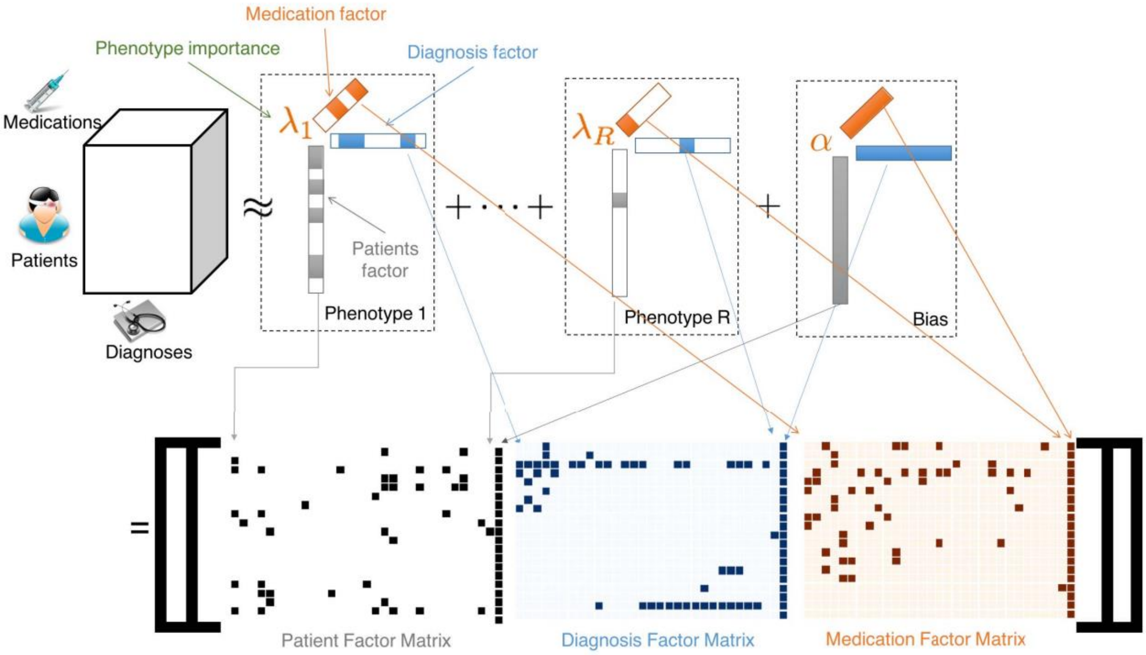


Figure 15:Phenotyping via Tensor Decompositions (Henderson et al., 2017)

Optimization problem for NMF is :

$$\min_{W,H} \|V - WH\|_F^2 = \sum_{i,j} (V_{ij} - (WH)_{ij})^2$$

Equation 2: NMF Optimization Equation

where *V* is the TF-IDF matrix, *W* is the topic matrix, and *H* is the topic distribution.

3. **BERTopic with ClinicalBERT Embeddings:**

ClinicalBERT embeddings, fine-tuned for clinical documentation, effectively captures the complex language of medical information by converting structured and unstructured data, like clinical notes and ICD classifications into dense, context-rich vectors.

BERTopic utilized the ClinicalBERT embeddings to generate dense document clusters, which were then refined into specific topics. BERTopic algorithm leverages the embedded components, UMAP for dimensionality reduction and HDBSCAN for hierarchical clustering. This approach was chosen for its ability to exploit the deep semantic context contained within ClinicalBERT. The model effectively manages the high dimensionality of EHR data by conducting clustering within the embedding space before refining the topics.

This technique uses the concept of mixed-membership, which assigs each patient document to a distribution of multiple topics, which results in multi-faceted phenotypes discovery.

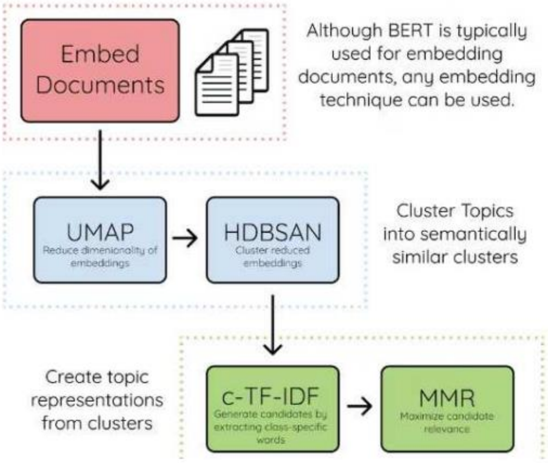

Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimenionality of embeddings
HDBSAN
Cluster reduced embeddings
Cluster Topics into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Figure 16:BERTopic Algorithm (Oveh et al., 2022)

2.9.2. Clustering for Phenotype Identification

After identifying topics through traditional and advanced topic modelling, the next step involves clustering to group patients based on their topic distributions, ultimately defining disease phenotypes

1. **K-means Clustering for LDA and NMF**

For Traditional approaches including LDA and NMF, K-means was utilized to partition patient records into clusters predicted on intrinsic similarities. When applied to **topic distributions**(the percentage membership of each topic within a patient document), This method was employed for its simplicity and efficiency in clustering high dimensional datasets.

K-means minimizes the within-cluster sum of squares(Equation 3) :

$$\min_{\{C_k\}_{k-1}^{K}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

Equation 3: K-means WCSS

Where $C_{k.}$ are clusters x_i are topic distributions of patients and μ_k ,cluster centroids.

2. **HDBSCAN for BERTopic**

   BERTopic incorporates HDBSCAN, a density-based clustering algorithm, in its topic modelling framework, HDBSCAN autonomously detects clusters(phenotypes) in the embedding space by locating high-density regions. HDBSCAN's parameters, including minimum cluster size and minimum samples size are finely adjustable, allowing for optimization of clustering.

2.9.3. Optimal Topics Determination and Model Evaluation:

To ensure robust model performance, various evaluation metrics and methods were employed to determine the optimal number of topics for each model as the choice of 'K' is a critical part of modelling phase.

- **Coherence Score** : Measures the semantic similarity of generated topics, with elevated scores reflecting enhanced interpretability.
- **Silhouette Score**: Evaluates clustering effectiveness by analysing how similar a patient’s topic distribution is to its assigned cluster, identifies optimal number of clusters in K-means.
- **Elbow Method:** Applied in K-means to evaluate the optimal number of clusters.
- **Topic Diversity** :This metric ensures that model produces distinct, comprehensive topics and reduces redundancy
- **Perplexity:** A measure of how well a probabilistic model predicts a sample, lower perplexity indicates enhanced performance.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results

This research examines the effectiveness of basic and advanced computational phenotyping techniques: LDA combined with K-means clustering, NMF in conjunction with K-means clustering and BERTopic with ClinicalBERT embeddings. Each method was evaluated based on its ability to generate coherent topics, maintain topic diversity, minimize perplexity, and achieve clear cluster separation (Table 4.1). These metrics provide critical insights into the quality of disease topics and uncover disease phenotypes that can enhance patient stratification and improve personalized treatment strategies.

4.1.1 Comparative Analysis of Phenotyping Pipelines

Method	Topic Coherence	Topic Diversity	Perplexity	Silhouette Score	Total Number of Phenotypes
LDA + K-means (TF-IDF)	0.6025	0.8909	1653.11	**0.0206**	11
NMF + K-means (TF-IDF)	**0.6604**	**0.8846**	810531	0.2932	13
BERTopic (ClinicalBERT)	0.5365	0.7368	**1.892**	0.0167	18

Table 4.1: Comparative Performance Evaluation of Phenotyping Pipelines

LDA + K-means(TF-IDF) yields a moderate topic coherence and high topic diversity, making it effective at generating a broad spectrum of distinct topics. Nevertheless, the relatively low silhouette scores indicates that the clusters produced are inadequately defined, signifying potential overlap or inconsistency within the topic assignments. This methodology, which results in the formation of 11 disease phenotypes, may be particularly advantageous for exploratory analyses where the topic breadth is prioritized over precision.

NMF + K-means stands out by exhibiting the **highest levels of topic coherence and silhouette score**, signifying that the topics generated possess semantic significance and that the clusters are well-defined. Such attributes render NMF suitable for applications that necessitate clear and interpretable topics. However, the elevated perplexity indicates poor generalization of the model.

BERTopic (ClinicalBERT), although exhibits lower coherence and diversity but achieves the lowest perplexity, indicating strong generalization to unseen data. The generation of **18 possibly overlapping topics** underscores, its nuanced approach to subtle clinical variations, making it valuable for detailed clinical evaluations where both interpretability and generalization are of paramount importance. (Refer Table 4.2 for clear view)

Method	**Data Representation**	**Embedding Technique**	**Clustering Algorithm**	**Topic Coherence**	**Topic Diversity**	**Interpretability**
LDA with K-means (TF-IDF)	Discharge Diagnoses, ICD Descriptions	TF-IDF	K-means	Moderate	High	Moderate
NMF with K-means (TF-IDF)	Discharge Diagnoses, ICD Descriptions	TF-IDF	K-means	**High**	**High**	High
BERTopic (ClinicalBERT)	Discharge Diagnoses, ICD Descriptions	ClinicalBERT	Hierarchical Clustering	Moderate	Moderate	**Very High**

2.Table 4.2: Method Comparison for Topic and Phenotype Interpretability

4.2.2 Detailed Analysis of Basic Phenotyping Method

Method 1: LDA with K-Means Clustering

1. Optimal Number of Topics

Through coherence analysis, it was determined that the optimal number of topics was **11,** which produced the highest coherence score (Figure). This suggests that at 11 topics LDA model was best at capturing coherent topics from HER.

