Assessing Translation Science in Exercise using Text Analytics

A Thesis

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Master of Science (Systems Engineering)

by

Kristin Diane Morrow

May 2023

Approval Sheet

This thesis is submitted in partial fulfillment of the requirements for the degree of

Master of Science (Systems Engineering)

Kristin Diane Morrow

This thesis has been read and approved by the Examining Committee:

Don Brown, Advisor

Afsaneh Doryab, Committee Chair

Rafael Alvarado, SDS

Dan Cooper, MD, UC, Irvine

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, Dean, School of Engineering and Applied Science

May 2023

Abstract

A critical challenge facing biomedical investigators and frontline clinicians is the significant delay between clinical research and its adoption in real-world medical settings. Through an investigation of the gap between clinical research related to exercise and its appearance in electronic health records (EHRs), this study aims to connect advances in research to EHRs and back to medical decision-making. Current medical knowledge states that cardiorespiratory fitness (CRF) is an essential component of healthy lifestyles and can play a beneficial role in a variety of specific disease states and conditions. Consequently, care providers should recommend regular exercise whenever appropriate in-patient interactions. However, our findings at the University of California, Irvine Medical Center show that the discussion of established guidelines regarding CRF recommendations to patients rarely occurs in the context of EHR in the emergency medicine department. For coronary artery disease, the most common type of heart disease, only 0.34% of EHRs mention exercise-related treatment plans. Nonetheless, emergency department visits could provide a valuable opportunity for physicians to influence patient well-being through health messaging. While the time between patients and physicians in this setting will not change, direct querying from PubMed, patient charts, and social determinants can be leveraged to find important information and deliver targeted messages. In this way, the emergency department can serve as a powerful tool for physicians to effect positive change in patients' lives, despite the challenges associated with emergency care. This study presents a robust, self-supervised framework that quantifies the relationship between clinical research and EHRs through semantic similarity analysis of mentions within EHRs, thereby enhancing our understanding of the connection between the two. We release the code and PubMed article data to facilitate further research. [1]

[1] https://github.com/kristinmorrow/CT-Abstract-Work.git

Acknowledgements

I would like to express my gratitude to my thesis advisor, Dr. Don Brown, for his support and guidance throughout my research. His guidance and expertise have been invaluable in shaping my ideas and refining my work.

I am also grateful to the members of my thesis committee, Dr. Dan Cooper, MD, Dr. Afsaneh Doryab, and Dr. Rafael Alvarado, for their insightful feedback and constructive criticism.

I am indebted to my research group, especially Debajyoti Datta, for his friendship, collaboration, and technical assistance. His mentorship was crucial in shaping my research questions and methodologies and developing my technical skills.

Finally, I would like to express my heartfelt appreciation to my family and friends, who provided unwavering support, encouragement, and love throughout my academic journey.

Thank you all.

Contents

List of Figures

Chapter 1

Introduction

With the recent and ongoing COVID‘19 pandemic, health has rightfully come to the forefront of many conversations. This newfound emphasis on getting and staying healthy has manifested throughout society in habits such as social distancing, washing hands, or adopting an active lifestyle [6]. On top of that, individuals who are more physically active and demonstrate optimal levels of cardio-respiratory fitness (CRF) experience less severe effects of COVID‘19 and many other medical conditions. [26]. However, healthcare providers rarely mention physical activity in medical records, even though increased exercise can ameliorate many of these conditions [15].

Figure 1.1: UCI Emergency Department

Moreover, a major challenge in the medical field is the gap between research discovery and its implementation in clinical practice, often resulting in delays that can sometimes take decades[16]. Despite the critical nature of this problem, few studies have attempted to quantify this gap [9]. The recent advancements in artificial intelligence have facilitated the analysis of large datasets like hospital clinician notes, providing novel venues to quantify the magnitude of dissonance between discovery and practice. Many researchers are using AI applications in medicine; for instance, at the University of Virginia, AI is used to identify gene markers for celiac disease and to find new ways to treat Crohn's disease [25].

Our research objective was to explore the delayed knowledge flow by connecting findings from clinical research articles around CRF to electronic health records (EHRs) and back to medical decision-making. Our objectives include identifying what is being recommended from clinical trial research findings, whether or not these recommendations find their way into EHR data, and if so, at what quantity. So far, previous research has focused on analyzing citation rates of PubMed articles to predict whether or not a paper will make a significant medical contribution in the future [9]. It is worth noting that PubMed is a public, online database of biomedical and life sciences literature with over 32 million citations and abstracts from journals, books, and online resources, making it a valuable tool for both researchers and healthcare professionals [21]. Our goal is to expand on previous work by creating a framework to connect clinical research to a large sample of EHR data using Natural Language Processing tools like topic modeling, word embedding, and semantic similarity search.

As far as we are aware, no other attempts have been made to quantify the gap between "bench to bedside" progress, thereby making our research a novel addition to the healthcare industry. A study of this kind has been hard to come by for a few reasons. To start, accessing a sizeable EHR dataset is difficult due to strict HIPAA guidelines and the need for de-identification of any EHR-based data. Further, being able to accurately identify, remove, or disguise patient identifiers on large amounts of data is known to be an arduous task. Lastly, the nature and composition of EHR data are inconsistent, especially in the presence of unstructured data. Fortunately for our research, UCI has implemented processes to gather large amounts of de-identified EHR data efficiently. That, coupled with advances in artificial intelligence and natural language processing, allowed us to successfully clean, parse, and employ the large amounts of de-identified data to explore for our research purposes.

1.1 Related Work

To investigate related work, we created a search query across PubMed and Google Scholar to capture journal titles where 'NLP' or 'text analysis' and 'physical activity' or 'lifestyle' were mentioned. Based on the query there were a total of 19 results, plus the addition of a Kaiser study sent over by the UCI team. Only 3 of the articles were relevant to the research question and methodology. One was on prescribing physical activity for healthy aging through periodic VO2 max testing [13], one was on linking physical activity reports to the classification of functioning, disability, and health to use for medical text analytics [18], and the last was the Kaiser study on establishing an exercise vital sign as a standard of care [27]. Given the lack of research articles on the topic of physical activity mentions and the underutilization of NLP methods in electronic healthcare, additional research seems beneficial to strengthen our understanding of these topics and pave the way for future developments in the healthcare field.

One research article of note used the amount of citations by published clinical articles to measure translational progress in healthcare [9]. Their hope was that if they identified the characteristics of clinical articles that make a significant impact, they could identify other clinical articles with those characteristics faster than natural research progression in healthcare. The study used multiple machine learning models with 22 features from the clinical articles and a binary output for whether the paper was cited. This article was the closest to our research question, yet still significantly different. Its goal was to predict which articles would have an impact in the future, but ours is to see if past work is even making its way into current medical practice.

Kaiser's study presents a compelling argument for incorporating a physical activity prescription at every patient visit as a medical standard of care. Blood pressure, heart rate, and other metrics are used to inform how likely future medical conditions might be for a patient as well as a patient's current state of health [27]. These metrics are also a sign of physical fitness levels. Thus, using physical activity as a vital sign would not only allow physicians to monitor those metrics easily but also ensure that each patient's physical activity is assessed and would open the door for counseling on exercise standards at each visit. The results of our study build on Kaiser's work as it quantifies the amount of exercise currently being mentioned in EHR treatment plans.

Lastly, we checked for any research articles that proposed a similar framework to ours by searching for articles that used cluster-based learning with unseen data types. The most similar approach was for cluster-based zero-shot learning for multivariate data, which uses classification and predicts whether the unseen data belongs to a certain cluster or not [7]. Although similar to our framework, ours did not make use of classification. Instead, we used unseen data types and clustered to pull out interventions for each disease classification. No further model training was done on the unseen data types.

1.2 Data

1.2.1 PubMed Abstracts

A preliminary search was done on PubMed journal abstracts to capture research articles that discuss physical fitness and disease types, whether physical fitness levels affect medical decision-making, and whether there are any findings or correlations when both are mentioned [21]. We created the corpus with the following

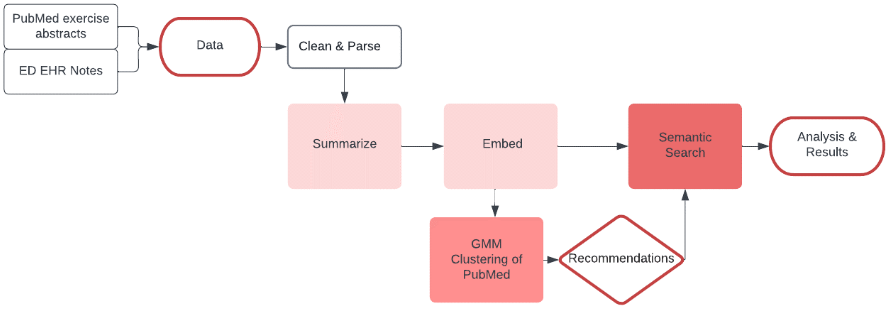


Figure 1.2: Approach

search query: (((“physical fitness”) OR (“physical activity”) OR (fitness) OR (exercise)) AND (health) AND (((diagnosis) OR (diagnose))) throughout the years from 2002 to 2022. The query returned around 109,000 abstracts. However, only abstracts containing research results were targeted for our research question, so the dataset was further filtered to contain Clinical Trial article types exclusively. By doing this, we ensured that we only analyzed medical abstracts containing findings from clinical trials, thus maximizing the accuracy and applicability of our results. This filtering reduced the number of abstracts to 20,311. The most occurring journals in the filtered dataset were the Journal of Strength and Conditioning Research with 564 [11]; Medicine and Science in Sports and Exercise with 432 [10]; and the European Journal of Applied Physics with 381 [28]. The top occurring affiliations were the School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, UK, with 12; the Faculty of Physical Education and Recreation, University of Alberta, Edmonton, Alberta, Canada, with 8; and the Department of Sport, Exercise and Health, University of Basel, Basel, Switzerland also with 8.

