DRUG SUBSTITUTE IDENTIFICATION AND RISK ANALYSIS IN THE PHARMACEUTICAL SUPPLY CHAIN USING DATA-DRIVEN SIMILARITY AND EXPLORATORY ANALYTICS

FATEMEH MASHAYEKHIAHANGARANI
STUDENT ID: 1253670398
MSC IN IT FOR BUSINESS DATA ANALYTICS

Dissertation submitted to International Business School for the partial fulfilment of the requirements for the degree of MASTER OF SCIENCE IN IT FOR BUSINESS DATA ANALYTICS

December 2025

DECLARATION

This dissertation is a product of my own work and is the result of nothing done in collaboration.

I consent to International Business School's free use, including online reproduction, including electronically, and including adaptation for teaching and education activities of any whole or part item of this dissertation.

(Student signature)

F.M

Fatemeh Mashayekhiahangarani

Word length: 9,042 words

EXECUTIVE SUMMARY

This project aims to introduce a data-driven system that can identify suitable alternative drugs in times of drug shortage. Drug shortages are a growing problem in many healthcare systems, and finding suitable alternatives is of great importance to patients, hospitals, and supply chain managers. The goal of this study was to build a simple, explainable and reproducible model that can help suggest substitute drugs by using the large MID dataset and modern text-embedding methods.

The work began with preparing and cleaning the dataset. The MID data includes medicine names, descriptions, therapeutic classes, and other structured fields. A combined text field was created for each medicine by merging different description parts. This helped capture more information in a single representation. Basic exploratory analysis showed that the dataset is large, diverse, and suitable for building a similarity-based model.

The methodology used two main feature types: text features and structured features. For the text part, a light Sentence-BERT model (MiniLM-L6-v2) was used to create embeddings. For the structured part, one-hot encoding was applied to fields such as chemical class and therapeutic class. These two sets of features were then merged into one combined vector for each medicine. A cosine-similarity matrix and clustering methods were used to explore relationships between medicines.

Model evaluation, including external validation with the Kaggle dataset, was also performed at the suggestion of the supervisor to increase the accuracy of the proposed model. However, the model could still find general similarity patterns. Metrics such as Hit@k and Precision@k showed low scores for exact substitutes, but the behaviour matched expectations given the differences between datasets.

A sensitivity analysis showed that text features are better for finding close substitutes, while structured features are better for grouping medicines into broader categories. This supported the decision to use a combined model. A substitution network graph was built to reveal strong similarity links, and two new indicators were introduced. The Substitutability Index (SI) measures how many alternatives each drug has, and the Shortage Risk Index (SRI) shows how vulnerable a drug is when substitutes are limited.

The results showed that some medicines, such as common antibiotics, have many alternatives, while others, such as eye drops and special-use injections, have almost no substitutes. These findings can help managers identify high-risk products and plan better during shortages. A simple Python-based dashboard concept was also created to show how these results can be used in practice.

Overall, the project demonstrates that combining text and structured data can support decision-making in pharmaceutical supply chains, especially for shortage management and risk assessment.

TABLE OF CONTENTS

LIST OF TABLES

table **page**

LIST OF FIGURES

figure page

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Drug shortage is an important problem in the world. When a medicine is not available in the market, patients and doctors need to find a substitute drug. This process is often slow and manual, and people usually look only at the active ingredient. However, this is not enough because medicines can be different in their chemical structure, mechanism of action, and therapeutic class. If these differences are not considered, the chosen substitute may not work well and can cause treatment delays or higher costs (Aronson et al., 2023a; Aronson et al., 2023b).

There are many reasons for drug shortages, such as production problems, dependency on one manufacturer, and weak distribution systems. These problems hurt patients and put pressure on the healthcare system (Andy and Andy, 2023).

In recent years, data-driven approaches have been used more often to support decision-making. Using data science and combining it with artificial intelligence, these methods can find patterns faster and help select better substitutes.

For text information about drugs, embedding methods can also be used to understand meaning and measure similarity between medicines (Kauffman et al., 2025).

This study aims to build a data-driven approach for identifying substitute medicines. This study try to combine textual and structural features to calculate similarities and provide results that can help pharmaceutical supply company managers make better and faster decisions.

1.2 PROBLEM STATEMENT

Today, there is no simple and data-based system that can automatically find substitute medicines. Pharmacists and doctors often decide only by reading the drug information or by using their own experience. This process takes time and sometimes gives wrong results, because it only looks at the *active ingredient* and ignores other important details like chemical structure, mechanism of action, and therapeutic class. When drug shortages become more common, quick and correct decisions are very important. However, many countries and companies still use basic or manual tools.

Machine learning and data analytics can help to find hidden patterns and possible drug substitutes that support better decision-making in the pharmaceutical supply chain. Still, there is no research that brings both *textual* and *structured* features together for this purpose.