2. Clustering Analysis

K-means clustering was applied to LDA topic distributions of each patient, identifying **11** as the optimal phenotypes count using Davies-Bouldin Index(Figure 4.1,4.2). The clustering results provided a basis for assigning phenotypic groups to patients based on their EHR data.

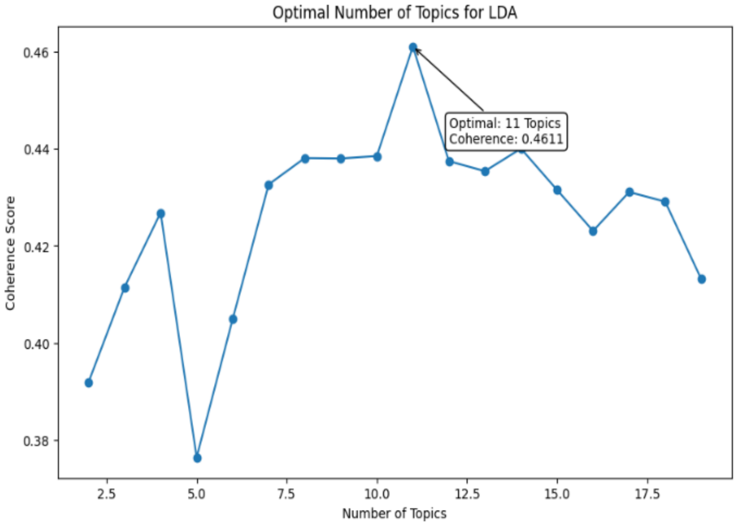


Figure 17:Coherence plot for Optimal LDA Topics

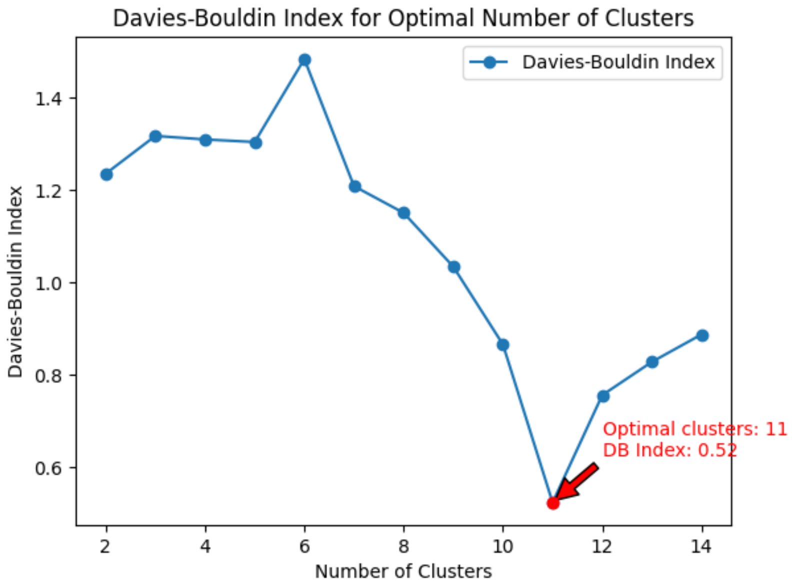
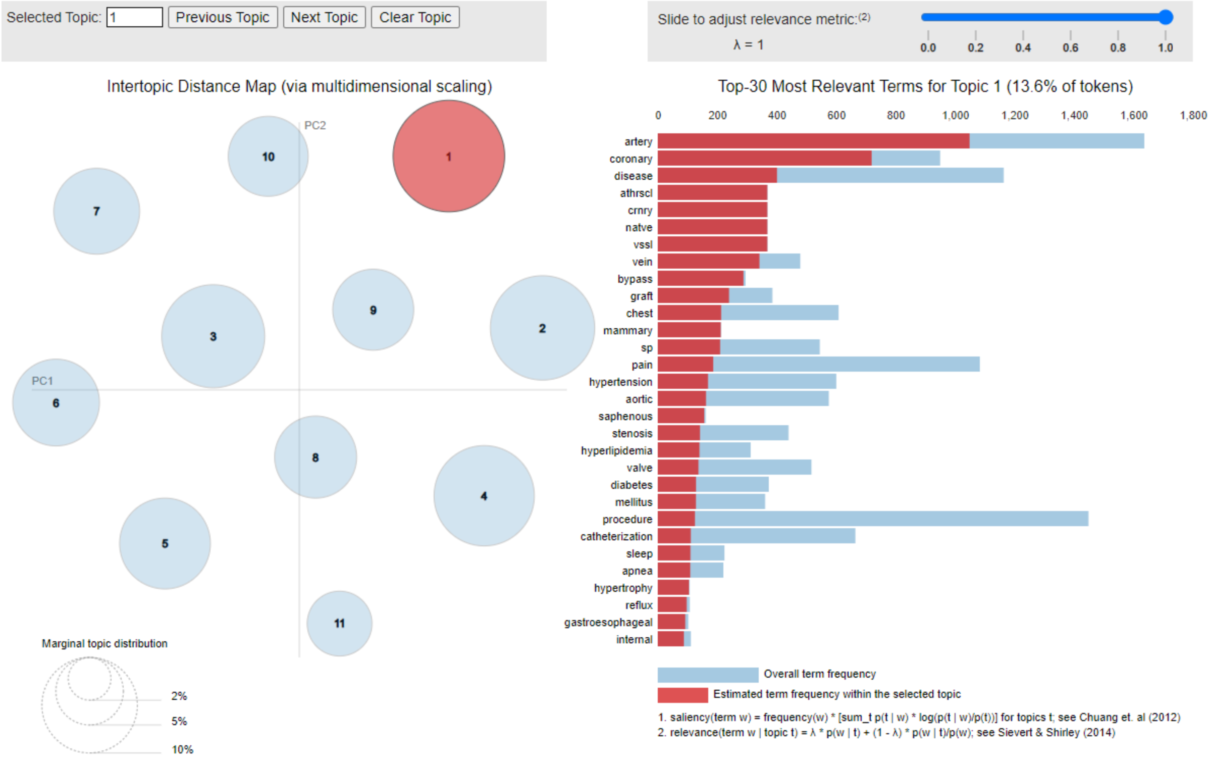


Figure 18: Davies-Bouldin Index for determining optimal clusters

Figure 4.3 shows intertopic distance which provides a clear representation of topic distributions, with an emphasis on most relevant words.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 19:Intertopic Distance Map (LDA)

3. Assigning Patients to Phenotypic groups

Each cluster was analysed to assign a disease phenotype name based on top topic and patient membership to each topic. Refer **Appendix B.2** for detailed cluster analysis.

- **Cluster 0 Analysis**
  - **Top Topics**: Heart failure, Respiratory failure, Cancer
  - **Phenotype Name**: **Cardiopulmonary and Oncological Disorders**
  - **Description**: This cluster encompasses patients with heart failure, respiratory and oncological disorders.

4. Common Patient Characteristics Within Phenotypic Groups

The analysis of patients within same phenotypic group revealed key demographic and clinical profiles for each phenotype in Table 4.3. For instance, cluster_0 had a mean age of 81.5 years, a nearly balanced gender distribution and was mainly linked to ICD9 code 389 (Figure 4.4).

Cluster	Cluster Names	Mean Age	Min Age	Max Age	Female Percentage	Male Percentage	Most Common ICD9	Patient Count
1	Cardiopulmonary and Oncological Disorders	81.15	18	309	45.84	54.16	389	3746
2	Abdominal and Vascular Conditions	71.71	16	307	52.18	47.82	389	2543
3	Neurological Hemorrhage and Respiratory Disorders	77.58	0	308	46.47	53.53	431	2524
4	Respiratory and Cardiovascular Conditions	79.78	17	310	45.29	54.71	389	4003

5	Ischemic Heart Disease	72.49	22	308	29.72	70.28	41401	3853
6	Stroke and Respiratory Failure	83.16	19	306	49.54	50.46	43411	985
7	Hepatic and Respiratory Disorders	57.60	18	306	39.68	60.32	5712	2243
8	Oncological and Abdominal Disorders	71.60	17	308	46.00	54.00	1983	3959
9	Cardiovascular Disorders	73.27	18	309	47.95	52.05	4241	2559
10	Trauma and Hemorrhagic Disorders	66.58	15	306	38.44	61.56	41519	2716
11	Advanced Cardiac Disease	78.14	22	309	37.86	62.14	41071	1717

3.Table 4.3: Patient Characteristics Within Phenotypic Groups(LDA+Kmeans)

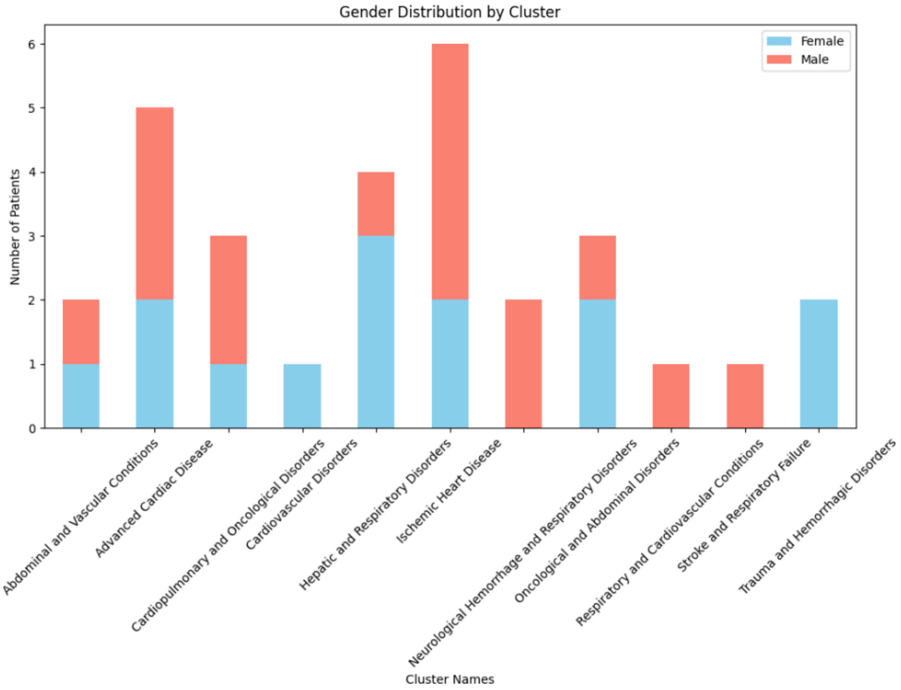


Figure 20:Gender Distribution across LDA Clusters.