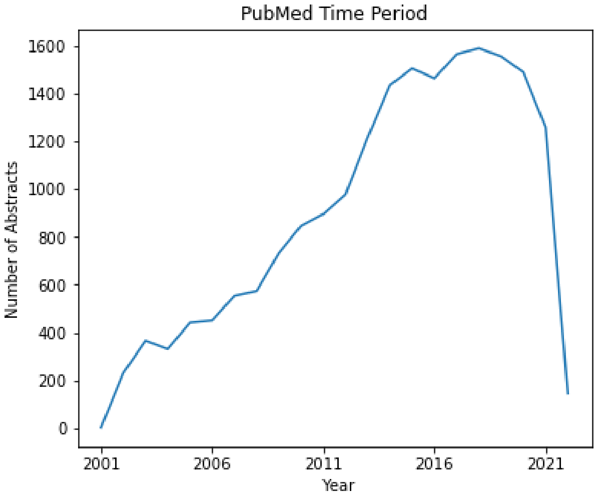


Figure 1.3: Volume of PubMed Clinical Trial Data by Year

Abstracts pulled from PubMed varied in length ranging from a minimum length of 98 words to a maximum length of 62,290 words. We decided to keep only the results sections of the clinical trial abstracts to optimize computational expense and maintain a more consistent length. By doing so, we ensured the corpus contained the key findings and recommendations to the medical community. To remove all other sections, we parsed the abstracts by common abstract formats found on PubMed. A majority of results sections had headings titled with 'Results,' 'Conclusions,' and 'Discussion.' We used regular expressions to locate these headings and then to strip everything that came before them. After parsing, the dataset included 15,639 results sections from the PubMed clinical trial abstracts.

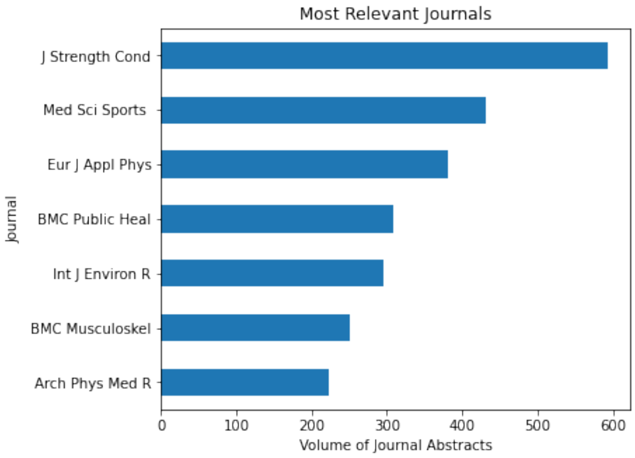


Figure 1.4: Most Occurring Journals

1.2.2 EHR Data

We acquired electronic health record data from the University of California, Irvine Medical Center. The data was comprised solely of UCI's emergency department, which according to the UCI medical staff, includes a diverse range of visit types and provides the most comprehensive representation of patient experiences. Around 50% to 60% of these visits are non-emergencies. Moreover, these visits tend to be the only time most patients even meet with a medical professional. Additionally, the UCI emergency department is the only level 1 trauma center in the area, is open 24/7, employs 23 different emergency specialists, and has a walk-in clinic for patients without appointments. They are also a designated receiving center for cardiac patients[29].

UCI gets its electronic health records in real-time from Epic, a leading medical software. Clinical data moves via ETL (extract, transform, and load) from Epic to Clarity, a relational database with a normalized data model, on a nightly basis. Data is loaded into the Observational Medical Outcomes Partnership (OMOP) from the relational database regularly downstream. UCI can then query the tables in these databases using

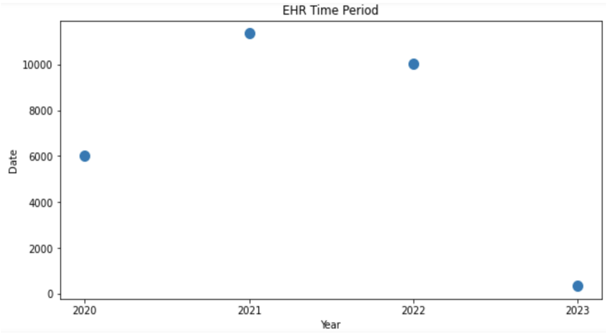


Figure 1.5: Volume of Total EHR Data in Years

SQL and generate de-identified output for research purposes. The de-identified data is created using a formulaic alias specific to each patient, which provides a consistent, de-identified patient ID across reporting. From there, researchers use a self-serving database to query data extracted from Clarity without patient identifiers.

We filtered health records for exercise keywords and for records between February 2019 and February 2022. This query resulted in 24,198 unique health records. Other metadata included were the date of patient visit and ICD-10 codes. ICD-10 is a system for classifying diagnoses, symptoms, and procedures for the processing of insurance claims. Similarly to the PubMed abstracts, a majority of the health records contained phrasing around treatment plans, therapy recommendations, or patient instructions. We again utilized regular expressions to locate the physician's instructions and to strip everything that came before. This process left us with a dataset of 13,720 electronic health records.

Chapter 2

Methods

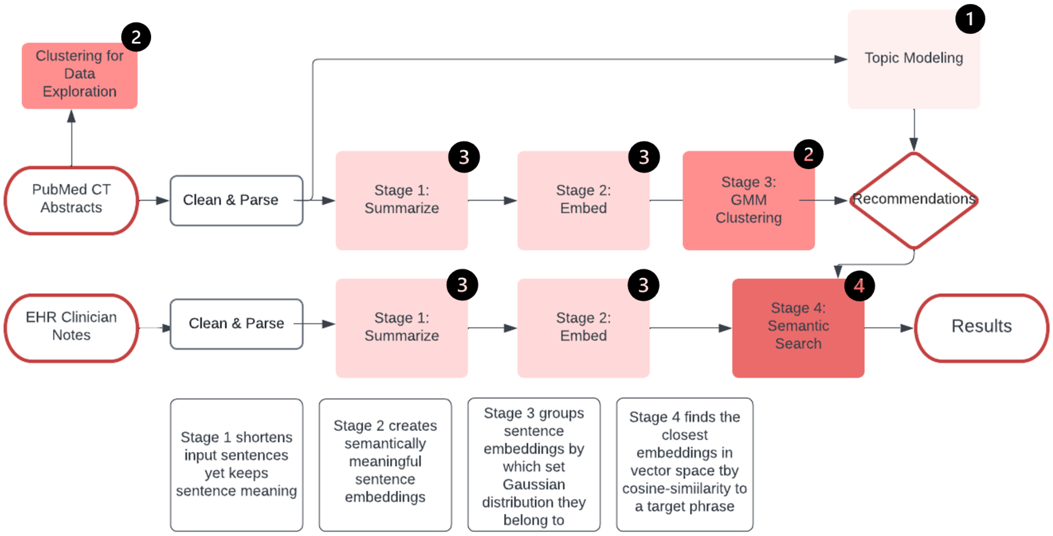

Clustering for Data Exploration
2
Topic Modeling
1
PubMed CT Abstracts
Clean & Parse
Stage 1: Summarize
3
Stage 2: Embed
3
Stage 3: GMM Clustering
2
Recommendations
EHR Clinician Notes
Clean & Parse
Stage 1: Summarize
3
Stage 2: Embed
3
Stage 4: Semantic Search
4
Results
Stage 1 shortens input sentences yet keeps sentence meaning
Stage 2 creates semantically meaningful sentence embeddings
Stage 3 groups sentence embeddings by which set Gaussian distribution they belong to
Stage 4 finds the closest embeddings in vector space tby cosine-simiilarity to a target phrase

Figure 2.1: Approach

Our primary goal in this study is to leverage self-supervised learning to improve query retrieval without training on PubMed or EHR data. Self-supervised learning is a machine learning technique that allows a model to process data and perform a task without labels and without having seen the data type before. The motivations for using self-supervised learning were:

- In healthcare research, numerous areas could greatly benefit from direct query-based approaches, mainly since pre-training domain-specific data is difficult to gather and acquire. Therefore, removing the pre-training step altogether allows researchers to work with the data they do have and still obtain meaningful results.

- EHRs contain protected health information (PHI) that risk violating HIPAA regulations in upstream or downstream tasks. By simplifying the process and minimizing upstream taks, HIPAA violations are less likely to occur.

Figure 2.1 walks through the entire methodology of our approach.

2.1 Data Exploration

2.1.1 Topic Modeling

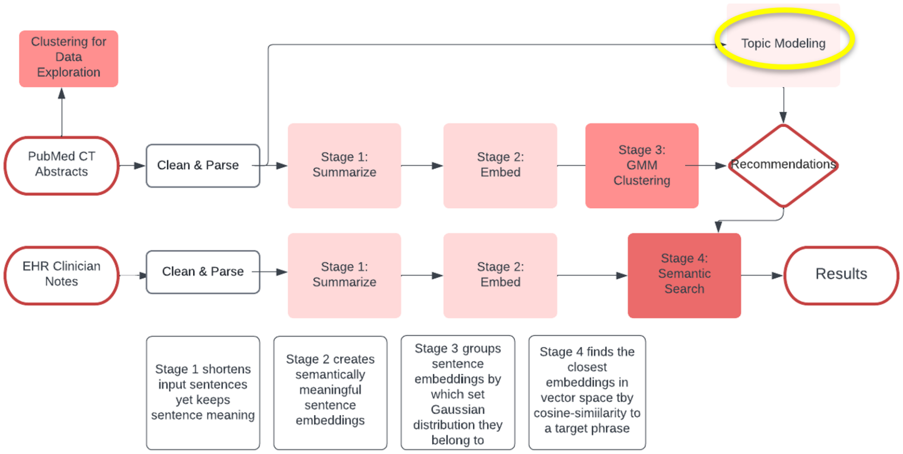

Clustering for Data Exploration
Topic Modeling
PubMed CT Abstracts
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 3: GMM Clustering
Recommendations
EHR Clinician Notes
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 4: Semantic Search
Results
Stage 1 shortens input sentences yet keeps sentence meaning
Stage 2 creates semantically meaningful sentence embeddings
Stage 3 groups sentence embeddings by which set Gaussian distribution they belong to
Stage 4 finds the closest embeddings in vector space tby cosine-simiilarity to a target phrase

Figure 2.2: Topic Modeling

To explore topics and themes appearing in the PubMed clinical trial dataset, we started by using topic modeling. Topic modeling looks at the whole collection of documents to discover underlying themes, which helps to discover main themes in the overall corpus. We also used it to select the main disease types we were going to focus on throughout our research. We began by choosing to use scikit-learn's Latent Dirichlet Allocation (LDA) topic modeling to identify and explore the underlying topics that occur throughout the clinical trial abstracts. LDA is an unsupervised machine learning algorithm that uses a probabilistic model to build on Dirichlet distributions. LDA allows for multiple topics in a document, and each topic is considered as a probability distribution over the set of words [2]. The topics are derived from the co-occurrence of words in the documents. We explored the top 15 topics represented by each topic's ten most occurring words. Three different disease types appeared: heart disease in topic 3, diabetes in topic 6, and mental health in topic 8. We subsequently used these three disease types moving forward throughout PubMed analysis due to their prevalence, and heart disease was used even further with UCI's electronic health records.

