MANCHESTER 1824
The University of Manchester

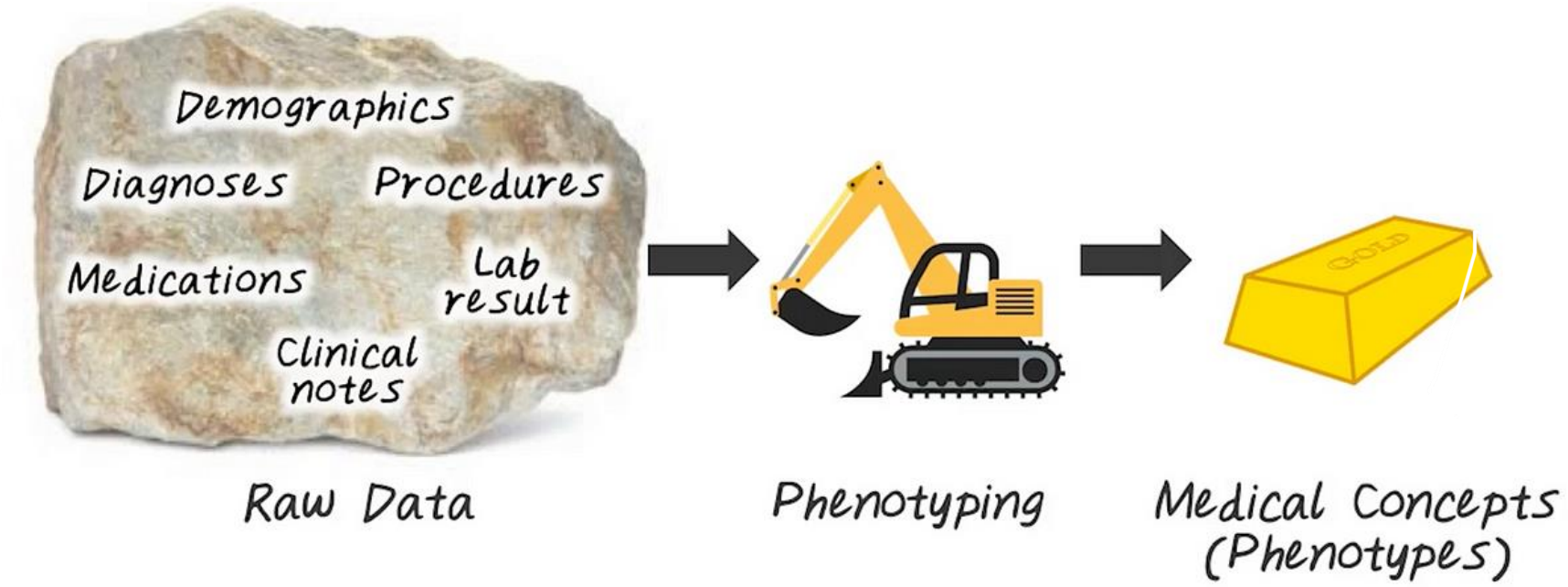
Computational phenotyping patients using multimodal EHRs

Student ID: 11356488

Supervisor: Dr. Xian Yang

Introduction to Computational Phenotyping

Phenotyping is essential for discovering patterns in patient data to improve diagnosis and treatment.