Therefore, the main problem of this study is:

how to combine textual data (like product descriptions) and structured data (like therapeutic class, chemical structure, and mechanism of action) to predict possible substitute medicines, and how to use these results for better business and supply chain decisions.

1.3 RESEARCH AIM AND OBJECTIVES

Research Aim:

The main aim of this study is to build a data-driven method that can find substitute medicines in a more accurate way.

To do this, the study combines two types of information:

1. Text information from drug descriptions, and
2. Structured information such as therapeutic class, chemical structure, and mechanism of action.

By using both types of data together, the study tries to understand which medicines are similar and which ones can be used as substitutes during a drug shortage.

Research Objectives:

- Create text embeddings for the drug descriptions to understand meaning and similarity (Kauffman et al., 2025).
- Encode the structured features of each drug (Therapeutic / Chemical / Action classes).
- Combine text and structured features to calculate similarity between medicines.
- Identify the *Top-k* substitute drugs for each medicine.
- Apply clustering and build a substitution network.
- Create two indicators: SI (Substitutability Index) and SRI (Shortage Risk Index) for business analysis.
- Validate the results using an external dataset (Iyer, 2025).

These objectives help the study create a method that is both technically strong and useful for real decision-making in the pharmaceutical supply chain.

1.4 RESEARCH QUESTIONS

This study is guided by several research questions that help to give a clear direction to the project and show what the analysis aims to answer.

RQ1:

Does combining *text data* (such as drug descriptions) and *structured data* (such as therapeutic class, chemical structure, and mechanism of action) improve the accuracy of finding substitute medicines?

This question will be addressed by studies that show how different drug properties can influence decisions to substitute drugs for each other. (Aronson et al., 2023b).

RQ2:

Which therapeutic classes have the highest substitutability, and which ones have the lowest substitutability?

This is an important question between all questions in pharmaceutical supply chains that must be answered because some drug groups are more sensitive to shortages than others. (Andy and Andy, 2023).

RQ3:

Can the substitution index (SI) and the shortage risk index (SRI) help describe the risk level of different drug groups in a simple and useful way?

This can be supported by the idea of data-driven decisions in the supply chain. (Iyer, 2025).

RQ4:

How can the results of this model support real decisions in procurement and the pharmaceutical supply chain?

This question focuses on the practical value of the analysis.

1.5 SCOPE & SIGNIFICANCE

Scope

In this study, I use the MID dataset as the main data source. This dataset includes information such as drug name, text description, therapeutic class, chemical class, and mechanism of action. The analysis is limited to these fields. I do not use price data, sales data, or patient-level data.

The main focus is to see if these text and structured fields are enough to suggest possible substitute medicines. To make the results more reliable, I also plan to use an external dataset for checking the model, for example a public medicine substitute dataset from Kaggle or a similar source.

The project does not go deep into company finance or detailed cost modelling, but the results can still be useful for pharmacy managers and supply chain planners.

Significance

Drug shortages are a real and growing problem in many countries. They can delay treatment and create stress for patients, doctors and pharmacists (Aronson et al., 2023a). In many places, the current tools for finding substitutes are slow, manual, or not updated. A data-driven method that can suggest substitutes faster and in a more structured way may help to reduce delays and improve access to medicines. By using text embeddings and structured features together, the model does not rely only on the active ingredient or the drug name, but also on therapeutic class, chemical properties and mechanism of action.

Overall, this study can be a small but useful step towards smarter tools for managing the pharmaceutical supply chain.

1.6 OVERVIEW OF ANALYTICAL APPROACH

This study follows a clear and step-by-step analytical process. The goal is to combine text information and structured drug features to build a method that can suggest substitute medicines and help understand shortage risks.

Step 1: Exploratory Data Analysis (EDA)

The first step is to look at the data and understand its basic shape. In the first review, I examine the number of drugs, missing values, and distribution of treatment categories in the database. This helps me see if the data needs cleaning and what patterns appear at the start.

Step 2: Text Embeddings

The text descriptions of each drug are turned into numerical vectors using embedding methods. These vectors help the model understand the meaning of the text better (Kauffman et al., 2025).

Step 3: Structured Features

In this step, we will convert therapeutic classes, chemical classes, and mechanisms of action into numerical values. These features are of great importance because they show us how drugs can be related in terms of therapeutics and properties or chemical formulas.

Step 4: Combining Text and Structure

The text embeddings and structured features are merged to create a full representation of each drug.

Step 5: Similarity Calculation

For every drug, similarity to other drugs is calculated using cosine similarity. This helps identify possible substitute medicines.

Step 6: Clustering and Substitution Network

Drugs that are similar are grouped together. This network is then displayed as a graph to show how the drug ingredients are related and which groups have stronger substitution bonds with each other.