5. Patient Cohorts for each Phenotypic group

The t-SNE visualization in Figure 4.5 demonstrates topics derived from LDA, facilitating comprehensive view of topic relationships.

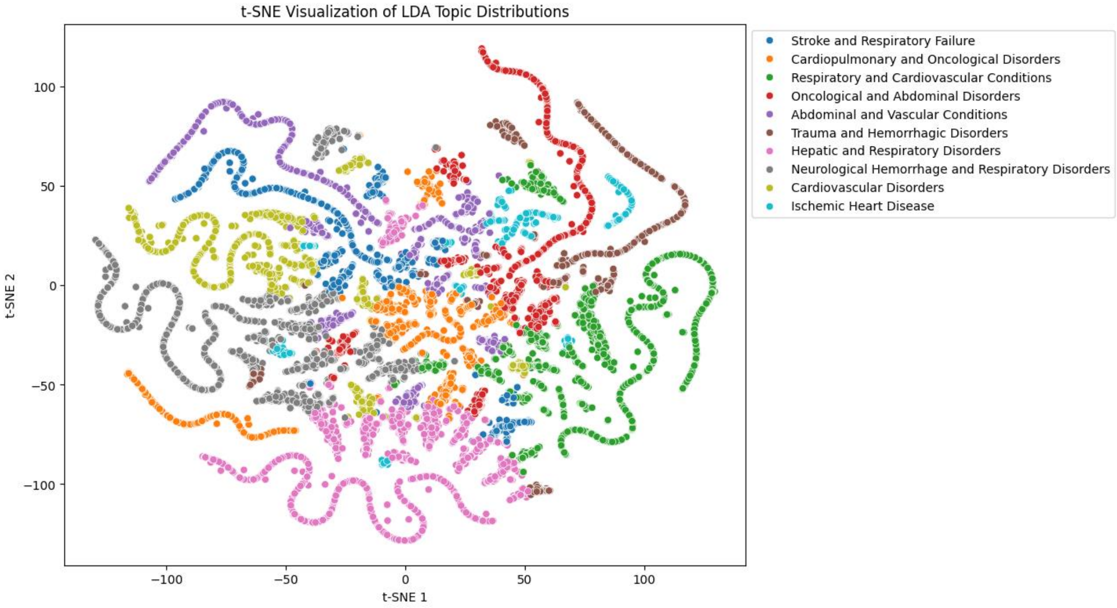


Figure 21: t-SNE Visualization of LDA Topic Distributions

Method 2: NMF with K-Means Clustering

1. Optimal Number of Topics

The NMF model was found to be the most interpretable with **13** topics as seen in (Figure 4.6). The ability of NMF to produce non-negative components facilitated a more interpretable factorization of the EHR data, leading to the emergence of coherent topics.

2. Clustering Analysis

Silhouette analysis revealed that **8** clusters achieved superior separation, denoting more distinct phenotypic groups. K-means resulted in non-overlapping phenotypes from the data.(Figure 4.7)

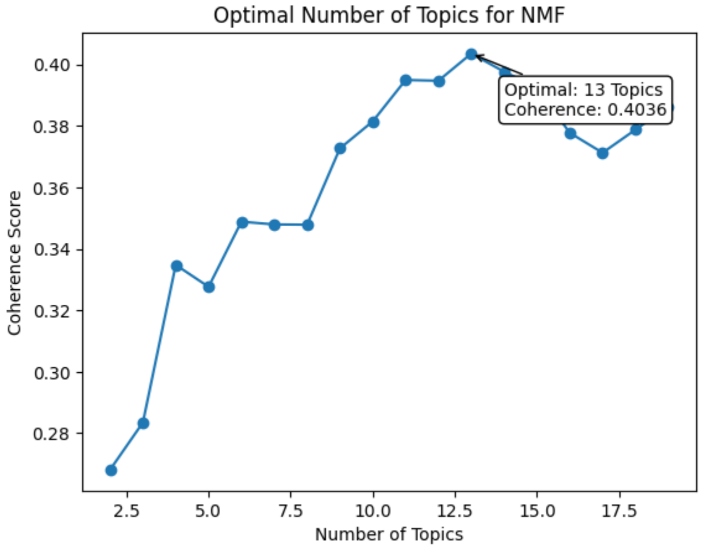


Figure 22:Optimal number of Topics in NMF

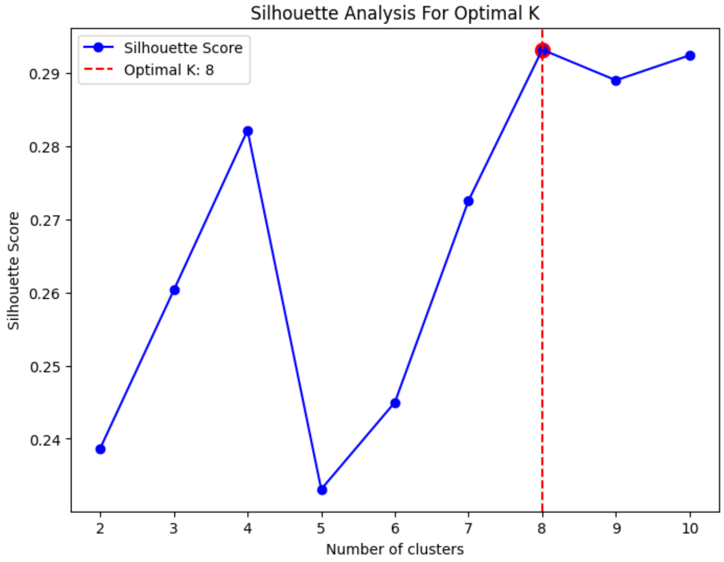


Figure 23:Silhouette Analysis for Optimal Clusters

3. Assigning Patients to Phenotypic groups

As with LDA, each cluster was analysed to determine the phenotype names

Cluster 0 Analysis:

- **Top Topics**: Heart failure, pulmonary disease, and Renal Disorders
- **Phenotype**: **Cardiopulmonary and Oncological Disorders**
- **Description**: This cluster reflects patients with multiple chronic ailments impacting the key organs in heart, lungs and kidneys are interconnected.

Similarly, Refer **Appendix B.2** for detailed analysis.

4. Common Patient Characteristics Within Phenotypic Groups

The phenotypic groups highlight distinct clusters of patients on Table 4.4. For instance, ''Chronic Cardiopulmonary and Renal Disorders'' group has a higher mean age of 85 years and a male-dominated demographic, reflecting patient profiles with chronic conditions.

Cluster	Phenotype Name	Number of Patients	Mean Age	Min Age	Max Age	Female Percentage	Male Percentage	Most Common ICD9	Patient Count
0	Chronic Cardiopulmonary and Renal Disorders	3533	85.33	17	308	43.62	56.38	51881	3533
1	Cardiovascular and Gastrointestinal Conditions	2720	70.68	25	306	24.96	75.04	41401	2720
2	Infectious Diseases and Oncological Conditions	18052	70.19	0	310	45.49	54.51	389	18052
3	Neurological and Hemorrhagic Conditions	801	84.06	17	305	35.71	64.29	85221	801
4	Cerebrovascular and Infectious Disorders	1017	73.80	17	306	48.97	51.03	431	1017

5	Cardiovascular and Pulmonary Disorders	2033	74.27	21	309	41.42	58.58	4241	2033
6	Renal and Infectious Conditions	1059	91.77	17	309	60.06	39.94	389	1059
7	Cardiac and Gastrointestinal Conditions	1633	77.48	23	309	36.80	63.20	41071	1633

4.Table 4.4: Patient Characteristics Within Phenotypic Groups(NMF+Kmeans)

4. Patient Cohorts for each Phenotypic group

The t-SNE plot in Figure 4.8 reveals how NMF topics form clusters when projected into a 2-dimensional space. Each cluster, indicates grouping of patients with similar context and gives significantly better distributions than LDA.