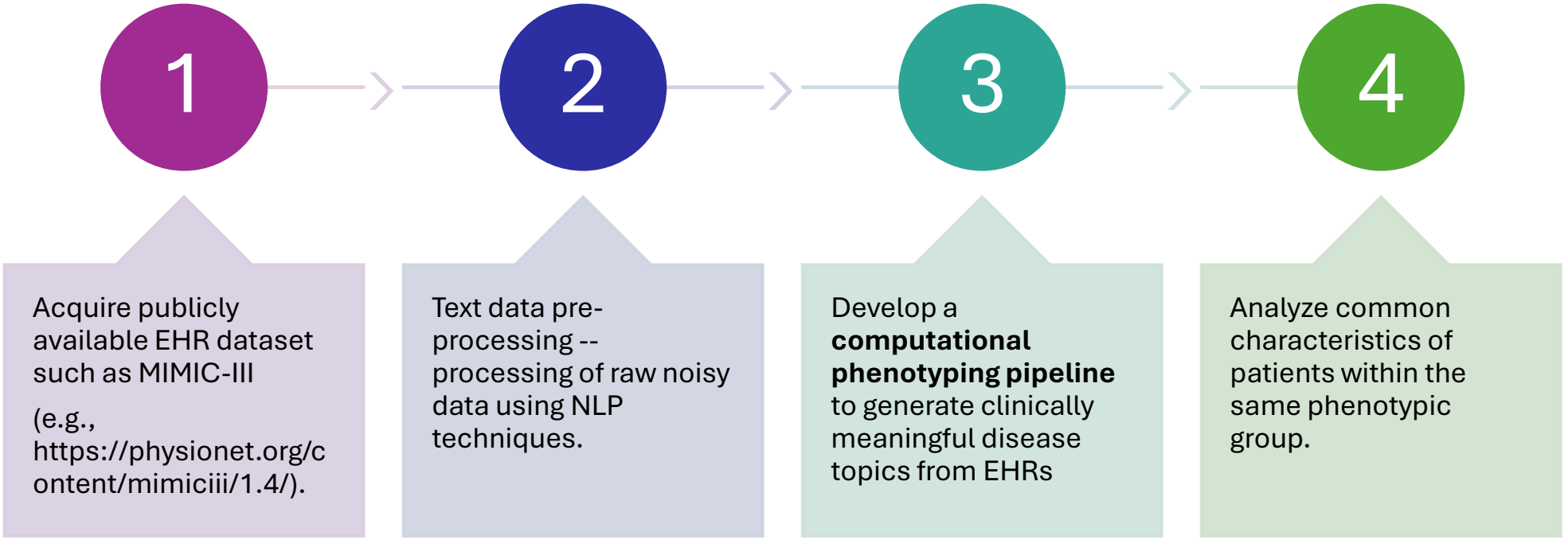


Research Question

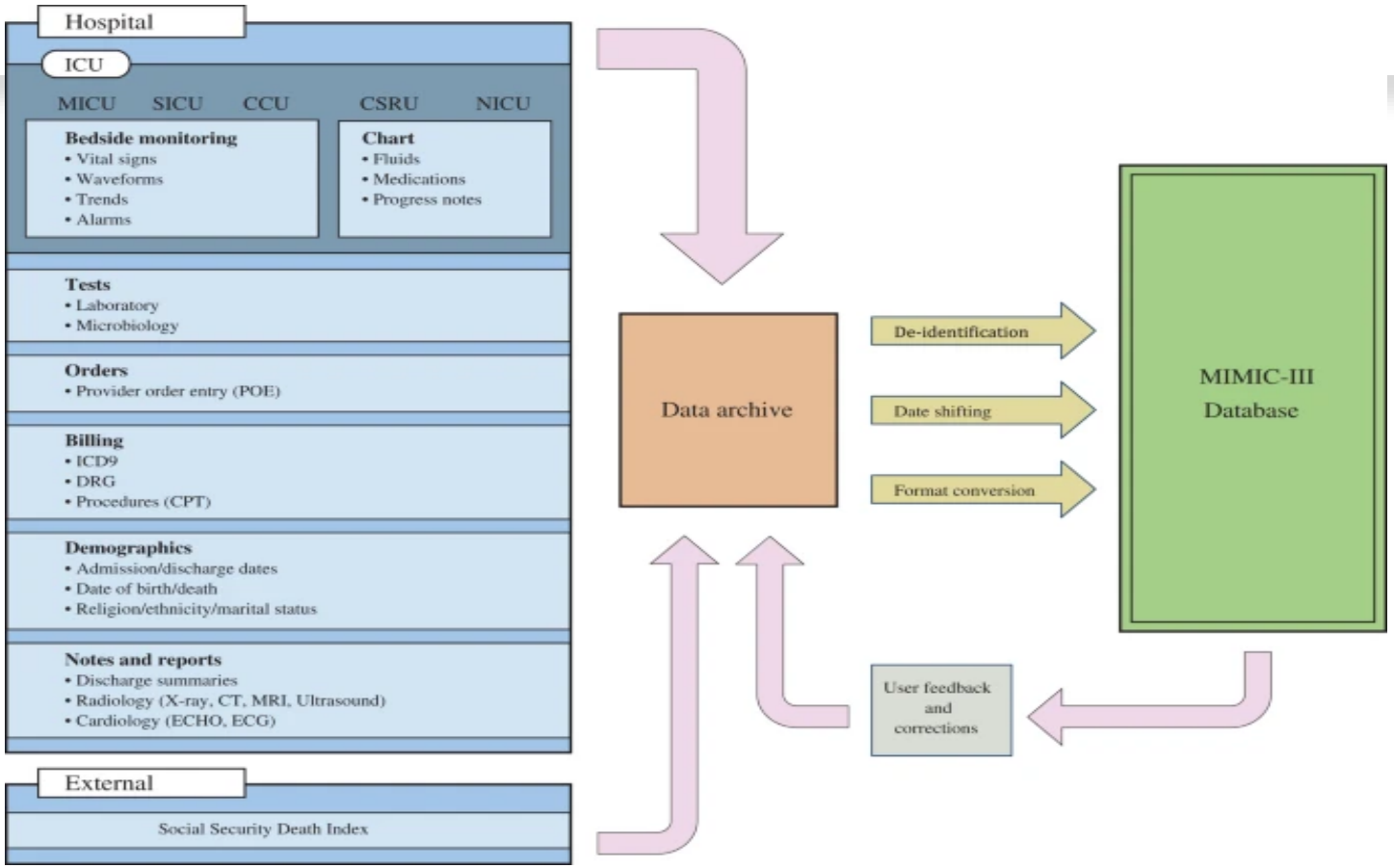
Research Question 1: How can multimodal EHR data be effectively integrated to enhance computational phenotyping?

Research Question 2: Which machine learning and NLP techniques are most effective for identifying disease phenotypes(Clusters) in multimodal EHR data?

Objectives


1
2
3
4
Acquire publicly available EHR dataset such as MIMIC-III
(e.g., https://physionet.org/content/mimiciii/1.4/).
Text data pre-processing -- processing of raw noisy data using NLP techniques.
Develop a **computational phenotyping pipeline** to generate clinically meaningful disease topics from EHRs
Analyze common characteristics of patients within the same phenotypic group.

Understanding the MIMIC-III Dataset



Dataset Overview

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)

Methodology

1

1. Data Acquisition:

Collected the 6 tables including structured and unstructured data from MIMIC-III dataset.

2

2. Data Preprocessing:
Extraction and cleaning of structured and unstructured data (ICD codes, discharge summaries).

3

3.Feature Engineering and Data Fusion:
Conversion of text data into numerical vectors (TF-IDF, ClinicalBERT embeddings).