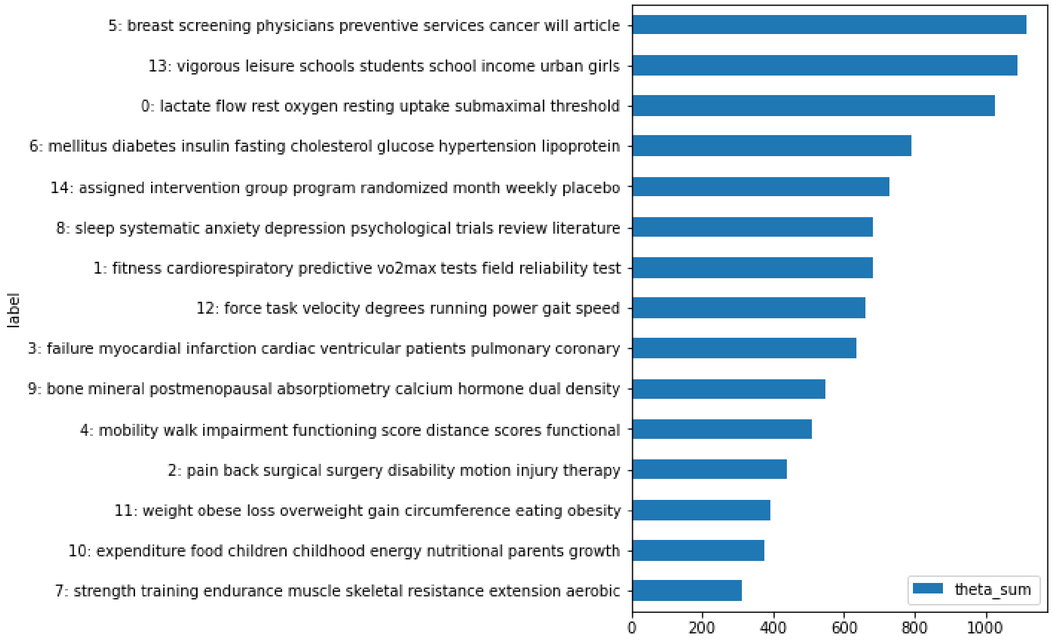


Figure 2.3: Top 15 Topics in PubMed Clinical Trial Data: The Y axis shows the words that make up each topic, while the X axis represents how prevalent that topic is to the overall corpus

2.1.2 Clustering

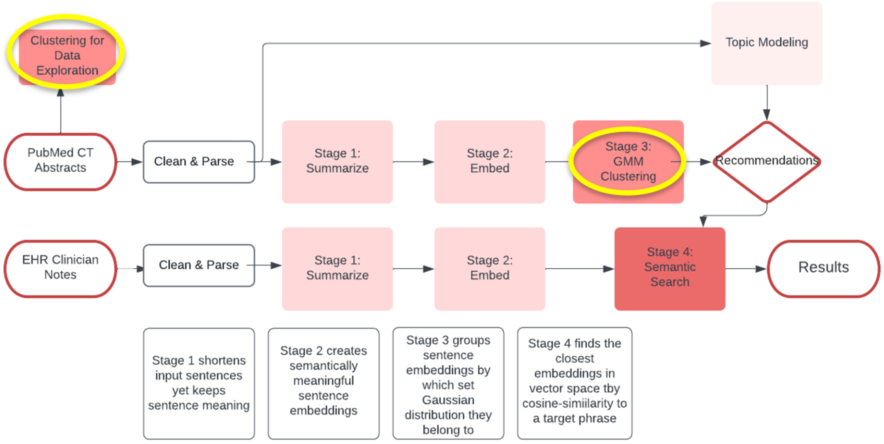

Clustering for Data Exploration
Topic Modeling
PubMed CT Abstracts
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 3: GMM Clustering
Recommendations
EHR Clinician Notes
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 4: Semantic Search
Results
Stage 1 shortens input sentences yet keeps sentence meaning
Stage 2 creates semantically meaningful sentence embeddings
Stage 3 groups sentence embeddings by which set Gaussian distribution they belong to
Stage 4 finds the closest embeddings in vector space tby cosine-simiilarity to a target phrase

Figure 2.4: Clustering for Data Exploration

Next, in our data exploration, we tested different clustering algorithms. Unlike topic modeling, clustering first splits documents into a number of groups, or clusters, based on a similarity score. We tested multiple clustering algorithms to see which worked best with our datasets and then explored the clusters of the best-performing technique. We completed additional rounds of clustering in further analysis, which is discussed in greater detail later on in this section.

We chose three clustering algorithms to feed word tokens to: K Means, Gaussian Mixture Models (GMM), and spectral clustering. K Means selects k number of centroids and assigns every data point to the nearest cluster while keeping the centroids as small as possible. GMMs assume all data points are generated from a mixture of Gaussian distributions with unknown parameters. Lastly, spectral clustering treats clustering as a graph partitioning problem, where nodes map to a low-dimensional space that can be easily separated to create clusters. We evaluated all models for clusters ranging from three to ten and ranked them by highest silhouette score and lowest Davies Bouldin score. The Silhouette score measures cluster quality by how well the data fits within each cluster; higher valued scores mean the clusters are strong and the points within that cluster are the most similar to the other points within that cluster. The Davies Bouldin score measures the average similarity between one cluster and the closest neighboring cluster; lower values mean stronger clusters because it means the clusters are more distinct from each other. We chose the silhouette and Davies Bouldin scores as our our definition of cluster quality.

K Means and GMM performed best with nine clusters, while spectral performed best with 7. Table 2.1 shows the results from clustering for data exploration.

Cluster Type	**n**	**Silhouette Score**	**Davies Bouldin Score**
K Means	9	0.055	2.982
GMM ”full”	9	0.066	2.948
GMM ”diagonal”	9	0.063	3.006
Spectral	7	0.057	3.100

Table 2.1: Clustering Algorithms and Results for Data Exploration of PubMed CT Abstracts

Of the three clustering algorithms tested, the GMM with full covariance performed the best overall, with a silhouette score of 0.066 and a Davies Bouldin score of 2.948. Full covariance is when cluster components can independently adopt any shape or position. We further explored the three clustering techniques by generating word clouds for each of the three algorithms with the best-performing number of clusters to get a sense of the different characteristics that make up each cluster. Word clouds generated were the most used words within each cluster.

The word clouds revealed that the GMM with full covariance outperformed K Means, despite both methods showing the best performance with nine clusters and similar scores for cluster quality. The GMM with full covariance produced more conceptually coherent clusters, with distinct word clusters emerging around types of exercise, disease interventions, and their results. Within these word clusters, words around the three disease types from topic modeling also surfaced, like ‘heart,’ ‘cardiovascular,’ ‘insulin,’ and ‘depression.’ Spectral clustering had a higher silhouette score compared to K Means and GMM. However, it also had the highest Davies Bouldin score indicating poorer performance than the GMM with full covariance. Figure 2.6 shows the word clouds from the GMM with full covariance.

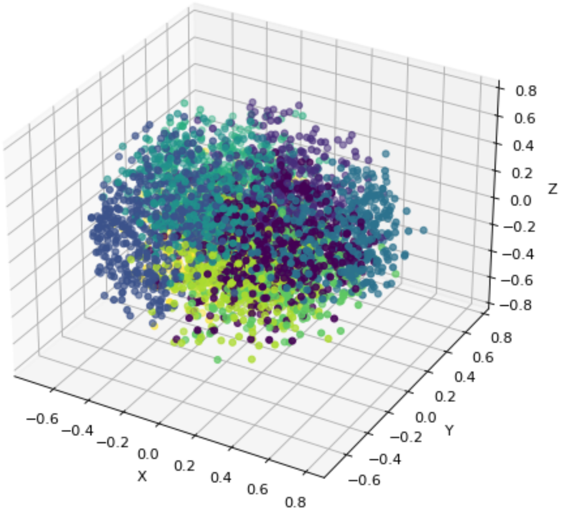


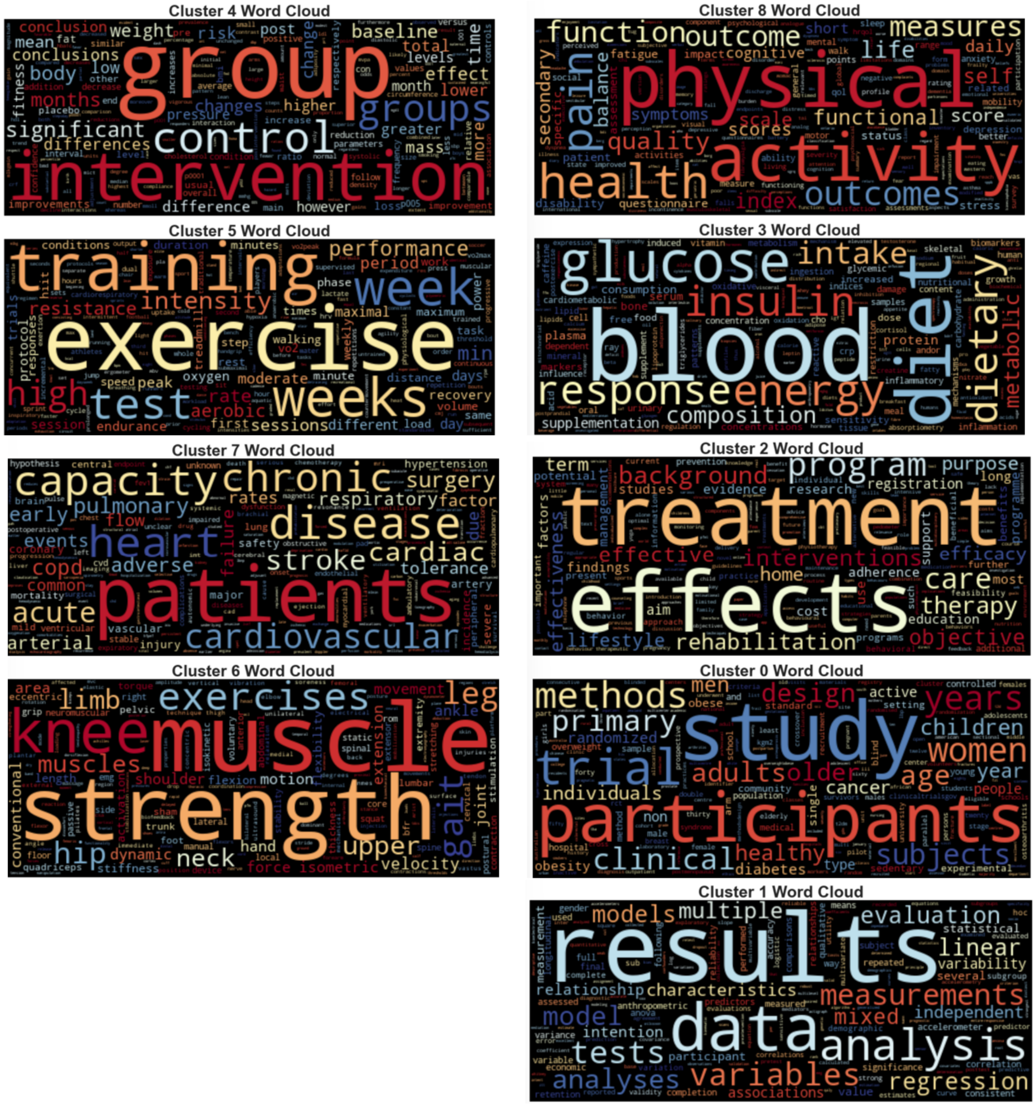
Figure 2.5: PubMed CT Data: GMM “Full”