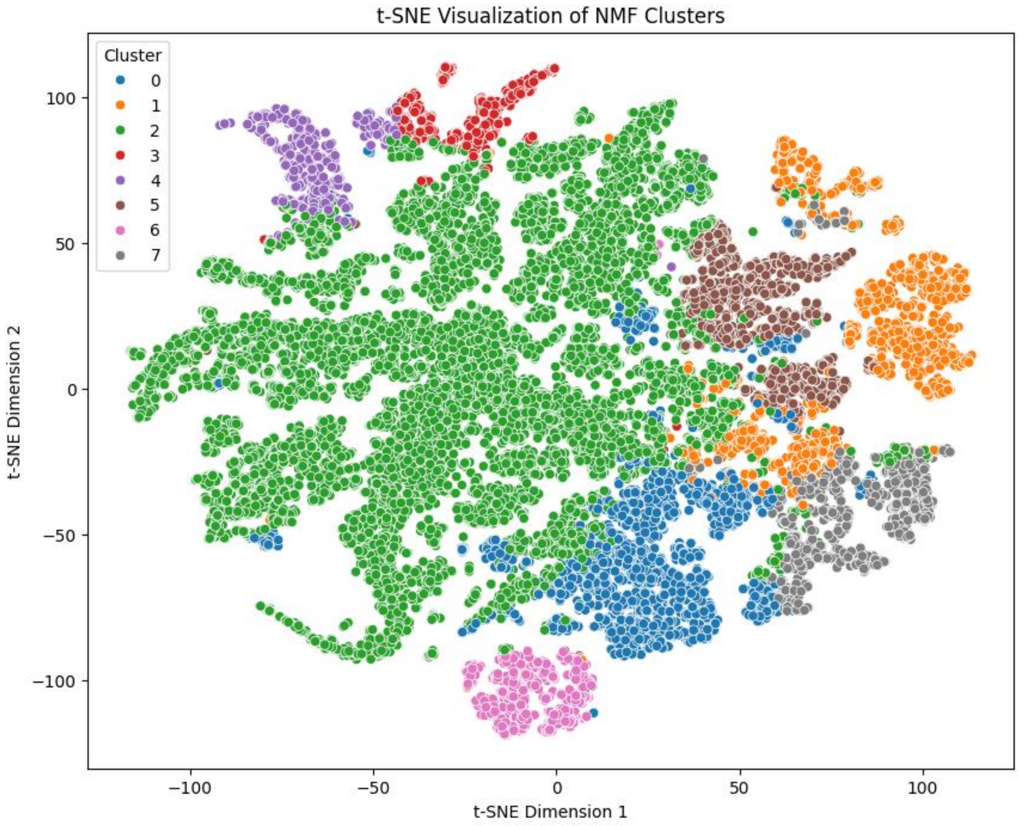


Figure 24: t-SNE Visualization of NMF Clusters

4.1.3 Detailed Analysis of Advanced Phenotypic Method.

Method 3 : BERTopic Modelling using ClinicalBERT Embeddings

1. Optimal Hyperparameters

BERTopic was applied to ClinicalBERT embeddings to capture intricate semantic relationships in clinical text. Hyperparameter tuning determined **15** neighbours and a minimum distance of **0.1** for UMAP (Dimensionality Reduction), with HDBSCAN (Hierarchical Clustering) set to a minimum cluster size of **500** and a minimum sample size of **100**. This configuration enabled the model to effectively capture dense, nuanced clinical topics.

2. Topic and Cluster Analysis

The model identified **18** topics, each representing distinct disease phenotypes, achieving a moderate score of **0.53.** It reflects that while the topics were **less cohesive** than those identified by NMF, they represented more specific and nuanced phenotypes, reflecting the complex nature of clinical language and the detailed understanding provided by ClinicalBERT embeddings. (Refer Figure 4.9)

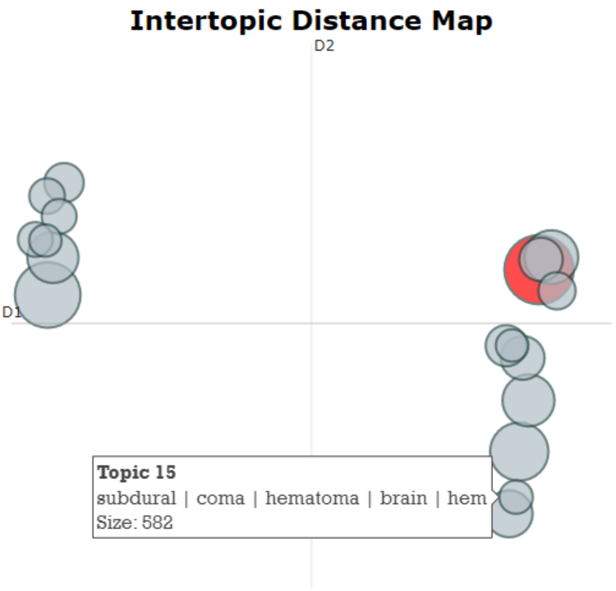


Figure 25:Intertopic Distance Map(BERTopic)

3. Analysis of Common Patient Characteristics Within Phenotypic Groups

A thorough analysis of patient characteristics in each phenotypic group from Table 4.5, showed that BERTopic effectively identified distinct topics with significant semantic meaning. For instance, the 'Acute Cardiac Conditions' topic was associated with a high incidence of acute cardiac events in both genders of average age of 69.

Refer **Appendix C.3** for detailed cluster analysis.

Topic	Count	BERTopic Name	Prevalence (%)	Patient Count	Mean Patient age
0	2493	Acute Cardiac Conditions	9.24533195	2316	69.4
1	2240	Cardiovascular and Vascular Disorders	7.35217842	2852	72.4
2	1767	Chronic Heart Failure and Disorders	7.20954357	930	76.9
3	1503	Gastrointestinal Disorders	5.54979253	1320	79.4
4	1408	Acute and Chronic Respiratory Conditions	9.02165456	1150	82.7
5	1361	Bacterial Infections and Sepsis	7.50778008	1155	80.0
6	1150	Hepatic and Renal Disorders	5.09595436	1572	80.3
7	998	Aneurysms and Myocardial Infarction	8.49649896	2621	80.8
8	991	Colonic and Hemorrhagic Conditions	7.05070021	912	77.0
9	882	Vascular Graft and Cardiac Device Complications	2.95643154	1201	68.4
10	806	Neurological Trauma and Hemorrhages	4.27904564	999	73.1
11	720	Sepsis and Systemic Infections	3.72795643	2268	76.7
12	657	Valvular and Aortic Disorders	3.01478216	2783	77.5

13	632	Intracerebral and Subarachnoid Hemorrhages	5.02787863	2175	68.5
14	626	Renal Failure and Acute Conditions	3.74416494	1551	78.3
15	582	Stroke and Cerebral Infarctions	3.58856328	1107	66.6
16	550	Trauma and Fracture Injuries	3.23845954	2224	65.0
17	542	Oncological Disorders	3.8932832	1712	71.1

5.Table 4.5: Patient Characteristics Within Phenotypic Groups(BERTopic+HDBSCAN)

3. Patient Cohorts for each Phenotypic group

UMAP in Figure 4.10 effectively clusters topics at granular level, allowing visualisation of topic similarity and density. The distinct clusters indicate clearly defined topics, while dense regions may suggest common or overlapping themes.

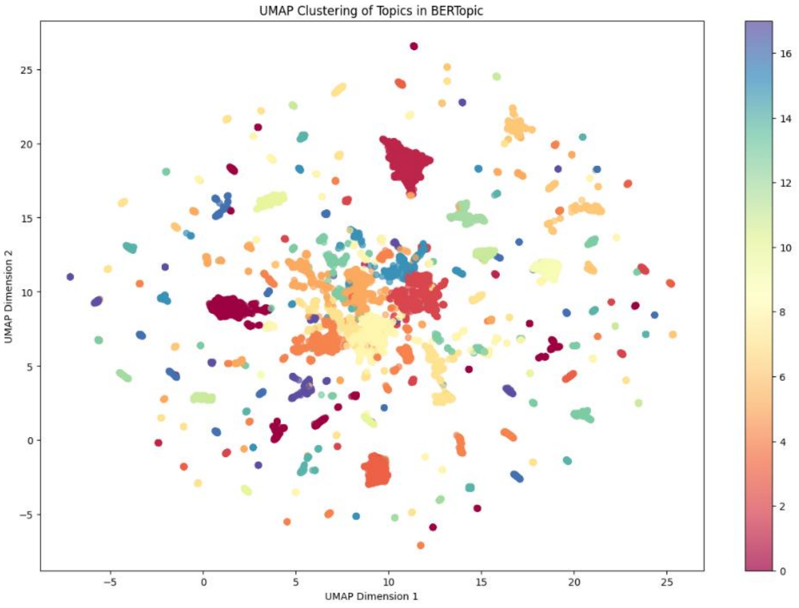


Figure 26: UMAP Clustering of Topics in BERTopic

4. Similarity Matrix of Topics

The heatmap in Figure 4.11 reveals relationships between topics, with higher similarity scores(darker shades) specifying more **overlap** or **shared patient characteristics**