4

4. Topic Modeling:

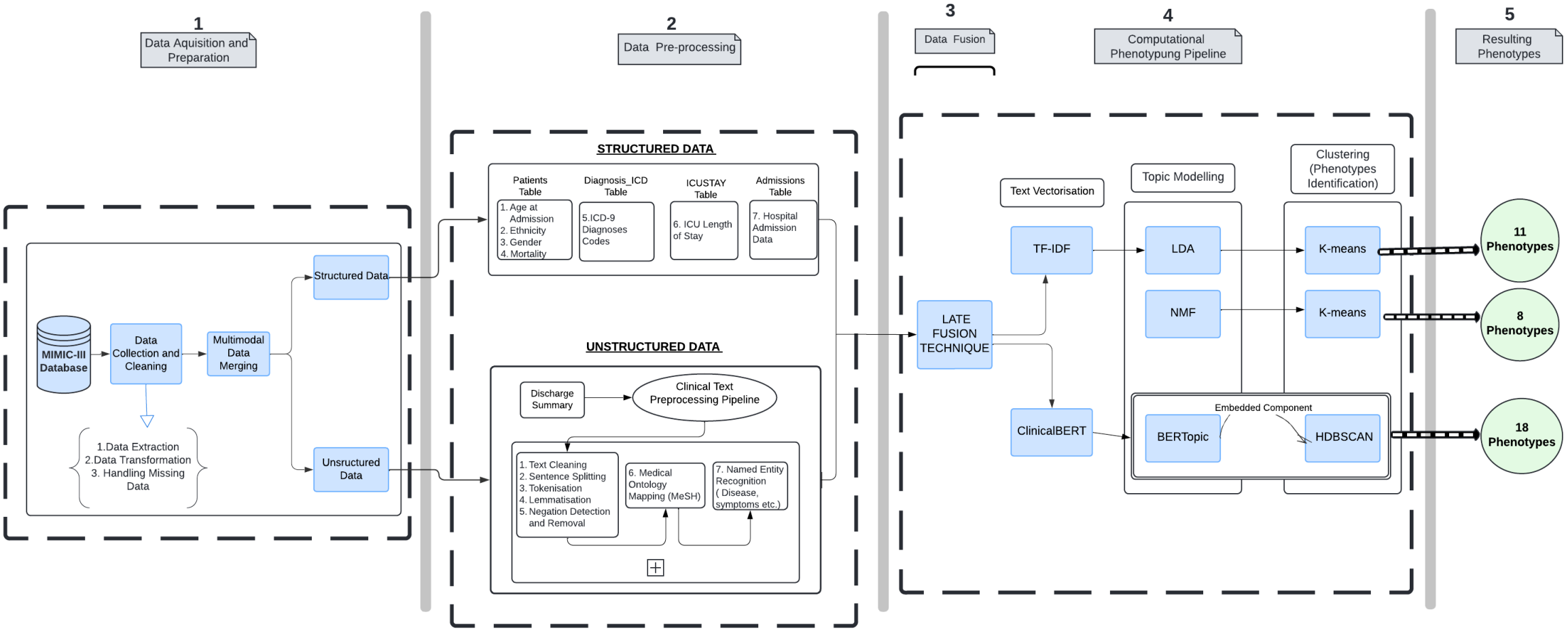
- LDA (Latent Dirichlet Allocation)
- NMF (Non-negative Matrix Factorization)
- BERTopic with ClinicalBERT

5

5. Clustering for Phenotyping:

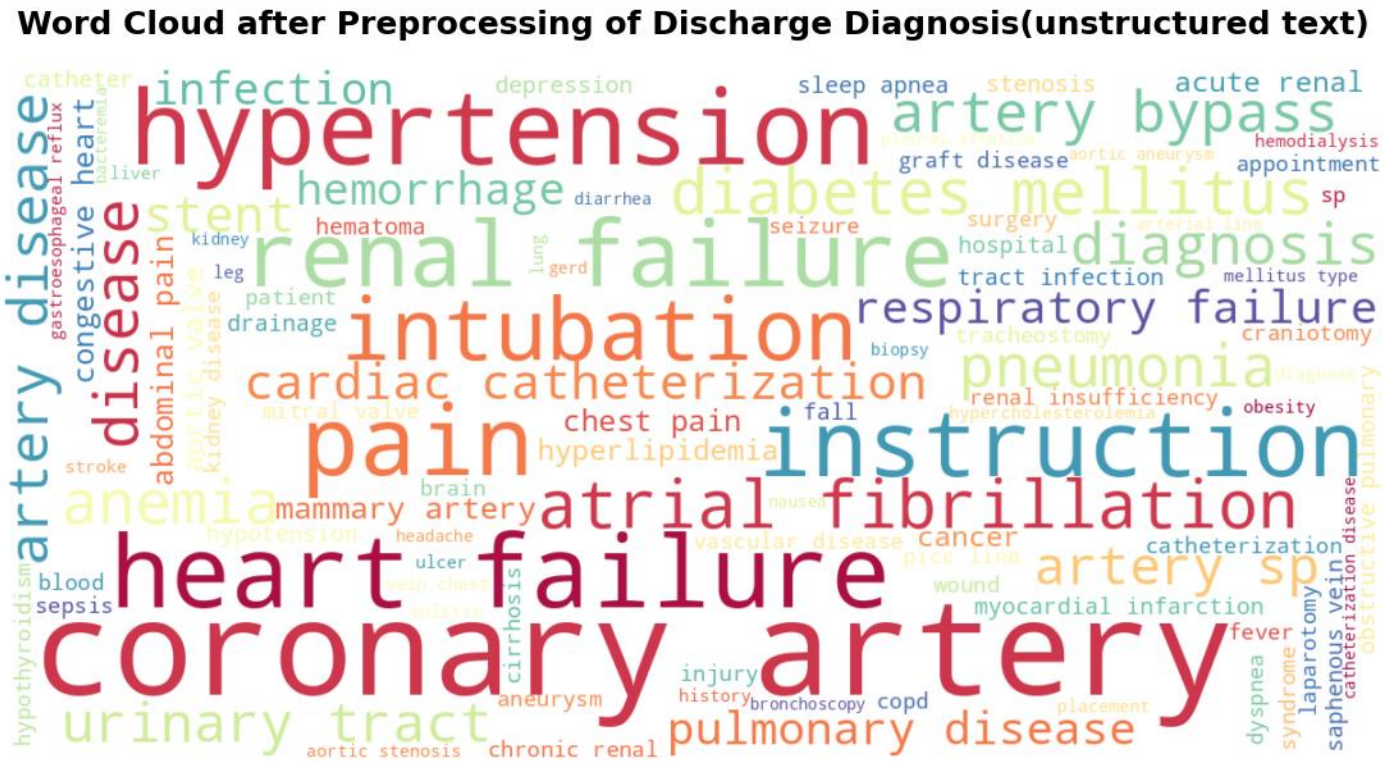
- K-means for LDA and NMF
- HDBSCAN for BERTopic