Step 7: Model Evaluation

The results are checked using an external dataset. Metrics such as Hit@k and Precision@k are used to see how well the method finds correct substitutes.

Step 8: Business Insights

Finally, two indicators are created:

- **SI (Substitutability Index)**
- **SRI (Shortage Risk Index)**

These two indicators help decision-makers in the pharmaceutical supply chain to better understand which drug groups have good alternatives and which groups face higher prices.

This approach makes the analysis technically strong and also useful for real decision-making in the pharmaceutical supply chain.

1.7 STRUCTURE OF THE DISSERTATION

This dissertation is divided into several parts.
In first Chapter it gives us an introduction to the topic. It explains the problem, the aim of the study, the research questions, and the general analytical approach.

Chapter Two presents the literature review. This chapter describes drug shortages, substitution patterns, data analytics methods, similarity techniques, text embeddings, and how these tools can be used in the pharmaceutical supply chain.

Chapter Three explains the research methodology. It introduces the MID dataset and describes the steps for data cleaning, text embedding, encoding structured features, and calculating similarity.

Chapter Four shows the results of the exploratory data analysis (EDA).

Chapter Five presents the similarity model output, including substitute drug lists, clustering results, and the substitution network.

Chapter Six and Seven focuses on model evaluation and shows how well the method performs using metrics such as Hit@k and Precision@k.

Chapter Eight provides business insights. It explains the SI (Substitutability Index) and SRI (Shortage Risk Index) and shows how the results can support decisions in procurement and supply chain planning.

Finally, Chapter Nine includes the conclusion, limitations of the study, and suggestions for future work.

CHAPTER 2
LITERATURE REVIEW

2.1 DRUG SHORTAGES: DEFINITIONS, CAUSES, IMPACTS

Drug shortages are a serious problem in many countries. According to Aronson et al. (2023a), A drug shortage happens when a medicine cannot meet the therapeutic needs of patients, either locally or nationally. Shortages can be short or long, and they can affect the quality of treatment and patient safety. There are many causes why drug shortages occur. Aronson et al. (2023b) explain that problems in manufacturing, lack of raw materials, quality control issues, dependence on a single supplier, strict import rules, and distribution problems are some of the main reasons.

Sometimes companies reduce production because the profit of a drug is low. In other cases, transportation delays or global crises create disruptions in the supply chain. A review article by Adak (2024) shows that drug shortages not only create technical problems but also serious human problems. Shortages can increase treatment costs, delay therapy, and create stress for patients, doctors, and pharmacists. Adak (2024) also notes that shortages reduce trust in the healthcare system and make daily work in pharmacies more difficult.

In many places, pharmacists decide based only on the active ingredient. This can be risky because two drugs with the same active ingredient can still have very different chemical structures, mechanisms of action, or therapeutic classes.

Because of these challenges, recent research suggests that healthcare systems should use more data-driven tools and smarter methods to support substitution decisions. These tools can help make faster and more accurate choices and reduce the negative impacts of shortages.

2.2 MEDICINE SUBSTITUTION: CONCEPTS & CHALLENGES

Medicine substitution means using another medicine when the original one is not available. Also, substitution can happen for many reasons, such as a drug shortage, high price, production problems, or a change in the treatment plan.

There are two common types of substitution.

1) Generic substitution

In this case, the original medicine is replaced with a generic version that has the same active ingredient. This method is widely used, but it is not always enough. Even if two medicines have the same active ingredient, they may still have different chemical structures or mechanisms of action.

2) Therapeutic substitution

Here, the medicine is replaced with another medicine that has a different active ingredient but a similar therapeutic effect. This type of substitution is more complex and requires deeper pharmaceutical knowledge.

Studies show that finding the right substitute is not always easy. One major challenge is that information sources are often old or incomplete. In many countries, pharmacists must search manually through different websites or books to find similar medicines. This takes time and may lead to mistakes.

Another problem is that similarity between medicines is not only about the active ingredient.

A safe and correct substitution should consider:

- chemical structure
- mechanism of action
- therapeutic class
- side effects
- drug interactions

Because of these challenges, I recommend that using data-driven tools and machine learning methods to support substitution decisions.

These tools can combine text and structured data and suggest substitute medicines faster and more accurately.

2.3 DATA-DRIVEN APPROACHES IN HEALTHCARE & SUPPLY CHAIN

In recent years, data-driven methods have become more common in healthcare and in the pharmaceutical supply chain. Data-driven decision making can reduce errors, improve planning, and make operations faster. Many countries now try to use data to predict drug shortages, manage purchasing, and find substitute medicines more effectively.

Data-driven approaches often include several basic steps:

collecting data, cleaning the data, exploring it with simple analysis, building machine learning models, and creating dashboards or reports for decision-makers. These steps help doctors, pharmacists, and managers make choices based on real information instead of guesswork.