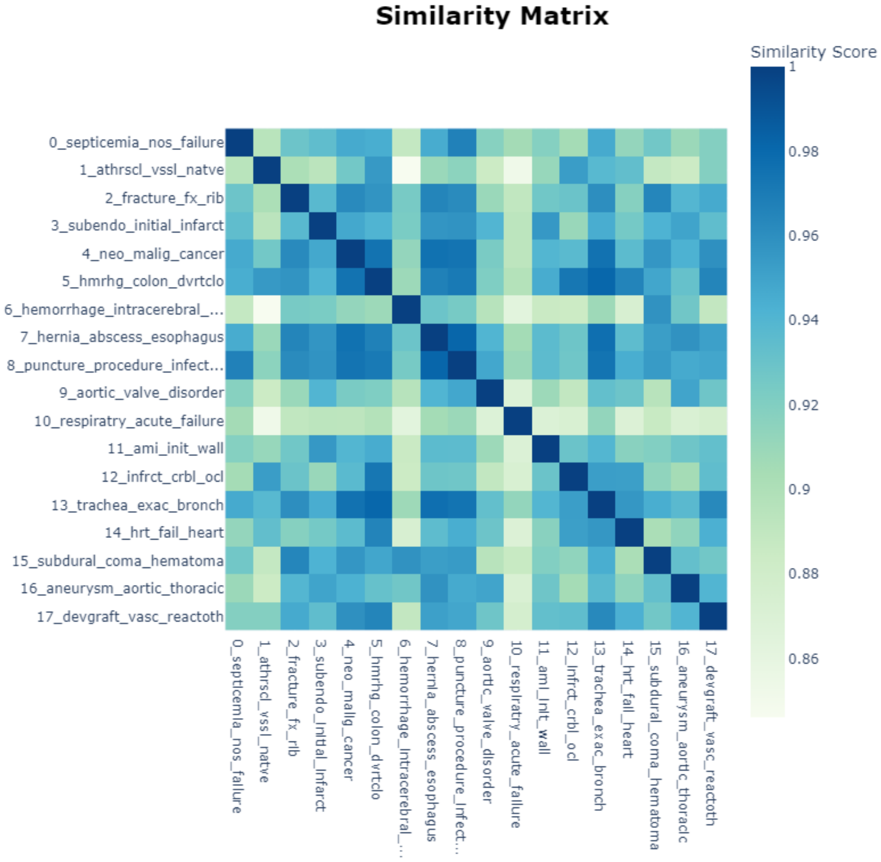


Figure 27:Heatmap :Similarity Matrix of BERTopic-Topics

4.2 Discussion

The research analysis evaluates the strengths and limitations of various computational methods for phenotyping electronic health records (EHR) data. **NMF with K-means** clustering emerged as the most effective method overall with high topic coherence, demonstrating efficacy in identifying distinct and broader disease phenotypes, aiding in patient categorization. Conversely, **BERTopic utilizing ClinicalBERT** embeddings exhibits superior performance in detecting subtle, context-dependent patterns within clinical text, offering nuanced insights into complex medical conditions. This approach is particularly adept at evaluating detailed phenotypic patterns that may not be evident through traditional methods. LDA with k-means, while being a more basic method compared to NMF and BERTopic , still provides wide range of latent topics appropriate for exploratory analysis aimed at identifying a wide range of simpler phenotypes.

CHAPTER 5

CONCLUSION

The set research established a computational phenotyping framework that integrates both structured and unstructured modality data derived from the MIMIC-III dataset to evaluate clinically pertinent disease topics and patient characteristics. The incorporation of ICD codes alongside clinical narratives significantly enhances the accuracy and depth of phenotyping, offering a more precise depiction of patient health complexities and enabling the extraction of more comprehensive phenotypes. The use of NMF alongside K-means clustering effectively identified broad disease phenotypes, while advanced deep learning techniques, such as BERTopic utilizing ClinicalBERT embeddings, outperformed traditional techniques like LDA in recognizing nuanced clinical patterns in EHR data. This validates the hypothesis that advanced models with domain-specific embeddings excel at recognizing complex phenotypes, highlighting their potential for precision medicine and tailored treatment plans for each patient profile.

5.1 Implications for Research and Practice

The research outcomes carry substantial implications both scientific inquiry and medical application. In the realm of research, this study establishes the groundwork for developing advanced phenotyping models that effectively utilize a variety of EHR datasets. Subsequent future studies should aim to integrate additional data sources, such as genetic information **or advanced imaging data, to refine phenotyping** techniques and broaden their applicability across varied patient cohorts.

From a clinical standpoint, generating accurate and relevant phenotypes from the EHRs can significantly improve patient care. Enhanced phenotyping facilitates precise patient classification, enabling healthcare practitioners to customize treatment plans more effectively, predict disease progression risk and mortality chances based on disease phenotypes. Insights derived from advanced methodologies like BERTopic can also contribute to the development of targeted interventions and aid medical decision-making in complex clinical scenarios.

5.2 Proposed Future Research

Future research efforts should focus on expanding **advanced multimodal fusion** techniques to encompass seamless real-time data integration and advanced predictive analytics, building on the findings of this study. It is essential to assess these models in various healthcare environments and among patient populations to ensure their efficacy and broader applicability. Moreover, investigating the interpretability and transparency of computational phenotyping models is vital for clinical adoption, ensuring that healthcare providers can rely on and comprehend the insights generated by these technologies. These measures will further amplify computational phenotyping's capacity to advance personalized medicine and elevate patient health.

REFERENCES

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jindi, D., Naumann, T. and McDermott, M. (2019) 'Publicly available clinical BERT embeddings', arXiv preprint, arXiv:1904.03323.

Angelov, D. (2020) 'Top2Vec: Distributed representations of topics', arXiv preprint, arXiv:2008.09470.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) 'Latent Dirichlet Allocation', Journal of Machine Learning Research, 3, pp. 993–1022.

Bodenreider, O. (2004) 'The Unified Medical Language System (UMLS): Integrating biomedical terminology', Nucleic Acids Research, 32(suppl_1), pp. D267-D270.

Bodenreider, O. (2008) 'Biomedical ontologies in action: role in knowledge management, data integration and decision support', Yearbook of Medical Informatics, 17(01), pp. 67-792E.

Che, Z., Kale, D., Li, W., Bahadori, M.T. and Liu, Y. (2015) 'Deep computational phenotyping', in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 507-516.

Che, Z. and Liu, Y. (2017) 'Deep learning solutions to computational phenotyping in healthcare', in 2017 IEEE International Conference on Data Mining Workshops (ICDMW). New York: IEEE, pp. 1100-1109.

Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F. and Sun, J. (2016) 'Doctor AI: Predicting clinical events via recurrent neural networks', in Machine Learning for Healthcare Conference. PMLR, pp. 301-318.

Devarajan, K. (2008) 'Nonnegative matrix factorization: an analytical and interpretive tool in computational biology', PLoS Computational Biology, 4(7), p. e1000029.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) 'BERT: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint, arXiv:1810.04805.

Gao, J., Li, P., Chen, Z. and Zhang, J. (2020) 'A survey on deep learning for multimodal data fusion', Neural Computation, 32(5), pp. 829-864.

Grootendorst, M. (2020) 'BERTopic: Neural topic modeling with transformers', arXiv preprint, arXiv:2010.02151.

Grootendorst, M. (2022) 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure', arXiv preprint, arXiv:2203.05794.

Henderson, J., Ho, J.C., Kho, A.N., Denny, J.C., Malin, B.A., Sun, J. and Ghosh, J. (2017) 'Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping', in 2017 IEEE International Conference on Healthcare Informatics (ICHI). New York: IEEE, pp. 214–223.

Ho, J.C., Ghosh, J., Steinhubl, S.R., Stewart, W.F., Denny, J.C., Malin, B.A. and Sun, J. (2014) 'Limestone: High-throughput candidate phenotype generation via tensor factorization', Journal of Biomedical Informatics, 52, pp. 199–211.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J. (2023) 'K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data', Information Sciences, 622, pp. 178-210.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', Pattern Recognition Letters, 31(8), pp. 651–666.

Kim, Y., El-Kareh, R., Sun, J., Yu, H. and Jiang, X. (2017) 'Discriminative and distinct phenotyping by constrained tensor factorization', Scientific Reports, 7(1), p. 1114.

Kim, Y., Sun, J., Yu, H. and Jiang, X. (2017) 'Federated tensor factorization for computational phenotyping', in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 887–895.

Koutsomitropoulos, D.A. and Andriopoulos, A.D. (2020) 'Automated MeSH indexing of biomedical literature using contextualized word representations', in Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I. Cham: Springer International Publishing, pp. 343-354.

Lee, D.D. and Seung, H.S. (1999) 'Learning the parts of objects by non-negative matrix factorization', Nature, 401(6755), pp. 788–791.

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

Liu, Y., Ma, X., Zhang, Y. and Guo, H. (2019) 'Multimodal data fusion for medical diagnosis and analysis: A survey', IEEE Reviews in Biomedical Engineering, 13, pp. 135–145.

Liu, Z., Zhang, J., Hou, Y., Zhang, X., Li, G. and Xiang, Y. (2022, October) 'Machine learning for multimodal electronic health records-based research: Challenges and perspectives', in China Health Information Processing Conference. Singapore: Springer Nature Singapore, pp. 135-155.

Lipscomb, C.E. (2000) 'Medical subject headings (MeSH)', Bulletin of the Medical Library Association, 88(3), pp. 265–266.

MacQueen, J. (1967) 'Some methods for classification and analysis of multivariate observations', in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, pp. 281–297.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient estimation of word representations in vector space', arXiv preprint, arXiv:1301.3781.

Mohsen, F., Ali, H., El Hajj, N. and Shah, Z. (2022) 'Artificial intelligence-based methods for fusion of electronic health records and imaging data', Scientific Reports, 12(1), p. 17981.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y. (2011) 'Multimodal deep learning', in Proceedings of the 28th International Conference on Machine Learning (ICML-11). New York: ACM, pp. 689-696.

Oveh, R.O., Adewunmi, M.A. and Aziken, G.O. (2022, November) 'BERTopic Modelling with P53 in Ovarian Cancer', in 2022 5th Information Technology for Education and Development (ITED). New York: IEEE, pp. 1-4.