Methodology Flow

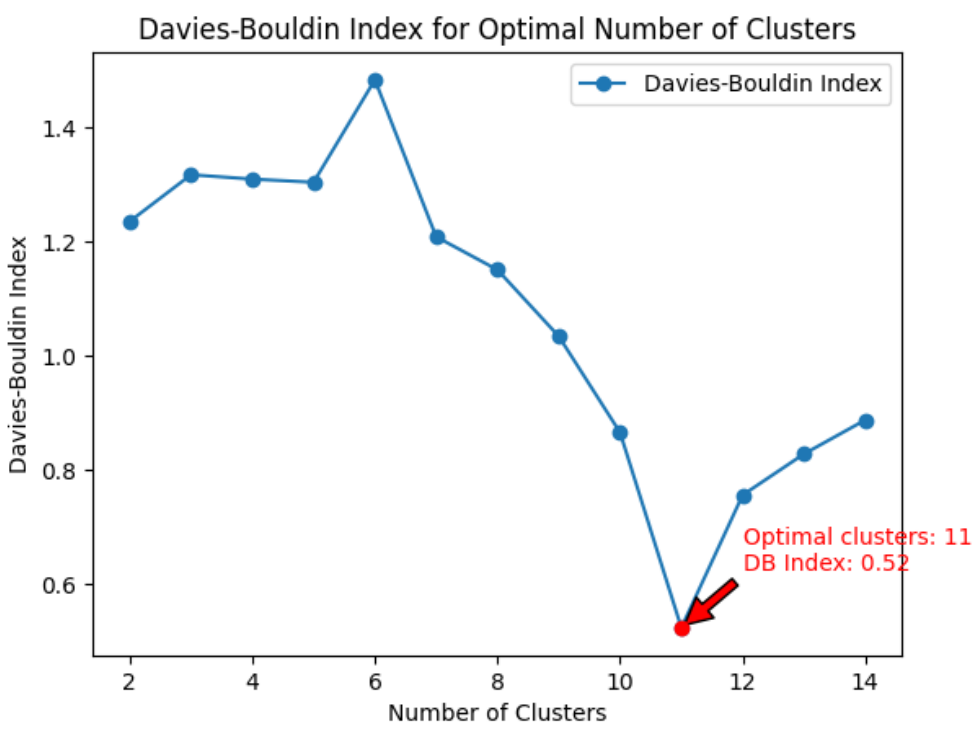


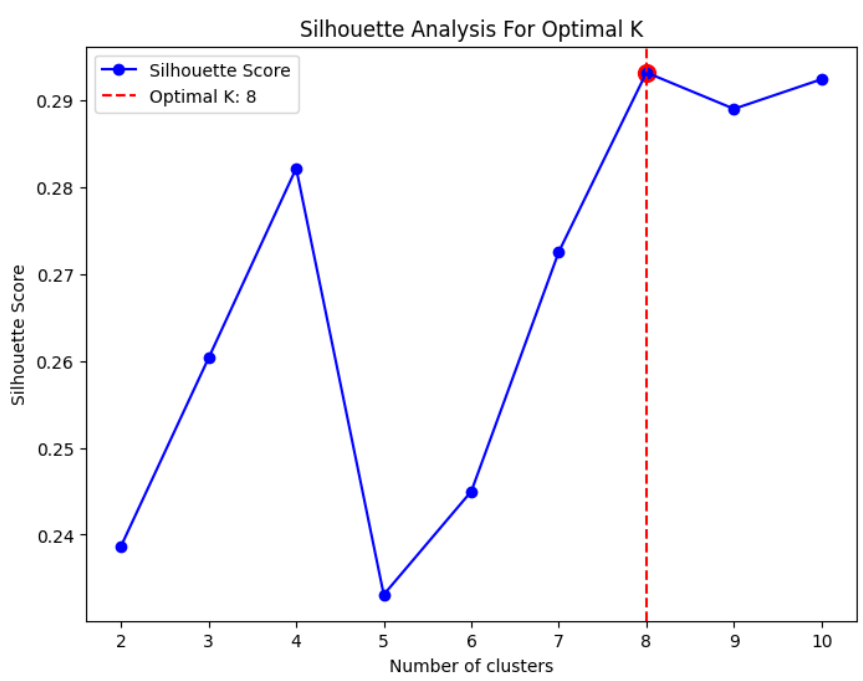
Results and Analysis

Word Cloud : After Text Pre-processing



Optimal Number of Topics & Clusters





LDA

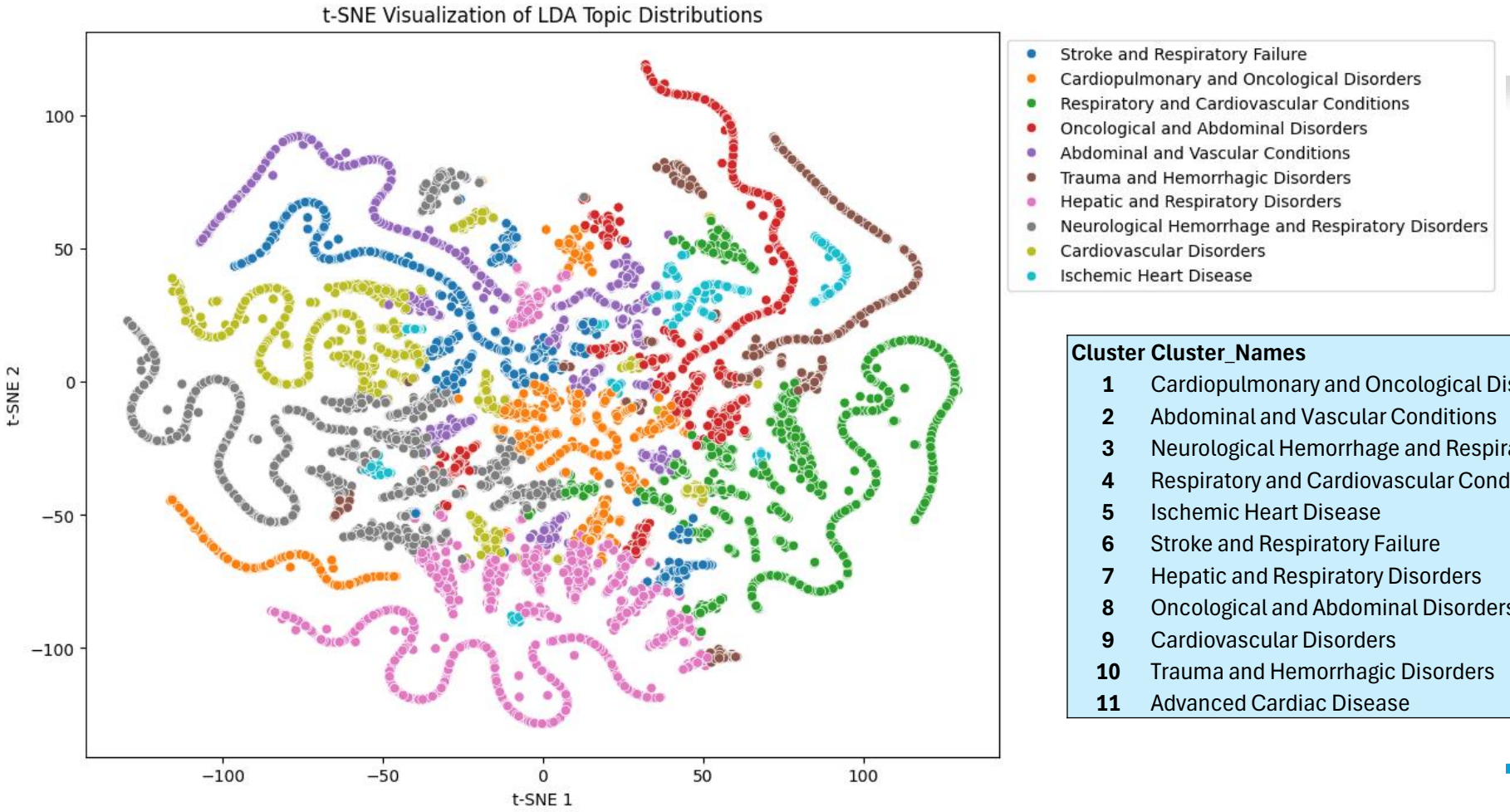
Topic Modelling Outcome [LDA]

- Top words per Topic

	Top words in LDA
Topic #0	liver, cirrhosis, hepatitis, alcohol, procedure, encephalopathy, dev, transplant, alcoholic, oth
Topic #1	catheterization, infarction, initial, infarct, subendo, cardiac, pain, ami, myocardial, chest
Topic #2	hemorrhage, subdural, hem, intracerebral, hematoma, procedure, subarachnoid, ulcer, coma, headache
Topic #3	artery, coronary, disease, natve, vssl, athrscl, crnry, vein, bypass, graft
Topic #4	aortic, valve, mitral, aneurysm, disorder, stenosis, atrial, procedure, heart, fibrillation
Topic #5	cancer, neo, lung, tract, urinary, mal, cell, procedure, infection, brain
Topic #6	failure, pneumonia, septicemia, nos, intubation, acute, procedure, instruction, respiratry, respiratory
Topic #7	heart, failure, disease, renal, kidney, pulmonary, chronic, artery, diabetes, mellitus
Topic #8	stroke, crbl, infrct, ocl, art, artery, emblsm, carotid, cerebral, infrc
Topic #9	fracture, fx, rib, cl, injury, fall, wound, procedure, bone, closed
Topic #10	pain, hernia, abdominal, procedure, laparotomy, colon, dissection, lymph, fistula, colectomy

Clustering for Phenotypes Identification[LDA]