2.2 Sentence Transformers

Sentence transformers are a form of deep learning that provide streamlined methods to compute dense vector representations, or embeddings, for text and image data [8]. Vector representations enable the processing of large amounts of unstructured data, making it easier to work with and analyze text data. They are used in a variety of natural language processing tasks, such as semantic similarity, summarization, classification, and information retrieval as they are able to capture the meaning and context of the data and maintain it throughout the transformation process. Tranformers use self-attention, a variant of attention, to process a sequence by replacing each element with a weighted average of the rest of the sequence [22]. They represent text in high-dimensional space, with each dimension corresponding to a specific word. By representing text as numeric vectors in dimensional space, a piece of text is positioned closer to other parts of text that have a similar meaning. We measure closeness by cosine similarity:

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{t} \cdot \mathbf{e}}{\|\mathbf{t}\| \|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{2.1}$$

In this equation, $\cos(t, e)$ represents the cosine of the angle between vectors t and e, which is equal to the dot product of t and e divided by the product of their magnitudes. The vectors t and e can be represented



Cluster 4 Word Cloud

Cluster 8 Word Cloud

Cluster 5 Word Cloud

Cluster 3 Word Cloud

Cluster 7 Word Cloud

Cluster 2 Word Cloud

Cluster 6 Word Cloud

Cluster 0 Word Cloud

Cluster 1 Word Cloud

Figure 2.6: Word clouds generated for each of the 9 clusters of the best performing GMM with full covariance

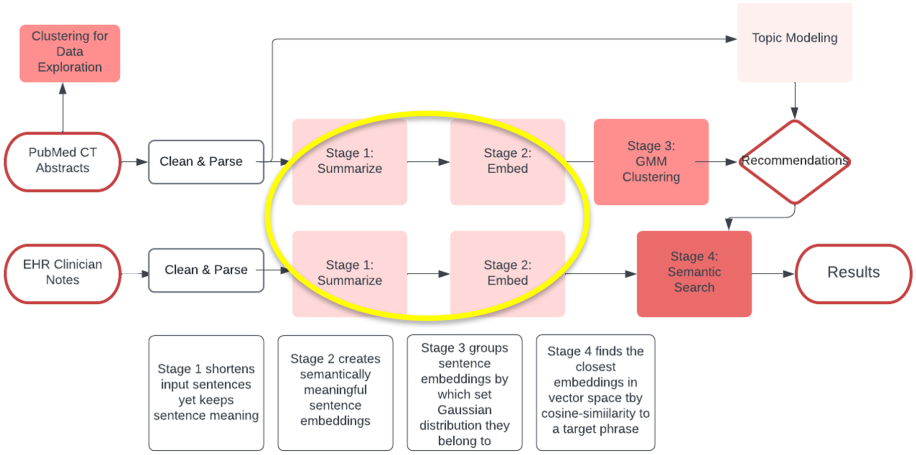

Clustering for Data Exploration
Topic Modeling
PubMed CT Abstracts
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 3: GMM Clustering
Recommendations
EHR Clinician Notes
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 4: Semantic Search
Results
Stage 1 shortens input sentences yet keeps sentence meaning
Stage 2 creates semantically meaningful sentence embeddings
Stage 3 groups sentence embeddings by which set Gaussian distribution they belong to
Stage 4 finds the closest embeddings in vector space tby cosine-simiilarity to a target phrase

Figure 2.7: Sentence Transformers for Summarization and Embedding

as column vectors of their respective components. Sentence transformers make it easy to quickly compare, cluster, and retrieve pieces of text. Our research makes use of sentence transformers for summarization, embedding, and semantic similarity searches.

2.3 Summarization and Embedding

Building off of the previous section on sentence transformers, we next review transformers as they pertain to text summarization. The text summarization process creates new sentences based on input text that are shorter and more concise while still capturing the main ideas from the original text. Summarizing text is one of the most challenging NLP tasks. An effective summarizing model must process the input and generate coherent text that captures the same meaning while removing unnecessary filler. There are two primary forms of summarization: extractive and abstractive. Extractive summarization uses only words found in the original sentence, while abstractive summarization can augment based on new words.

For our research, we made use of extractive summarization. We first fed the abstract results, and EHR treatment plans to the T5-base model. All results and treatment plans were then summarized and capped at a max length=25. By making use of summarization, sentences were ensured to be roughly equal in size and easier on computation. An example of text summarization is shown in table 2.2.

After using sentence tranformers to summarize the abstract results and treatment plans, we next used them to embed the summarized sentences with the all-MiniLM-L6-v2 model. Embedding is an NLP technique that transforms unstructured data, such as text or images, into a structured and numerical form, in this case, as a high-dimensional numerical vector. After embedding, each piece of text is mapped into vector space where each word, or other specified characteristic, represents a dimension. Once embedded, we can analyze

Note Text	**Summarized Text**
Exercise weekly, and reducing saturated fats in the diet #HTN–On amlodipine 10 mg daily –Blood pressure still elevated, reporting improved blood pressures at home but –Patient counseled on the following lifestyle modifications for non-pharmacologic treatment of hypertension. Dietary salt restriction, potassium supplementation, weight loss, DASH diet, moderate intensity aerobic exercise five times per week for 30 minutest at a time, and limiting alcohol intake. Return to clinic in three months...	Patient counseled on lifestyle modifications for non-pharmacologic treatment of hypertension

Table 2.2
An Example of Text Summarization

the relationships and similarities between different pieces of text, and various computational or mathematical techniques can be applied, such as text classification, sentiment analysis, or semantic searches. Our research makes use of clustering on embeddings and semantic similarity search.

The embedding process was identical for both datasets, except that duplicates were not removed in the EHR data. By the nature of EHRs, and especially in the emergency department's fast-paced environment, physicians often copy and paste where applicable in their notes to save time. Therefore, we kept all 4,939 duplicates in the treatment plans in order to not skew counts where duplicate plans were made for different visits or patients.

No additional, task-specific training was done on either the PubMed abstracts or EHR treatment plans. By relying solely on the all-MiniLM-L6-v2 model's pre-training, we utilized a self-supervised learning framework that could generate meaningful results without the need to acquire additional domain-specific data or provide labeled data. This approach is particularly powerful in the healthcare domain, where acquiring and labeling medical data is difficult. Our self-supervised learning framework saved time and resources and allowed us to demonstrate the effectiveness of a self-supervised learning framework for other highly specialized domains beyond healthcare.

2.4 Clustering Analysis

Next, we used the summarized and embeded sentences from the previous section with GMM clustering to group similar abstracts and treatment plans into distinct clusters. We chose to use GMM Clustering since it performed best with our datasets during initial data exploration and because they allow for overlapping clusters, which provide a more insightful analysis. Our goal with clustering was to create clusters based on relationships within the data by grouping by cosine similarity to identify underlying trends. As a reminder,

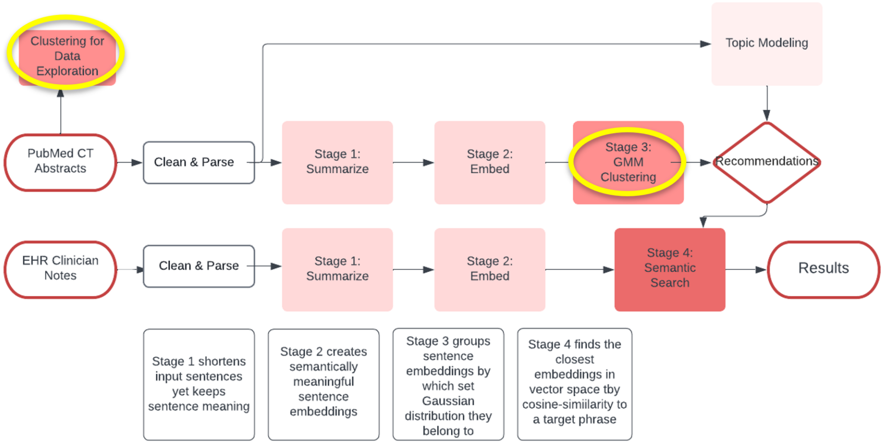

Clustering for Data Exploration
Topic Modeling
PubMed CT Abstracts
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 3: GMM Clustering
Recommendations
EHR Clinician Notes
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 4: Semantic Search
Results
Stage 1 shortens input sentences yet keeps sentence meaning
Stage 2 creates semantically meaningful sentence embeddings
Stage 3 groups sentence embeddings by which set Gaussian distribution they belong to
Stage 4 finds the closest embeddings in vector space tby cosine-simiilarity to a target phrase

Figure 2.8: GMM Clustering

the equation for cosine similarity is shown below.