Richesson, R.L., Sun, J., Pathak, J., Kho, A.N. and Denny, J.C. (2016) 'Clinical phenotyping in selected national networks: Demonstrating the need for high-throughput, portable, and computational methods', Artificial Intelligence in Medicine, 71, pp. 57-61.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', Information Processing & Management, 24(5), pp. 513-523.

Srivastava, A. and Sutton, C. (2017) 'Autoencoding variational inference for topic models', arXiv preprint, arXiv:1703.01488.

Stearns, M.Q., Price, C., Spackman, K.A. and Wang, A.Y. (2001) 'SNOMED clinical terms: Overview of the development process and project status', in Proceedings of the AMIA

Symposium. Bethesda, MD: American Medical Informatics Association, p. 662.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', arXiv preprint, arXiv:1706.03762.

Wu, D., Zhang, Y., Wang, X. and Tan, T. (2014) 'An overview on multimodal data fusion for healthcare', IEEE Journal of Biomedical and Health Informatics, 21(4), pp. 916-926.

Xu, D. and Tian, Y. (2015) 'A comprehensive survey of clustering algorithms', Annals of Data Science, 2(2), pp. 165–193.

Zeng, J., Yu, H. and Pan, Y. (2019) 'A comprehensive literature review on integrating multiple types of clinical data for phenotyping in medical research', Journal of Biomedical Informatics, 98, p. 103287.

Zeng, Z., Deng, Y., Li, X., Naumann, T. and Luo, Y. (2018) 'Natural language processing for EHR-based computational phenotyping', IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(1), pp. 139-153.

Zhang, S., Li, H., Tang, R., Ding, S., Rasmy, L., Zhi, D., Zou, N. and Hu, X. (2023, June) 'PheME: A deep ensemble framework for improving phenotype prediction from multi-modal data', in 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI). New York: IEEE, pp. 268-275.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L. and Buntine, W. (2021) 'Topic modelling meets deep neural networks: A survey', arXiv preprint, arXiv:2103.00498.

APPENDIX

APPENDIX A

Code Details

The Code Link, Technical Appendix and Readme file are attached in the Additional Materials on Blackboard**. (BM-11356488-am.docx)**

APPENDIX B

1. Additional EDA:

ICU Length of Stay (LOS) and ICU Outcomes

Most ICU stays are short under 10 days, with a notable peak at shorter durations (Figure 3.5). Majority ICU patients survive their stay, as reflected by the survival statistics in Figure 3.6, indicating effective resource utility.

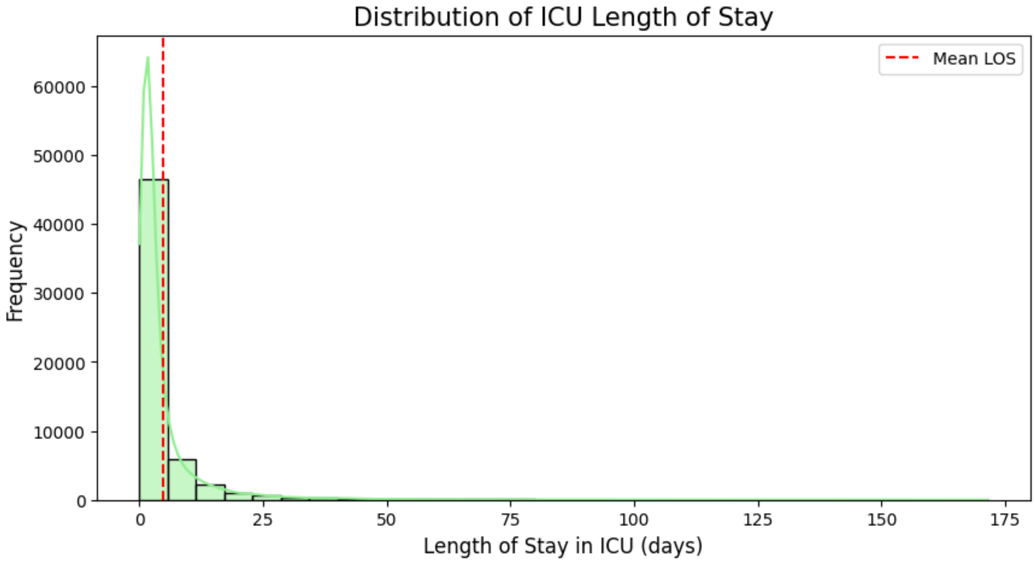


Figure 3.5 Distribution of ICU LOS

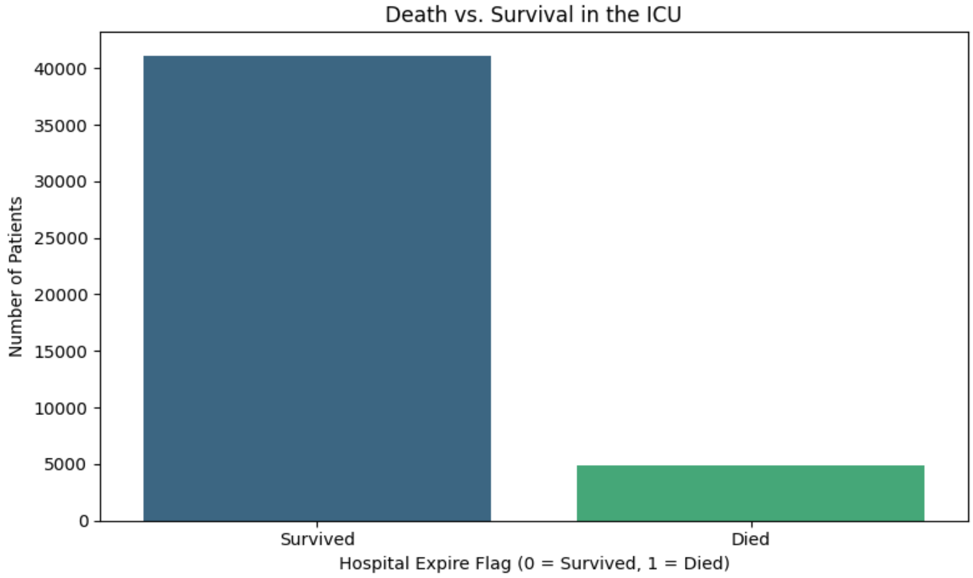


Figure 3.6 Death vs Survival in ICU

1. Admission Types and Patient Ethnicity

Figure 3.10 reflects the dominance of emergency admissions, highlighting the acute nature of the patient population. The dataset is ethnically diverse, with the majority being White, followed by Black/African and Hispanic (Figure 3.11).

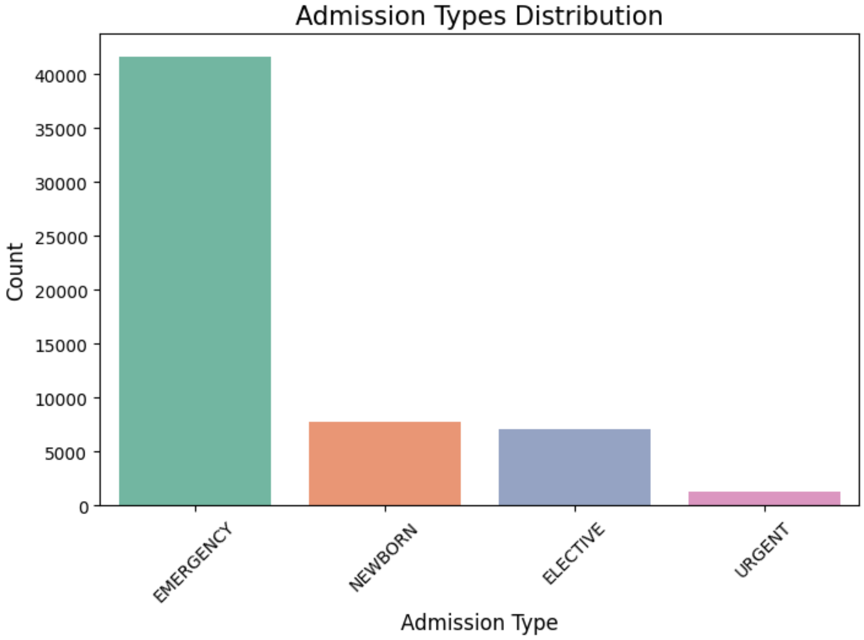


Figure 28: Admission Types Countplot

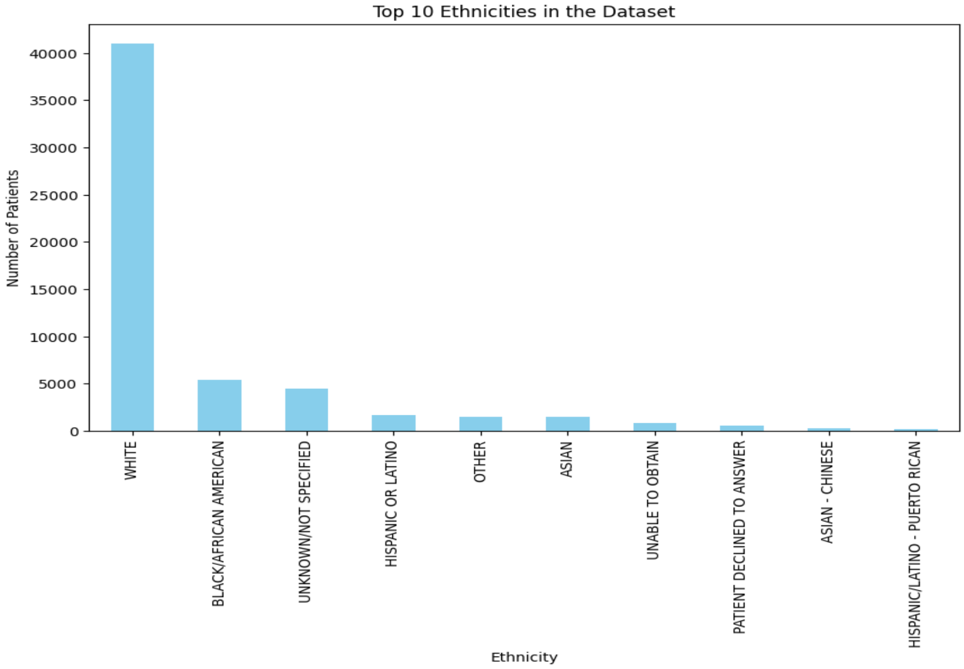


Figure 29: Top 10 Ethnicities in Dataset

Detailed Phenotype Cluster Analysis

C.1. Phenotypes Cluster Analysis for LDA:

1. **Cluster 0**:
    1. **Top Topics**: Heart failure, Respiratory failure, Cancer
    2. **Phenotype**: **Cardiopulmonary and Oncological Disorders**
    3. **Description**: This cluster seems to represent patients primarily dealing with heart failure and respiratory issues, alongside some oncological conditions.
2. **Cluster 1**:
    1. **Top Topics**: Abdominal Pain, Cancer, Vascular Diseases
    2. **Phenotype**: **Abdominal and Vascular Conditions**

    3. **Description**: This cluster is characterized by conditions related to abdominal pain (hernia, laparotomy), cancer, and vascular diseases (aortic aneurysms, valve disorders).

3. **Cluster 2**:
    1. **Top Topics**: Hemorrhage, Respiratory failure, Cancer
    2. **Phenotype**: **Neurological Hemorrhage and Respiratory Disorders**
    3. **Description**: Patients in this cluster likely present with intracranial hemorrhages and respiratory issues, often with some overlap in cancer-related conditions.
4. **Cluster 3**:
    1. **Top Topics**: Respiratory failure, Cancer, Coronary Artery Disease
    2. **Phenotype**: **Respiratory and Cardiovascular Conditions**
    3. **Description**: This cluster indicates a patient group with both respiratory and cardiovascular conditions, and a notable presence of cancer.
5. **Cluster 4**:
    1. **Top Topics**: Coronary Artery Disease, Myocardial Infarction, Heart Failure
    2. **Phenotype**: **Ischemic Heart Disease**
    3. **Description**: Primarily dealing with heart disease, this cluster groups patients with coronary artery disease and myocardial infarctions.
6. **Cluster 5**:
    1. **Top Topics**: Stroke, Respiratory failure, Hemorrhage
    2. **Phenotype**: **Stroke and Respiratory Failure**
    3. **Description**: Focuses on neurological conditions (stroke) and respiratory failures.
7. **Cluster 6**:
    1. **Top Topics**: Liver Disease, Cancer, Respiratory failure
    2. **Phenotype**: **Hepatic and Respiratory Disorders**
    3. **Description**: Patients with liver diseases (cirrhosis, hepatitis), cancer, and respiratory failures.
8. **Cluster 7**:
    1. **Top Topics**: Cancer, Respiratory failure, Abdominal Pain

   2. **Phenotype**: **Oncological and Abdominal Disorders**
   3. **Description**: This group includes patients with cancer, respiratory issues, and abdominal conditions like hernia or abdominal pain.

9. **Cluster 8**:
   1. **Top Topics**: Vascular Disorders, Coronary Artery Disease, Respiratory failure
   2. **Phenotype**: **Cardiovascular Disorders**
   3. **Description**: Focuses on various cardiovascular diseases, including coronary artery disease and valve disorders.
10. **Cluster 9**:
    1. **Top Topics**: Fracture, Hemorrhage, Cancer
    2. **Phenotype**: **Trauma and Haemorrhagic Disorders**
    3. **Description**: Patients experiencing trauma-related fractures and haemorrhages, alongside cancer conditions.
11. **Cluster 10**:
    1. **Top Topics**: Myocardial Infarction, Heart Failure, Coronary Artery Disease
    2. **Phenotype**: **Advanced Cardiac Disease**
    3. **Description**: This cluster represents patients with advanced stages of cardiac diseases, including myocardial infarctions and heart failure.

C.2. Phenotypes Cluster Analysis for NMF:

1. **Cluster 0: Chronic Cardiopulmonary and Renal Disorders**
   - **Phenotype Characteristics:** This cluster predominantly includes patients with chronic cardiopulmonary and renal disorders. Conditions like heart failure, renal failure, and respiratory issues are common.
   - **Top Topics:** The topics mainly revolve around heart and renal failure, respiratory problems, and chronic obstructive pulmonary diseases. Terms like "failure," "heart," "renal," "acute," and "respiratory" are frequently mentioned, reflecting a group of patients with multi-organ

involvement and chronic conditions.

2. **Cluster 1: Cardiovascular and Gastrointestinal Conditions**
   - **Phenotype Characteristics:** This group is characterized by a combination of cardiovascular and gastrointestinal issues. The conditions often include coronary artery disease and various gastrointestinal disorders.
   - **Top Topics:** The topics in this cluster are centered around coronary artery bypass grafting, chest pain, and abdominal conditions. Common terms include "artery," "coronary," "pain," and "chest," indicating a mix of heart-related diseases and gastrointestinal complications.
3. **Cluster 2: Infectious Diseases and Oncological Conditions**
   - **Phenotype Characteristics:** Patients in this cluster generally present with infectious diseases and oncological conditions. It suggests a high prevalence of sepsis, pneumonia, and various cancers.
   - **Top Topics:** Frequent topics involve infections and cancer types, such as septicemia, pneumonia, lung cancer, and brain malignancies. Words like "septicemia," "cancer," "sepsis," and "procedure" highlight the focus on managing both infections and oncological conditions.
4. **Cluster 3: Neurological and Hemorrhagic Conditions**
   - **Phenotype Characteristics:** This cluster comprises patients with neurological and hemorrhagic disorders, including conditions like subdural hematoma, intracerebral hemorrhage, and brain injuries.
   - **Top Topics:** The topics are related to neurological emergencies, such as hematomas, brain injuries, and subarachnoid hemorrhage. Terms like "hematoma," "hemorrhage," "brain," and "craniotomy" indicate a patient group with severe neurological complications.
5. **Cluster 4: Cerebrovascular and Infectious Disorders**
   - **Phenotype Characteristics:** Patients in this cluster have cerebrovascular conditions combined with infectious diseases, suggesting complications like hemorrhages that may be exacerbated by infections.

  - **Top Topics:** The topics include intracerebral hemorrhage, subdural hematomas, and septicemia, emphasizing a mix of severe cerebrovascular events and infections. Common words are "hemorrhage," "subdural," "septicemia," and "intubation."

6. **Cluster 5: Cardiovascular and Pulmonary Disorders**
   - **Phenotype Characteristics:** This cluster focuses on cardiovascular diseases coupled with pulmonary disorders. It includes conditions like aortic valve disease and chronic obstructive pulmonary disease (COPD).
   - **Top Topics:** The topics often mention aortic valve issues, coronary artery disease, and pulmonary diseases. Keywords such as "aortic," "valve," "disease," "pulmonary," and "obstructive" denote patients with combined heart and lung conditions.
7. **Cluster 6: Renal and Infectious Conditions**
   - **Phenotype Characteristics:** This group is primarily made up of patients with renal issues and associated infections, such as urinary tract infections and wound infections.
   - **Top Topics:** Frequent topics cover renal failure, urinary tract infections, and septic conditions. Terms like "infection," "renal," "failure," and "sepsis" indicate a population dealing with both kidney diseases and infection management.
8. **Cluster 7: Cardiac and Gastrointestinal Conditions**
   - **Phenotype Characteristics:** This cluster highlights patients with cardiac conditions alongside gastrointestinal problems, such as myocardial infarction and abdominal pain.
   - **Top Topics:** Topics focus on cardiac catheterization, myocardial infarction, and abdominal surgeries. Words like "infarction," "cardiac," "abdominal," and "pain" suggest a patient group facing both heart and gastrointestinal issues.

C.2. Phenotypes Cluster Analysis for BERTopic:

This is the analysis of the BERTopic clusters. For methodology overview refer BERTOPIC in Section 4.