LDA + K-Means Clustering



Cluster	Cluster_Names	Patient Count
1	Cardiopulmonary and Oncological Disorders	3746
2	Abdominal and Vascular Conditions	2543
3	Neurological Hemorrhage and Respiratory Disorders	2524
4	Respiratory and Cardiovascular Conditions	4003
5	Ischemic Heart Disease	3853
6	Stroke and Respiratory Failure	985
7	Hepatic and Respiratory Disorders	2243
8	Oncological and Abdominal Disorders	3959
9	Cardiovascular Disorders	2559
10	Trauma and Hemorrhagic Disorders	2716
11	Advanced Cardiac Disease	1717

NMF

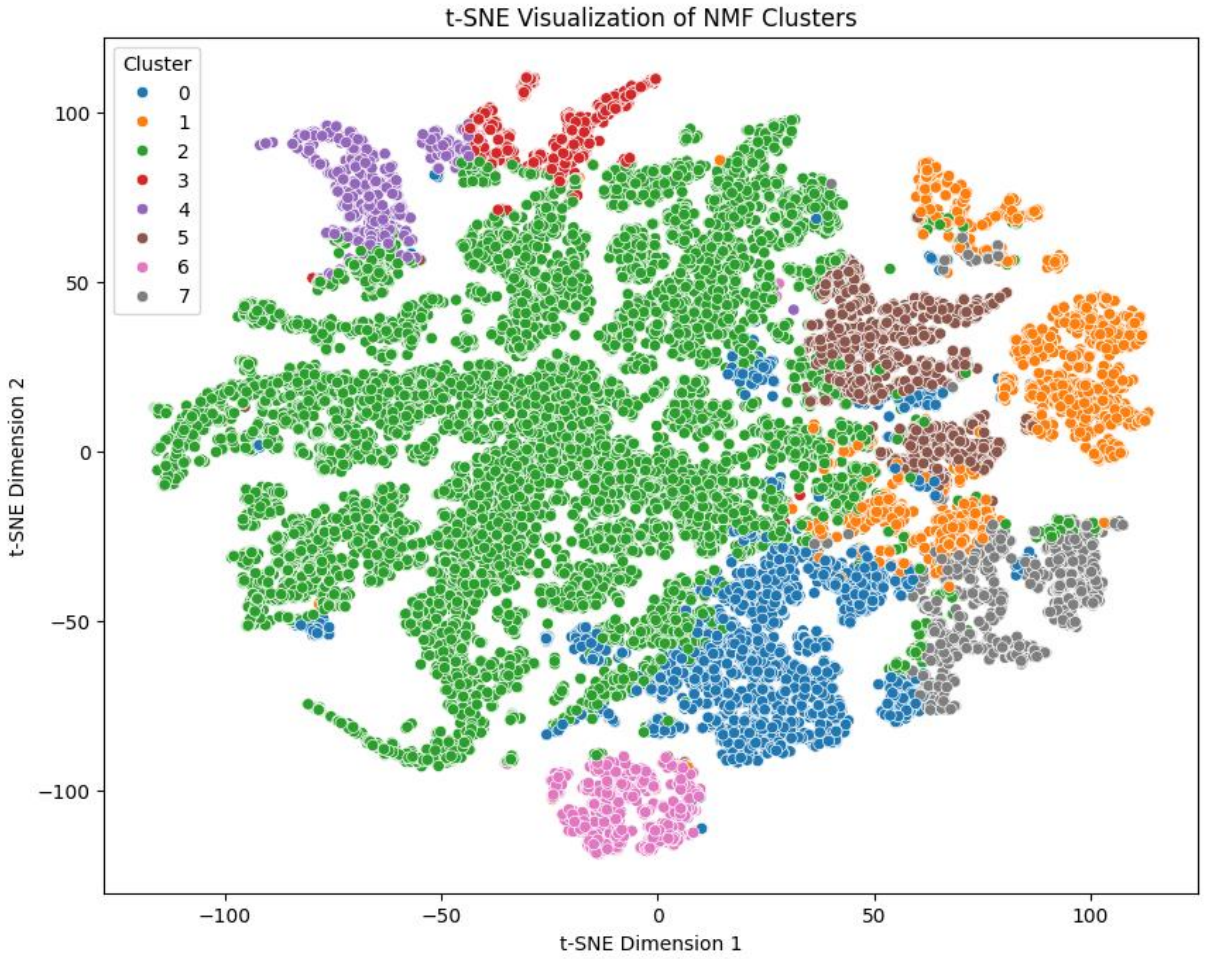
Topic Modelling Outcomes [NMF]

- Top words per Topic

Topic	Top Words
0	artery, coronary, vein, bypass, disease, mammary, graft, saphenous, internal, sp
1	failure, heart, renal, acute, congestive, respiratry, respiratory, hrt, systolic, fail
3	catheterization, infarction, cardiac, initial, infarct, subendo, myocardial, ami, init, stent
4	aortic, valve, stenosis, disorder, mitral, aneurysm, dyspnea, tissue, aorta, atrial
5	hemorrhage, intracerebral, subarachnoid, procedure, angiogram, headache, instruction, aneurysm, intracranial, cerebral
6	disease, pulmonary, obstructive, kidney, chronic, diabetes, mellitus, apnea, sleep, hypertension
7	infection, tract, urinary, wound, secondary, postop, urin, picc, blood, react
8	cancer, lung, neo, mal, brain, cell, sec, malig, procedure, nec
9	pain, chest, abdominal, pancreatitis, procedure, blood, hernia, laparotomy, appointment, acute
10	nos, septicemia, pneumonia, instruction, intubation, procedure, hypotension, sepsis, shock, line
11	subdural, hematoma, coma, hem, fall, craniotomy, fracture, brain, fx, injury
12	crnry, vssl, athrscl, natve, sp, chest, procedure, hyperlipidemia, catheterization, graft