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{t} \cdot \mathbf{e}}{\|\mathbf{t}\| \|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{2.2}$$

We completed multiple rounds of GMM clustering, but each round started the same way. GMMs were evaluated with components ranging from three to ten and ranked by the highest silhouette score. The generated GMMs were the ones that had the best cluster quality, defined as having the highest silhouette score and lowest Davies Bouldin score. Once again, the silhouette score measures cluster quality by how well the data fits within each cluster; higher valued scores mean the clusters are strong and the points within that cluster are the most similar to the other points within that cluster. While, the Davies Bouldin score measures the average similarity between one cluster and the closest neighboring cluster; lower values mean stronger clusters because it indicates the clusters are more distinct from each other.

2.5 Similarity Search

The last step in our methodology was to implement a series of semantic similarity searches. A similarity search matches relevant pieces of data together using features or characteristics of the item of interest, in this case, a target sentence. We selected target sentences around different exercise interventions recommended from PubMed clinical trials that we found through the previous section's clustering analysis. We then searched for these exercise interventions in the summarized treatment plans from the EHR data. These searches returned the most similar treatment plans to whichever target sentence around an exercise recommendation was used.

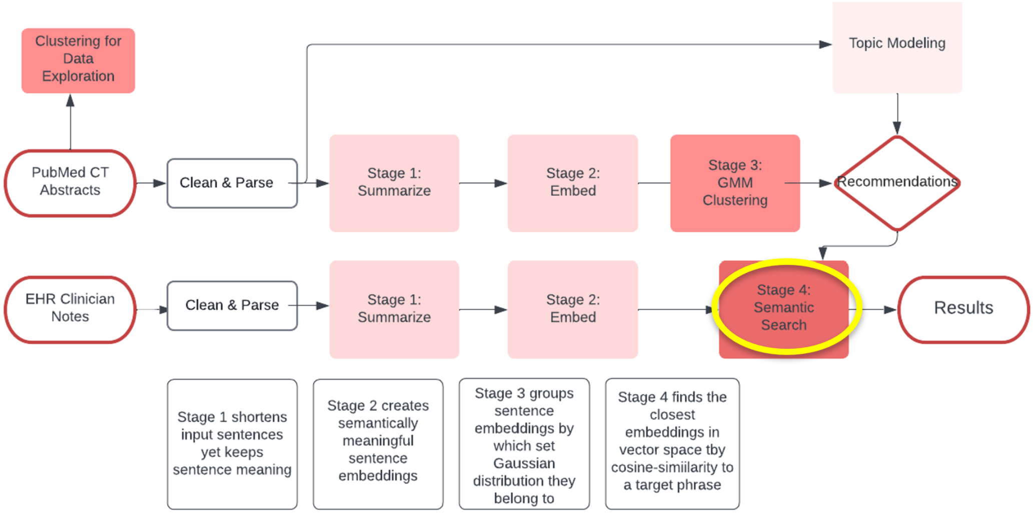

Clustering for Data Exploration
Topic Modeling
PubMed CT Abstracts
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 3: GMM Clustering
Recommendations
EHR Clinician Notes
Clean & Parse
Stage 1: Summarize
Stage 2: Embed
Stage 4: Semantic Search
Results
Stage 1 shortens input sentences yet keeps sentence meaning
Stage 2 creates semantically meaningful sentence embeddings
Stage 3 groups sentence embeddings by which set Gaussian distribution they belong to
Stage 4 finds the closest embeddings in vector space tby cosine-similiarity to a target phrase

Figure 2.9: Semantic Similarity Search

To implement this, we utilized Facebook artificial intelligence similarity search, or FAISS, as it is widely recognized as a very computationally efficient vector database [20]. FAISS has a set of algorithms to search in dense vector sets of any size to compute the argmin on the data based on the target sentence or query vector. From there, the distance is measured between each returned sentence and the target sentence, and the returned sentence with the smallest distance from the target is considered the most similar. For our research, we chose FAISS to measure the Euclidean distance between all points between the target sentence and the vectors in the index. Euclidean distance uses the Pythagorean theorem to measure and calculate the straight line distance between point A and point B.

Euclidean distance is measured as:

$$\textsc{Distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.3}$$

Where x and y are vectors with n elements each, and i represents the i-th element of the vectors.

Chapter 3

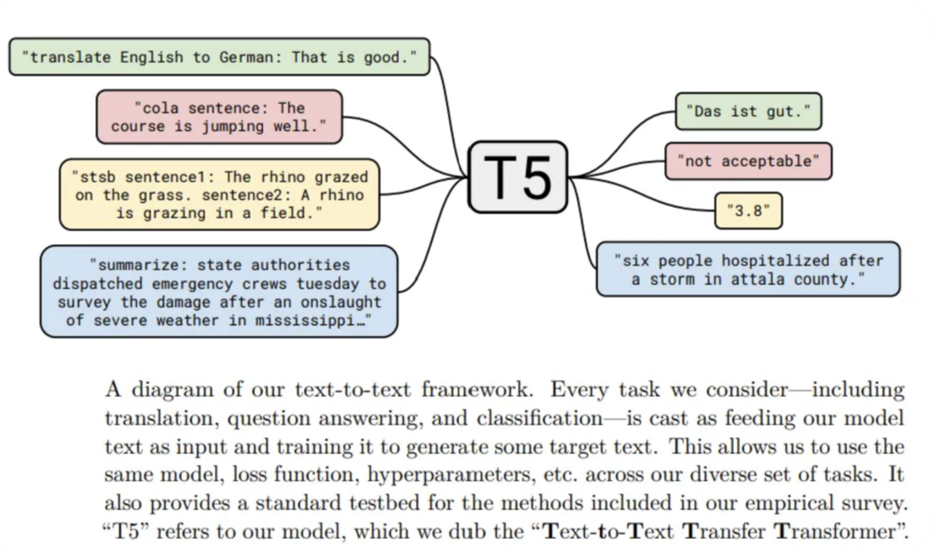
Models

3.1 Summarization Models

For multiple reasons, the T5-Base transformer model was a strong choice for summarizing the abstract results sections and EHR treatment plans. First, we could use it without specific training on our data and still produce accurate and intelligent summaries. This made our approach much simpler and more easily reproducible. Second, the T5-base model is highly compatible with many systems due to its unified text-to-text format, which utilizes inputs and outputs as strings. This allows for the same model, loss function, and hyperparameters throughout all NLP tasks. Every task for the model is considered as input that iteratively feeds the model and trains the input to generate target text [5]. Third, pre-training included both supervised text-to-text language modeling and unsupervised de-noising tasks [5]. And lastly, it produces state-of-the-art results for multiple NLP tasks [22]. All of these considerations made the T5-base model a good choice for our self-supervised learning framework.

Specifically, the T5-base model passes an input sequence of tokens into embeddings which then pass to an encoder. The encoder is made up of a self-attention layer and a small feed-forward network layer with layer normalization applied to each. Next, a residual skip connection adds each layer's input to its output and applies dropout to the feed-forward network, the skip connection, the attention weights, and the input and output of the entire stack. The decoder is structured similarly to the encoder. However, it includes a standard attention mechanism after every self-attention layer to attend to the output of the encoder and only to other past outputs. Lastly, the final decoder block output feeds to a dense layer with a soft-max output whose weights are shared with the input embeddings [22].

Based on our qualitative review of summarized sentences, we found no serious loss of information. This is


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

Figure 3.1: T5 Model Task Formulation: Figure from the T5 paper

especially true for the already condensed PubMed results sections. However, future work for this research can include a more exhaustive exploration of potential losses in summarized treatment plans.

3.2 Embedding Models

As stated previously, for our framework, we wanted to make use of embedding models without needing to do additional training on the data, yet still obtain meaningful results in the healthcare field. By doing this, our proposed framework would save resources by not needing to find and acquire large amounts of labeled, domain-specific data and would be more reproducible for future research and application in other fields.

The primary goal of an embedding model is to transform entire sentences into a structured representation while keeping their semantic meaning. We considered three different embedding models to see which performed the best in a self-supervised learning approach. All of the models considered for embedding purposes utilize a self-supervised contrastive learning objective. Contrastive learning is when given a sentence, the model should accurately predict which, out of a randomly sampled set, other sentence it is paired with in the dataset. We provide the contrastive learning equation below:

$$\mathcal{L} = -\log \frac{\exp(\textsc{Sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} [k \neq i] \exp(\textsc{Sim}(z_i, z_k)/\tau)} \tag{3.1}$$

Where $\mathcal{L}$ is the loss function, z_i and z_j are embeddings of two augmented versions of an input, $\textsc{Sim}(z_i, z_j)$ is a similarity function that measures the similarity between the two embeddings, τ is a temperature parameter, and N is the number of samples [12].

The three sentence transformers analyzed for embedding purposes were the all-MiniLM-L6-v2, multi-qa-mpnet-base-dot-v1, and all-mpnet-base-v2. The all-MiniLM-L6-v2 was created with the intention of being used for both clustering and semantic search. It uses a 384-dimensional vector space and was trained on a dataset of 1 billion sentence pairs [23]. The multi-qa-mpnet-base-dot-v1 was created with the intention of being used for purely semantic search. It uses a 768-dimensional dense vector space and was trained on 215 million questions and answers from diverse sources [4]. The all-mpnet-base-v2 was created with the intention of being used for clustering and semantic search. It also uses a 768-dimensional dense vector space and was trained on over 1 billion sentences [3].

The performances of the embedding models were all comparable. This did not come as a surprise since all of the models were pre-trained on massive corpora, but none included any domains specific to healthcare. All things considered, we decided to use the all-MiniLM-L6-v2 model for several reasons. First, the all-MiniLM-L6-v2 is derived from the all-mpnet-base-v2 model, which has the best quality as rated by SBERT [24]. It performs the highest in both performance sentence embeddings and performance semantic search. Second, it is a smaller version of the all-mpnet-base-v2 model, performing 5x faster while retaining sound quality. And third, the all-MiniLM-L6-v2 was intended for both clustering and semantic search, which are both used throughout our research.