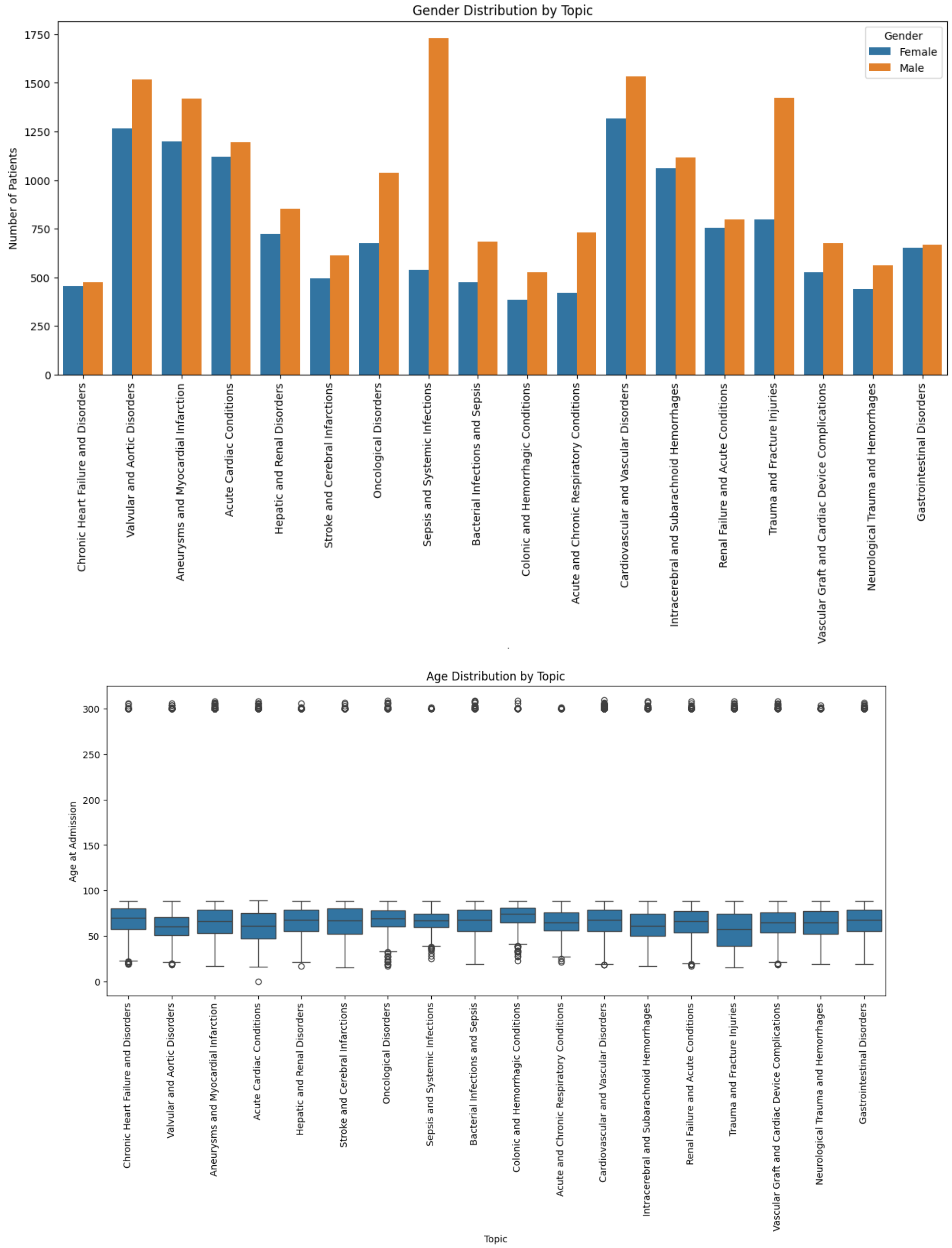


Figure C.1: Gender and Age Distribution by Topic (BERTopic)

□ **Topic 0: Cardiovascular and Vascular Disorders**

- **Prevalence:** 9.25%
- **Characteristics:** This topic covers a range of cardiovascular and vascular disorders, focusing on conditions like septicemia, heart failure, and pneumonia. The high prevalence suggests a significant portion of the patient population is affected by cardiovascular and infectious conditions, making this a critical area for disease management.
- **Implications for Phenotyping:** Identifying patients with these conditions can help develop targeted treatments and management strategies for cardiovascular diseases complicated by infections.

⍰ **Topic 1: Sepsis and Systemic Infections**

- **Prevalence:** 7.35%
- **Characteristics:** This topic centers on sepsis and systemic infections, with keywords like "athrscl," "vssl," and "natve" indicating conditions related to arterial sclerosis and coronary artery diseases.
- **Implications for Phenotyping:** This topic is crucial for identifying patients at high risk of severe infections, which can inform aggressive management strategies to reduce mortality and improve outcomes.

⍰ **Topic 2: Trauma and Fracture Injuries**

- **Prevalence:** 7.21%
- **Characteristics:** Focuses on injuries related to fractures, falls, and bone injuries, highlighting the importance of trauma care in the clinical setting.
- **Implications for Phenotyping:** Recognizing trauma patterns helps in triage and management, ensuring that patients receive appropriate and timely care to minimize complications.

⍰ **Topic 3: Oncological Disorders**

- **Prevalence:** 5.55%
- **Characteristics:** This topic involves various oncological conditions, including cancers of the brain, breast, and spine.
- **Implications for Phenotyping:** Accurate identification of oncological disorders supports personalized cancer treatments and monitoring, improving patient outcomes and resource allocation.

⍰ **Topic 4: Valvular and Aortic Disorders**

- **Prevalence:** 9.02%
- **Characteristics:** Includes conditions related to aortic and valvular diseases, such as stenosis and aortic aneurysms.
- **Implications for Phenotyping:** This topic is essential for managing patients with valvular heart diseases, guiding surgical decisions, and monitoring for complications like aneurysms.

⍰ **Topic 5: Acute Cardiac Conditions**

- **Prevalence:** 7.51%
- **Characteristics:** This topic focuses on acute cardiac conditions, such as myocardial infarction and cardiac arrest.
- **Implications for Phenotyping:** It helps in early detection and management of acute cardiac events, which is crucial for preventing long-term cardiac damage and improving survival rates.

⍰ **Topic 6: Hepatic and Renal Disorders**

- **Prevalence:** 5.10%
- **Characteristics:** Highlights disorders affecting the liver and kidneys, such as renal failure and hepatic conditions.
- **Implications for Phenotyping:** This is vital for patients with comorbid conditions affecting multiple organ systems, aiding in comprehensive care planning.

⍰ **Topic 7: Aneurysms and Myocardial Infarction**

- **Prevalence:** 8.50%
- **Characteristics:** Includes aneurysms and myocardial infarction, indicating a need for focused care on cardiovascular health and surgical interventions.
- **Implications for Phenotyping:** Helps identify patients at risk for aneurysms and myocardial infarctions, facilitating preventive measures and timely surgical interventions.

⍰ **Topic 8: Intracerebral and Subarachnoid Hemorrhages**

- **Prevalence:** 7.05%
- **Characteristics:** Focuses on neurological conditions involving brain

hemorrhages.

- **Implications for Phenotyping:** Identifying these conditions is crucial for managing acute neurological emergencies and improving recovery outcomes.

⍰ **Topic 9: Colonic and Hemorrhagic Conditions**

- **Prevalence:** 2.96%
- **Characteristics:** Covers gastrointestinal bleeding and colonic conditions.
- **Implications for Phenotyping:** It is vital for identifying patients with gastrointestinal bleeding risks, guiding interventions like endoscopy or surgery.

⍰ **Topic 10: Gastrointestinal Disorders**

- **Prevalence:** 4.28%
- **Characteristics:** Includes a range of gastrointestinal issues, focusing on both acute and chronic conditions.
- **Implications for Phenotyping:** Helps in managing gastrointestinal diseases by identifying the specific types and guiding appropriate treatment.

⍰ **Topic 11: Acute and Chronic Respiratory Conditions**

- **Prevalence:** 3.73%
- **Characteristics:** Focuses on respiratory conditions, including acute respiratory failure and chronic pulmonary diseases.
- **Implications for Phenotyping:** It aids in respiratory care planning, ensuring patients receive appropriate ventilation support and monitoring.

⍰ **Topic 12: Chronic Heart Failure and Disorders**

- **Prevalence:** 3.01%
- **Characteristics:** This topic addresses chronic heart failure and related disorders, highlighting the management of long-term heart conditions.
- **Implications for Phenotyping:** Identifying chronic heart failure patients allows for tailored long-term management and monitoring, reducing hospital readmissions.

⍰ **Topic 13: Renal Failure and Acute Conditions**

- **Prevalence:** 5.03%
- **Characteristics:** Focuses on acute renal failure and associated conditions.
- **Implications for Phenotyping:** Helps identify acute renal failure, ensuring rapid

intervention to prevent long-term kidney damage.

⍰ **Topic 14: Bacterial Infections and Sepsis**

- **Prevalence:** 3.74%
- **Characteristics:** Includes bacterial infections leading to sepsis, highlighting the need for infection control and aggressive treatment.
- **Implications for Phenotyping:** Identifying sepsis-prone patients helps in timely administration of antibiotics and supportive care, reducing sepsis-related mortality.

⍰ **Topic 15: Stroke and Cerebral Infarctions**

- **Prevalence:** 3.59%
- **Characteristics:** This topic revolves around stroke and cerebral infarction conditions, focusing on neurological emergencies.
- **Implications for Phenotyping:** Helps in stroke management and rehabilitation, ensuring timely thrombolysis and monitoring for complications.

⍰ **Topic 16: Neurological Trauma and Hemorrhages**

- **Prevalence:** 3.24%
- **Characteristics:** Focuses on trauma-related neurological injuries and hemorrhages, indicating the need for neurocritical care.
- **Implications for Phenotyping:** Identifies patients with neurological trauma, guiding neuro-intensive care and surgical interventions.

⍰ **Topic 17: Vascular Graft and Cardiac Device Complications**

- **Prevalence:** 3.89%
- **Characteristics:** Covers complications related to vascular grafts and cardiac devices, indicating post-surgical and interventional care needs.
- **Implications for Phenotyping:** Helps in managing complications from cardiovascular surgeries, ensuring timely identification and treatment of device-related issues.