Clustering for Phenotypes Identification[NMF]

NMF + K-Means Clustering

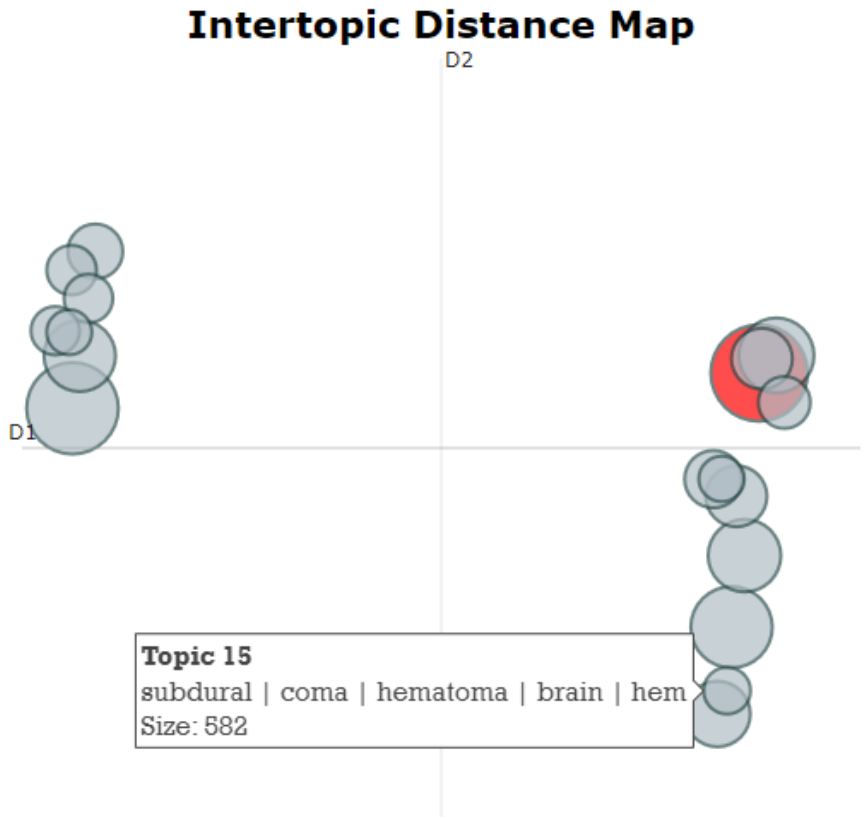


Cluster	Phenotype Name	Number of Patients	Top Topics
0	Chronic Cardiopulmonary and Renal Disorders	3533	[1, 6, 10]
1	Cardiovascular and Gastrointestinal Conditions	2720	[0, 12, 9]
2	Infectious Diseases and Oncological Conditions	18052	[10, 8, 9]
3	Neurological and Hemorrhagic Conditions	801	[11, 5, 10]
4	Cerebrovascular and Infectious Disorders	1017	[5, 11, 10]
5	Cardiovascular and Pulmonary Disorders	2033	[4, 0, 6]
6	Renal and Infectious Conditions	1059	[7, 1, 10]
7	Cardiac and Gastrointestinal Conditions	1633	[3, 9, 0]

BERTopic

Topic Modelling Outcomes [BERTopic]

- Top words per Topic



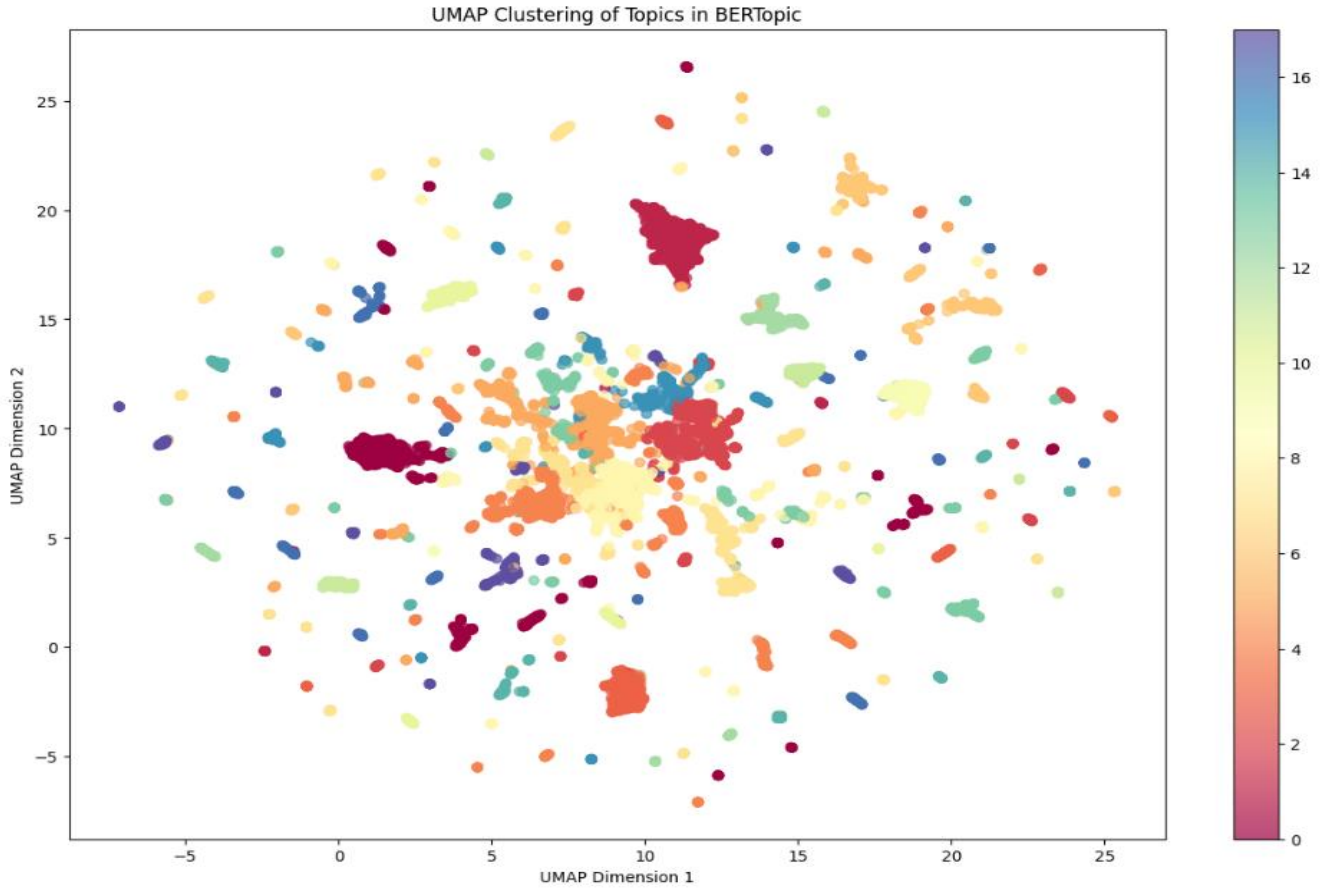
Topic Modelling Outcomes [BERTopic]