Chapter 4

Results

4.1 Clustering

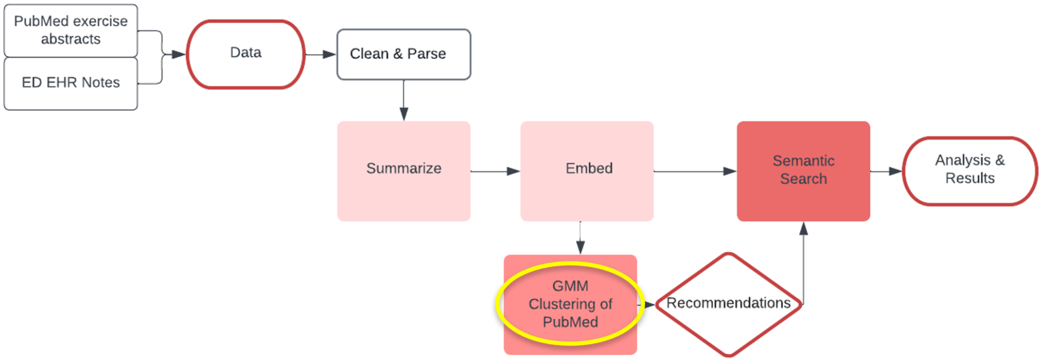
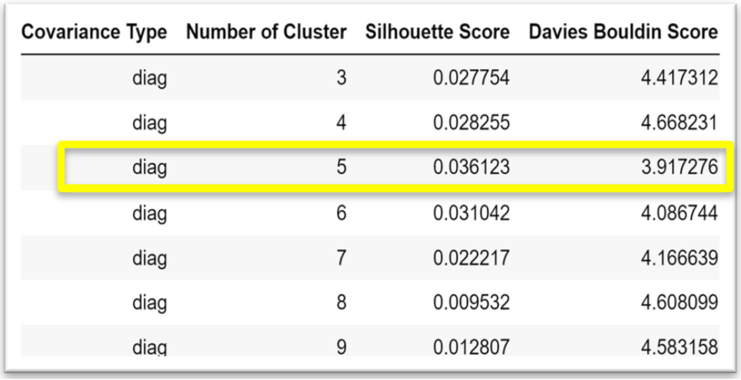


Figure 4.1: GMM Clustering Results

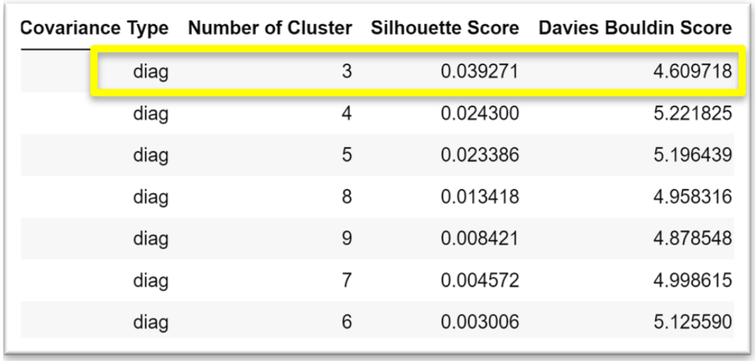
To group abstract results sections with similar themes, we used GMM Clustering on the word embeddings. The different clusters allowed us to explore and analyze similar recommendations in the clinical trial results that we could then look for in the EHR data. After our first round of GMM clustering, the GMM with five cluster components and a diagonal covariance performed the best, with a Silhouette Score of .036 and a Davies Bouldin score of 3.917. A diagonal covariance is when the contour axes are oriented along the coordinate axes, but other eccentricities may vary. GMM clustering with diagonal covariance performed the best on the word embeddings, whereas full covariance performed the best during data exploration on the word tokens. We generated the GMM, and assigned cluster labels to each abstract's results section so that the clusters could be grouped and explored. The cluster themes from the 5 component GMM focused primarily on the specific wording of the results sections, such as whether or not the findings were conclusive or significant or if any change was observed and not on the actual content of the results.



Covariance Type	Number of Cluster	Silhouette Score	Davies Bouldin Score
diag	3	0.027754	4.417312
diag	4	0.028255	4.668231
diag	5	0.036123	3.917276
diag	6	0.031042	4.086744
diag	7	0.022217	4.166639
diag	8	0.009532	4.608099
diag	9	0.012807	4.583158

Figure 4.2: GMM Clustering Results Round 1

To address this, we performed another round of clustering to group by whether the clusters from the GMM with five components had conclusive results; this left clusters 0, 1, and 2 and a total of 10,791 clinical trial abstract results. We combined clusters 0, 1, and 2, and again GMMs were evaluated based on components ranging from three through ten with diagonal covariance. The GMM with three cluster components performed the best, with a Silhouette score of .039 and a Davies Bouldin score of 4.609. We once again explored the clusters to pull out underlying themes, and this time clusters revolved around intervention types. Cluster 0 focused on weight management techniques and dieting, cluster 1 focused on counseling, therapy, breathing techniques, and yoga, and cluster 2 focused on exercise programs and improving muscle strength.



Covariance Type	Number of Cluster	Silhouette Score	Davies Bouldin Score
diag	3	0.039271	4.609718
diag	4	0.024300	5.221825
diag	5	0.023386	5.196439
diag	8	0.013418	4.958316
diag	9	0.008421	4.878548
diag	7	0.004572	4.998615
diag	6	0.003006	5.125590

Figure 4.3: GMM Clustering Results Round 2

The final round of clustering focused on the three disease types. Having selected the different diseases through topic modeling, we filtered for them directly through keyword searches. With the help of UCI's team,

Disease	**Intervention**
CAD	Aerobics, HIIT, Strength Training
Diabetes	Weight Management Techniques, Counseling, Endurance training
Mental Health	Breathing Exercises, Yoga, Web & Mobile Phone Apps

Table 4.1: Clustering on Disease Type

we compiled a list that included different naming conventions and symptoms for each condition. Clustering was performed a third time on each of the grouped abstracts by the disease-type filters. We saw similar cluster themes for each disease type around different interventions like education, diet, and exercise. We chose the intervention types for each condition based on the most occurring and most relevant interventions. Thus, we selected aerobics and strength training for heart disease, education and diet for diabetes, and mindfulness, and web / mobile phone apps for mental health. Not only did aerobics and strength training show up when clustering abstract results sections, but they also appeared in topic seven from topic modeling. Moving forward, we focused solely on heart disease, specifically coronary artery disease (CAD), and its associated interventions. We selected CAD because it is the most common heart disease, and UCI's emergency department is a cardiac patient-receiving center. This way, we had a more focused signal to follow in the EHR data.

```
HD = ['heart disease', 'cardiovascular disease', 'clogged arteri
es', 'coronary disease', 'angina', 'heart failure', 'peripheral
arterial disease', 'PAD', 'atherosclerosis', 'coronary', 'CAD',
'heart attack', 'arrhythmia', 'chest pain', 'heart’,
’tightness’ ]
```

Figure 4.4: Keywords used to capture mentions of heart disease

Clusters appeared for each disease around interventions found in the previous clustering rounds as shown in table 4.1.

4.2 PubMed Recommendation Table

To ensure we included all relevant, heart-disease-related PubMed clinical trial abstracts, we used the same keyword filter for heart disease that we did in the last round of clustering analysis and filtered the whole abstract text instead of just the summarized results sections. This made sure that we were not excluding any abstracts that related to heart disease that did not mention it explicitly in their summarized results section. Broadening our search criteria to the entire PubMed abstract increased the number of clinical trial abstracts pertaining to heart disease from 443 abstracts to 2,629. We then performed yet another round of keyword searches to get the updated counts for each recommended intervention for heart disease, as shown in Table 4.2.

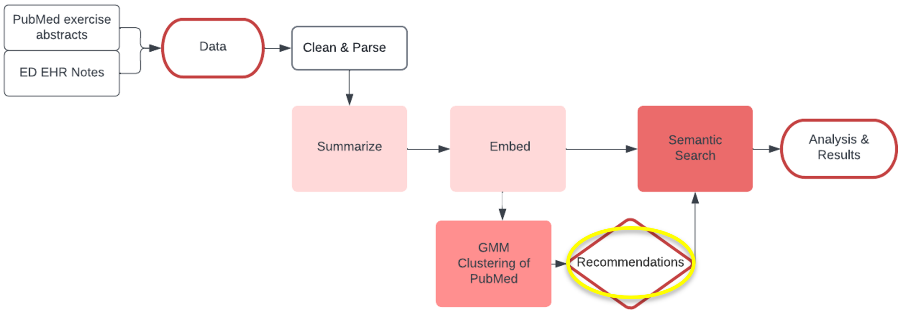


Figure 4.5: PubMed Recommendations

CAD Intervention	Count
Aerobics	228
HIIT	103
Strength Training	59

Table 4.2: Heart Disease PubMed Recommendation Table: Counts for each recommended intervention

4.3 Similarity Search

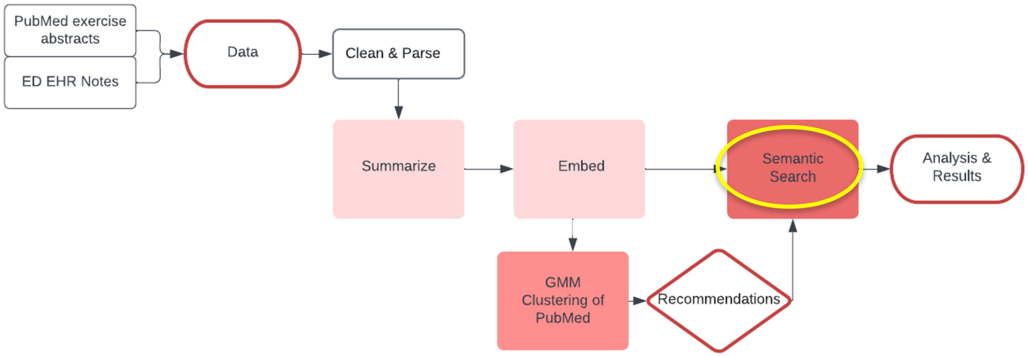


Figure 4.6: Semantic Similarity Search Results

Next, we needed to see if the recommendations found in the PubMed clinical trial abstract data from our clustering analysis appeared in UCI's emergency department EHR data. With the EHR data already summarized and embedded, we implemented the final methodology in our approach: semantic similarity search. We completed a series of semantic similarity searches based on each recommended exercise intervention for heart disease and compared with the embedded EHR treatment plans.