- Top words per Topic

Topic	Topic Name	Patients	Top Words
0	Cardiovascular and Vascular Disorders	2493	septicemia (0.1496), nos (0.077), failure (0.0459), sepsis (0.0442), shock (0.0435), pneumonia (0.038)
1	Sepsis and Systemic Infections	2240	athrscl (0.1092), vssl (0.1092), natve (0.1092), crnry (0.1092), artery (0.1089), coronary (0.09)
2	Trauma and Fracture Injuries	1767	fracture (0.1228), fx (0.0883), rib (0.0651), injury (0.053), fall (0.0516), bone (0.0498), cl (0.0447)
3	Oncological Disorders	1503	subendo (0.1212), initial (0.1154), infarct (0.1143), catheterization (0.0722), artery (0.061)
4	Valvular and Aortic Disorders	1408	neo (0.1406), malig (0.094), cancer (0.0857), mal (0.0806), sec (0.0724), brainspine (0.0551)
5	Acute Cardiac Conditions	1361	hmrhg (0.0967), colon (0.0574), dvrtclo (0.0513), procedure (0.0412), nec (0.0382), blood (0.034)
6	Hepatic and Renal Disorders	1150	hemorrhage (0.3367), intracerebral (0.1753), subarachnoid (0.1099), subdural (0.1033)
7	Aneurysms and Myocardial Infarction	998	hernia (0.0976), abscess (0.0597), esophagus (0.0594), obstr (0.0545), pain (0.0533), intestinal (0.048)
8	Intracerebral and Subarachnoid Hemorrhages	991	puncture (0.0395), procedure (0.0386), infection (0.033), tract (0.033), anemia (0.0324)
9	Colonic and Hemorrhagic Conditions	882	aortic (0.1875), valve (0.1655), disorder (0.1232), stenosis (0.1037), coronary (0.0647)
10	Gastrointestinal Disorders	806	respiratry (0.2096), acute (0.1379), failure (0.1363), intubation (0.0711), respiratory (0.0698)
11	Acute and Chronic Respiratory Conditions	720	ami (0.1992), init (0.191), wall (0.1542), catheterization (0.1335), infarction (0.1269)
12	Chronic Heart Failure and Disorders	657	infrct (0.1859), crbl (0.1723), ocl (0.1576), art (0.1467), stroke (0.1277), emblsm (0.1068)
13	Renal Failure and Acute Conditions	632	trachea (0.0695), exac (0.0676), bronch (0.0638), tracheostomy (0.0632), comp (0.0602)
14	Bacterial Infections and Sepsis	626	hrt (0.1521), fail (0.1145), heart (0.1036), on (0.1001), ac (0.0949), chr (0.0915)
15	Stroke and Cerebral Infarctions	582	subdural (0.1669), coma (0.1639), hematoma (0.1302), brain (0.1103), hem (0.0992), hemorrhage (0.0946)
16	Neurological Trauma and Hemorrhages	550	aneurysm (0.2962), aortic (0.1502), thoracic (0.0771), nonrupt (0.069)
17	Vascular Graft and Cardiac Device Complications	542	devgraft (0.1748), vasc (0.1411), reactoth (0.1042), malfunc (0.0537), reactcardiac (0.05)

Clustering for Phenotypes Identification [BERTopic]

BERTopic + K-Means Clustering



Comparative Analysis

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 + HDBSCAN (ClinicalBERT)	Discharge Diagnoses, ICD Descriptions	ClinicalBERT	Hierarchical Clustering	Moderate	Moderate	**Very High**

CONCLUSION

This study shows that integrating multimodal data enhances the quality of phenotyping, and advanced methods like BERTopic are highly effective in capturing complex disease patterns.

Implications for research and clinical practice: These models can be applied to real-world healthcare settings to stratify patients better and improve personalized medicine.

Future Direction: Further research could explore incorporating genetic or lifestyle data to refine the phenotypes further.

Thank You