We decided to break up our searches on the summarized EHR data into two main rounds: one on the overall dataset and one filtered specifically for CAD based on a combination of ICD-10 codes and mentions of CAD in the clinician note. The overall dataset had treatment plans that mentioned aerobic exercise or strength training, while the CAD specific treatment plans had mentions of CAD or a corresponding ICD-10 code and mentions of aerobic exercise or strength training. When filtering for just ICD-10 codes for CAD, only 82 records came back, likely because an underlying condition is less likely to be why patients visit the emergency department; hence these conditions are unlikely to be used for billing purposes. That being said, CAD as an underlying condition can cause symptoms that result in a patient visit, like chest pain or shortness of breath [14]. Therefore, we included the 82 records and filtered for mentions of CAD or associated symptoms within the clinician note.

CAD ICD-10 Codes	Count
I25.10, I25.83, I25.89, I25.118, I25.5, I25.81, I25.2, I25.119, I25.709, I25.719, I25.810, I25.89, I25.9	82

Table 4.3: Heart Disease ICD-10 codes found in the EHR data

We included both rounds of searches because each one told an interesting story. To start, the overall dataset returned the most occurrences of these recommendations as it contained more data and had less filtering on it. As a reminder, these records had aerobic exercise or strength training mentioned anywhere in their treatment plan and were not necessarily associated with coronary artery disease. These results act as a baseline for how often exercise is recommended as an intervention in emergency department treatment plans. We also used them to compare the number of mentions of these exercise interventions with the number of mentions in the clinical trial results sections.

Intervention	**Example Text**	**Distance**	**Count**
AER: "practice more aerobic exercise"	"exercise through combination of aerobic/resistance/weight training"	119.05 – 210.37	472
AER: "practice more aerobic exercise like walking or jogging"	"aerobic activities. These are exercises, such asbrisk walking or water aerobics"	109.11 – 208.11	389
AER: "exercises like walking or jogging"	"exercise/movement: fast paced walking, swimming, running, biking/cycling, aerobic circuits"	144.78 – 253.14	124
STR: "increase muscle strength"	"PT services to increase strength and endurance, increase static and dynamic balance"	61.87 – 223.34	381
STR: "exercise like strength training"	"exercise; Patient education; Progressive resistive training; Strengthening exercises."	109.17 – 210.78	195

Table 4.4: Similarity Search on Overall EHR Dataset

The second round of semantic similarity searches was on the combined filter of heart disease-related ICD-10 codes and CAD-related keywords. See Table 4.3 for ICD-10 codes found in the EHR data that contributed to the CAD filter. The filtered dataset contained 432 records, and we used the same target phrases from the first round of similarity searches on the overall dataset. None of these phrases returned treatment plans that shared a similar meaning regardless of distance. We continued the searches with different phrasing around aerobic exercise and strength training. These searches included mentions of starting or increasing said exercise as well as mentioning different types of each exercise. Even still, no semantically similar treatment plans were returned. To confirm this, we manually searched the CAD data to see whether these results were accurate. After reviewing each record captured by the CAD filter, a few treatment plans should have been returned for aerobic and strength training. However, these recommendations still appeared in less than 3% of CAD health records. Table 4.5 shows the results from the manual search.

Intervention Type	**Treatment Plan**	**Count**
Aerobics	”if you joints are healthy, you do running, stair-master, play sports” ”150 minutes of moderate, high intensity aerobic exercise weekly” ”walk 5-10 minutes twice a day”	7
Strength	”exercises; resistive band exercises; strengthening”	2

Table 4.5: Manual Search on CAD EHR Dataset

After reviewing our results, we found that 12.7% of clinical trial abstracts mentioned aerobic exercise or strength training. Of the treatment plans we reviewed from UCI’s emergency department, 19.2% related to heart disease. Of these, the treatment plans that mentioned CAD mentioned aerobic exercise or strength training only 0.34% of the time. Lastly, treatment plans from the overall dataset from UCI mentioned aerobic exercise or strength training less than 5% of the time. One thing to keep in mind is that both of these datasets were acquired with an exercise-focused lens.

Chapter 5

Discussion

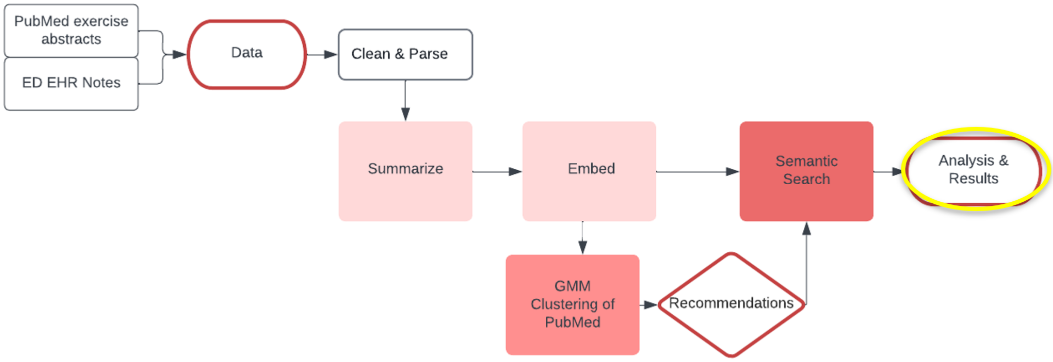


Figure 5.1: Discussion and Conclusion of Results

After reviewing the results, it is apparent that physicians in the emergency department rarely mention exercise prescriptions as a treatment plan. Specifically, treatment plans for coronary artery disease mentioned aerobic and strength training exercises only 0.34% of the time, and treatment plans for the overall dataset mentioned these exercises only 5% of the time. The results of our research are significant as they illustrate that clinical research can take years to make it into EHR practice, and even then, the results might be rarely used, especially in the case of exercise interventions. What's more, they show that exercise is seldom mentioned in the context of the emergency department.

There are many reasons that exercise may not appear in ED treatment plans, and exploring these reasons is another possible area of future research. However, after discussions with the UCI medical team, we identified a few potential reasons exercise does not appear in the emergency department.

To start, a significant percentage of emergency department patients seek care for non-urgent reasons.

However, emergency department physicians are under tremendous pressure to focus on the most seriously ill, which limits time for patients with less severe cases. Moreover, these physicians lack the necessary equipment and training to effectively provide advice on CRF, which is thought to be reserved for the primary care setting. This disconnect can result in missed opportunities to provide patients with guidance and support that could lead to improved health outcomes and a reduced need for emergency care down the line. These issues are particularly significant post-pandemic, where patients now present with a range of health concerns that were not addressed during the peak of the COVID'19 outbreak.

Additionally, walk-ins have become more convenient than scheduling appointments with primary care providers, which could result in waiting months before receiving medical attention. And in general, physicians may not have enough time to stay educated on the latest clinical research with their busy workloads and the amount of research coming out. Lastly, prescribing exercise interventions do not bring in money like prescribing medication or a particular procedure would from pharmaceutical and insurance companies, which can further discourage their adoption.

The points raised above substantiate the ongoing discussion on the interplay between theory and practice in the field of medicine. The theoretical knowledge found in clinical research does not make its way into actual medical practice, with this disconnect being further exacerbated in the emergency department setting. The disparity between theory and practice, or abstract and concrete, underscores a key systems integration challenge, where a loss of information occurs throughout the decomposition of a complex process. This challenge is perfectly exemplified by the angiogenesis case, which can be further read about here [19]. There are many proposed solutions for this disconnect, but none seem adequate, as evidenced by the persistent nature of the topic for discussion and research. Whether it is the aforementioned reasons above or because people who study theory are different from those who practice, it is imperative that medical professionals acknowledge and actively bridge this gap between translation science to provide the best possible care for patients and to improve processes within hospital systems.

Given all of these reasons, it is crucial to note that emergency department visits could provide a valuable opportunity for physicians to influence patient well-being through health messaging. Particularly since if an individual feels strongly enough to seek emergency care, they are typically experiencing a higher level of concern for their health and will likely be more receptive regarding their health.

Although the amount of time between a patient and a physician in this setting will not change, the use of direct querying from PubMed, patient charts, and social determinants can be leveraged to find important information and deliver targeted health messages to the patient. To make the most of the limited time available, health messaging could be in the form of the three main takeaways the physician should discuss with each patient regarding their health. In this way, the emergency department can serve as a powerful tool for physicians to effect positive change in patients' lives, despite the challenges associated with emergency care.

To the best of our knowledge, there have been no other efforts to quantify the gap between "bench to bedside" practice, making our research a unique contribution to the healthcare industry. By quantifying this gap, we provide a benchmark to determine whether, over time, the use of CRF-related dialogue might increase even in the context of a busy emergency department. Our novel approach introduces a self-supervised framework that is extremely powerful for highly specialized domains or situations with limited time or resources. This approach enhances the effectiveness of research and ensures easy reproducibility. Furthermore, it can be easily modified to incorporate more advanced models, allowing for further improvements in analysis. Due to its versatility, this approach can be applied to any data type, making it a valuable tool for researchers across many fields.

5.1 Limitations

Limitations of this study include the potential information loss resulting from summarizing the EHR treatment plans. To check for loss, we manually audited a sample of the summarized treatment plans. We did not identify any major information losses, but further testing on potential loss could be a future step in this research. Summarized PubMed results sections are less of a concern as they are already condensed to fit into the abstract format making it harder to lose important information throughout summarization.

Another potential limitation is the low results from the semantic similarity searches on the EHR dataset. Comparing the semantic similarity search with a manual search showed nine misses. A potential solution could be to switch the embedding or semantic search models to see if a different one might perform better at accurately catching these occurrences. Another solution could be to work with medical teams to create more domain-specific search queries to identify interventions in the treatment plans. Involving the medical team could lead to more relevant search queries with more medical verbiage or formatted in ways more likely to be seen in ED EHRs. On the other hand, the margin of error is low, with less than 3% of abstracts containing these interventions after the manual search, so confirming with stakeholders what level or error they are comfortable with could also be the answer.

5.2 Future Work

Future steps to expand this research question within UCI include:

- Adding more departments to the EHR dataset

- Increasing the time period beyond the current one
- Looking at conditions aside from just CAD

Other hospital systems can utilize and expand upon the research question at hand, and moreover, other highly specialized domains can readily apply the proposed framework as well.

Chapter 6

Conclusions

Having access to publicly available clinical research and associated electronic health record data, including the clinician notes, provided a unique opportunity to assess the effectiveness between clinical research and its adoption in medical practice. Furthermore, using a specific condition, such as Coronary Artery Disease, allowed us to hone in on and track a distinct signal throughout the medical research process directly to a physician's treatment plans.

The central premise guiding this work was that although many conditions can be greatly ameliorated through physical activity, very rarely is it mentioned, evaluated, or prescribed to patients during a standard ED visit. To investigate this theory, we created a self-supervised approach to improve query retrieval of EHR data. We confirmed that although certain types of exercise are mentioned in the PubMed clinical trials abstract data, they rarely are prescribed for the prevention or mitigation of CAD diagnoses or symptoms in the emergency department. In order to fully evaluate the success of this study, a success measure needs to be defined, but ultimately that measure should be up to the stakeholders, in this case, the UCI Medical Team.

This study developed a system for using combined NLP approaches to generate scientific results. We provide a valuable and unprecedented approach to investigating translation science in medicine that can be leveraged for a wide range of other important applications. This self-supervised framework for direct query-based searching can be transferred across healthcare and beyond, including other highly specialized domains such as physics, law, actuarial science, and more. Ultimately, our approach can be used in any scenario where a machine-learning model is required to understand and respond to language queries. We release the queried PubMed dataset used for our research as well as the code used on the PubMed abstracts. This dataset can be used to answer further research questions around exercise and medical decision-making, and the released code can be used as a streamlined approach to parse and clean future PubMed datasets.

Code used on EHR data cannot be released due to privacy concerns regarding individual patient medical history.

Bibliography

[1] Rafael Alvarado. “DS5001-2022-01”. In: *GitHub* (2022). URL: https://github.com/ontoligent/DS5001-2022-01.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[3] Omar Espejel. “all-mpnet-base-v2”. In: *Hugging Face* (2023). URL: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[4] Omar Espejel. “multi-qa-mpnet-base-dot-v1”. In: *Hugging Face* (2022). URL: https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. “Simcse: Simple contrastive learning of sentence embeddings”. In: *arXiv preprint arXiv:2104.08821* (2021).

[6] Abdul Hafeez et al. “A review of COVID-19 (Coronavirus Disease-2019) diagnosis, treatments and prevention”. In: *Ejmo* 4.2 (2020), pp. 116–125.

[7] Toshitaka Hayashi and Hamido Fujita. “Cluster-based zero-shot learning for multivariate data”. In: *Journal of ambient intelligence and humanized computing* 12.2 (2021), pp. 1897–1911.

[8] Hugging-Face. “Using Sentence Transformers at Hugging Face”. In: *Hugging Face* (2023). URL: https://huggingface.co/docs/hub/sentence-transformers.

[9] B Ian Hutchins et al. “Predicting translational progress in biomedical research”. In: *PLoS biology* 17.10 (2019), e3000416.

[10] Andrew M. Jones. “Medicine & Science Sports & Exercise.” In: *American College of Sports Medicine* (2023). URL: https://www.acsm.org/education-resources/journals/medicine-science-in-sports-exercise.

[11] Nicholas A. Ratamess Jr. "The Journal of Strength and Conditioning Research." In: *National Strength and Conditioning Association* (2023). URL: https://journals.lww.com/nsca-jscr/pages/default.aspx.

[12] Prannay Khosla et al. "Supervised contrastive learning". In: *Advances in neural information processing systems* 33 (2020), pp. 18661–18673.

[13] Emily Knight and Robert J Petrella. "Prescribing physical activity for healthy aging: longitudinal follow-up and mixed method analysis of a primary care intervention". In: *The Physician and Sportsmedicine* 42.4 (2014), pp. 30–38.

[14] Keith A Kyker and Marian C Limacher. "Gender differences in the presentation and symptoms of coronary artery disease." In: *Current women's health reports* 2.2 (2002), pp. 115–119.

[15] Cliff Lindeman et al. "The extent to which family physicians record their patients' exercise in medical records: a scoping review". In: *BMJ open* 10.2 (2020), e034542.

[16] Fiona Milne et al. "Combatting Sedentary Lifestyles: Can Exercise Prescription in the Emergency Department Lead to Behavioral Change in Patients?" In: *Cureus* 12.2 (2020).

[17] muhammetbektas. "Segmentation of Credit Card Users in Python". In: *GitHub* (2022). URL: https://github.com/muhammetbektas/Unsupervised-Learning/blob/master/Segmentation_of_Credit_Card_Users.ipynb.

[18] Denis Newman-Griffis and Eric Fosler-Lussier. "Automated Coding of under-Studied Medical Concept Domains: Linking Physical Activity Reports to the International Classification of Functioning, Disability, and Health". In: *Frontiers in Digital Health* 3 (2021).

[19] Francesco Pezzella et al. "Blood vessels and cancer much more than just angiogenesis". In: *Cell death discovery* 1.1 (2015), pp. 1–2.

[20] Inc. Pinecone Systems. "Introduction to Facebook AI Similarity Search (Faiss)". In: *Pinecone* (2022). URL: https://www.pinecone.io/learn/faiss-tutorial/.

[21] PubMed. "PubMed". In: *National Library of Medicine* (2023). URL: https://pubmed.ncbi.nlm.nih.gov/.

[22] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[23] Nils Reimers. "all-MiniLM-L6-v2". In: *Hugging Face* (2022). URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[24] Nils Reimers. “Pretrained Models”. In: *SBERT* (2022). URL: https://www.sbert.net/docs/pretrained_models.html.

[25] Physician Resource. “Harnessing the Power of AI for Our Youngest Patients”. In: *UVA Health* (2021). URL: https://www.uvaphysicianresource.com/harnessing-power-ai/.

[26] Robert Sallis et al. “Physical inactivity is associated with a higher risk for severe COVID-19 outcomes: a study in 48 440 adult patients”. In: *British journal of sports medicine* 55.19 (2021), pp. 1099–1105.

[27] Robert E Sallis et al. “The call for a physical activity vital sign in clinical practice”. In: *The American journal of medicine* 129.9 (2016), pp. 903–905.

[28] Cesario de Souza et al. “European Journal of Applied Physics.” In: *European Open Science* (2023). URL: https://www.ej-physics.org/index.php/ejphysics.

[29] Medical Center UCI. “Emergency Medical Services”. In: *UCI Health* (2023). URL: https://www.ucihealth.org/medical-services/emergency-services.

Appendix A: Wordclouds from Clustering

Figure 1 and Figure 2 are the wordclouds generated during data exploration for the best-performing K Means and Spectral clustering algorithms.

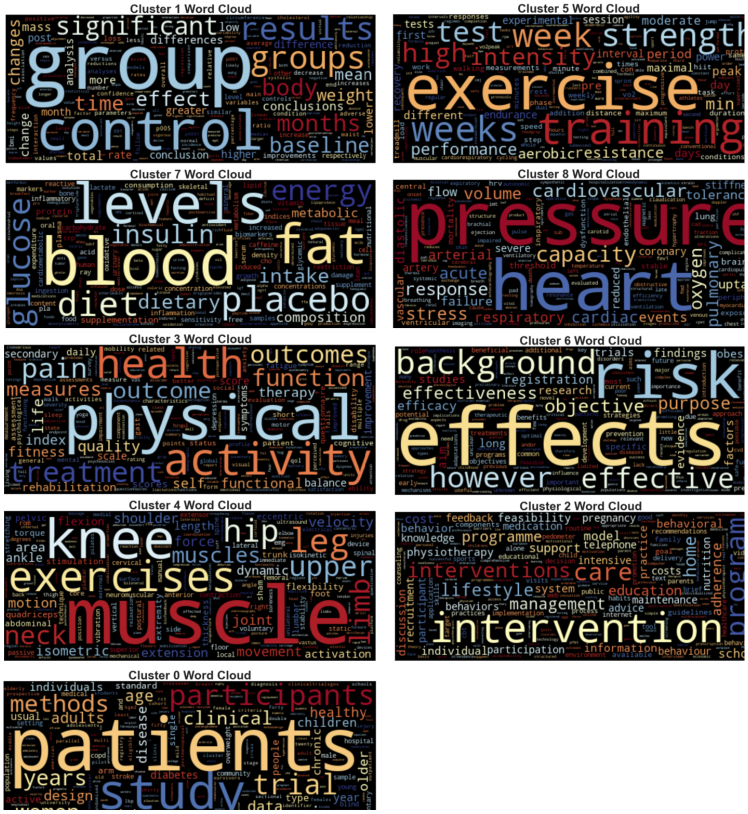


Figure 1: Word clouds generated for each of the 9 clusters of best performing K Means

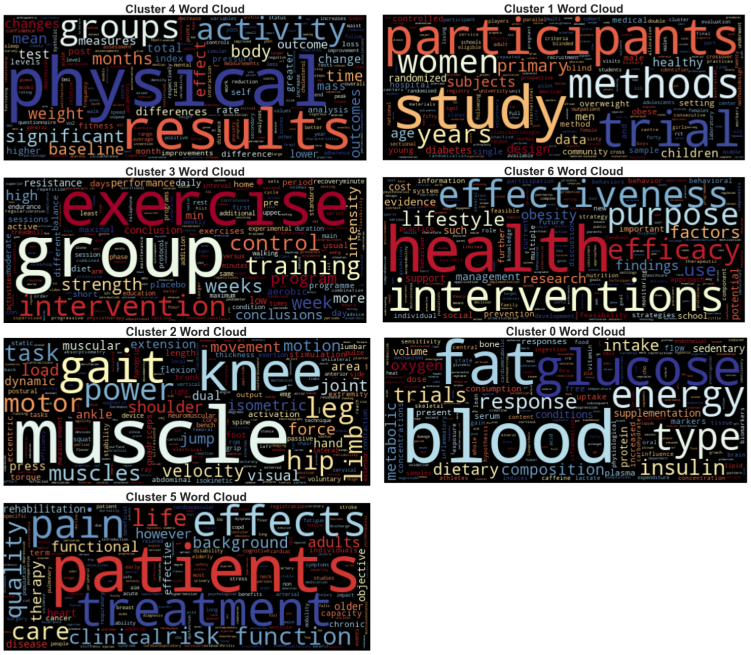


Figure 2: Word clouds generated for each of the 7 clusters of best performing Spectral

Appendix B: Code

Some of the code used throughout our research was based on implementations found in other research articles. We used code snippets from a “Segmentation of Credit Card Users in Python” to test the number of components and which covariance type performed the best for GMMs [17]. The best performing GMM was then generated for our clustering analyses. We used sample code found on Pinecone to implement FAISS for semantic similarity searches [20]. Lastly, Rafael Alvarado wrote the code to import the PubMed articles into Jupyter notebooks [1].