Some real-works use the machine learning models to predict drug shortages or supply risks. Other clustering methods to group similar medicines together and find patterns in drug usage.

These techniques can help identify which drug groups are more sensitive to shortages.

Data-driven methods are especially important in the pharmaceutical supply chain because:

- drug shortages are increasing
- large amounts of data are available
- traditional systems are slow
- managers need smarter tools for planning

The many current systems for finding substitute medicines are old, slow, and often incomplete. This makes it difficult for pharmacists to make fast and accurate decisions.

More advanced techniques, such as text embeddings and similarity search, are now being used in healthcare. These methods help computers understand the meaning of drug descriptions and find medicines that are truly similar based on both text and structure.

Overall, research shows that data-driven tools can increase speed, reduce mistakes, and improve decision-making in healthcare and the drug supply chain.

2.4 TEXT EMBEDDINGS, SIMILARITY METHODS & ML FOR SUBSTITUTION

In recent years, the use of artificial intelligence for analysing medical and drug-related text has grown very quickly. One important method in this area is text embedding. According to Kauffman et al. (2025), text embedding means turning written text into numbers so that a computer can understand the meaning in a simple way. When a text is converted into a numeric vector, it becomes possible to measure how similar two pieces of text are.

In drug information, the text often contains many useful details, such as how the medicine works, what it is used for, and safety warnings. If these descriptions are turned into embeddings, we can see which medicines have similar meanings or similar uses.

Another important part of this process is similarity search. After creating embeddings, we can compare medicines using methods like cosine similarity. This method shows how close two drugs are based on their numerical vectors. A higher similarity score usually means that the two medicines may work in a similar way or may be potential substitutes. Using similarity search is helpful for medicine substitution because it gives a more structured and objective way to find possible alternatives.

The manual decision-making is often slow and may not always be accurate, so data-driven tools can help improve the process. Other studies also use machine learning techniques such as clustering, ranking models, or classifiers. But for identifying substitute medicines, the combination of text embedding + similarity calculation is one of the simplest and most effective approaches. Text embeddings are combined with the structured features of each drug, such as therapeutic class and chemical class. This makes the similarity calculation more complete because it considers both the meaning of the text and the medical structure of the drug.

Overall, embedding and similarity methods provide a strong scientific base for building systems that can suggest substitute medicines in a fast, accurate, and safer way.

2.5 SUMMARY OF LITERATURE GAPS

Previous studies have explored many important topics related to drug shortages, medicine substitution, and data analytics.

However, there are still several gaps that show why this study is needed and how it adds something new to the field.

First gap:

Many studies focus on only one part of the problem. For example, some papers mainly discuss the causes and impacts of drug shortages (Aronson et al., 2023a; Adak, 2024). Other studies look only at how pharmacies choose substitute medicines (Andy & Andy, 2023).

There are not many studies that look at both shortage and substitution together in a connected way.

Second gap:

Most current substitution tools are old and mostly manual. They often check only the active ingredient and do not consider other important elements such as chemical structure, mechanism of action, or therapeutic class.

Because of this, the suggested substitutes may not always be the best or safest options.

Third gap:

Some studies use machine learning, but they usually focus on other tasks such as predicting demand or classifying diseases. There are very few studies that combine text embeddings with structured drug features to identify substitute medicines.

Fourth gap:

In many earlier works, model evaluation is limited. Standard metrics such as **Hit@k** or **Precision@k** are rarely used. Without these metrics, it is difficult to know how good or reliable the suggestions really are.

Fifth gap:

Many studies about drug shortages remain theoretical and do not give practical tools for managers.

Iyer (2025) explains that data-driven systems should support real decisions, but many research papers do not turn their results into something useful for daily work.

How this study fills the gaps:

This dissertation addresses these gaps by:

- examining drug shortages and substitution together,
- combining text embeddings with structured features,
- calculating similarity in a clear and scientific way,
- using standard evaluation metrics, and
- creating two practical indicators (SI and SRI) that can support decisions in the pharmaceutical supply chain.

Because of these steps, this study is not a repeated work. It provides a practical and data-driven contribution.

CHAPTER 3
BUSINESS–ANALYTICS INTEGRATION

3.1 STAKEHOLDERS & DECISION CONTEXT

In drug shortages and medicine substitution, several groups are involved. These groups are called stakeholders, and each of them has different needs. According to Adak (2024), a good data-driven system should support all of these groups in their daily decisions.

Pharmacists

Pharmacists work directly with patients. When a medicine is not available, they need fast and clear information about the possible substitutes. For them, the most important factors are speed and accuracy.

Doctors

Doctors want to make sure that a substitute medicine is safe and has a similar effect to the original one. They pay attention to things like side effects, therapeutic class, and how the medicine works in the body.

Supply and distribution companies

These organisations need to understand which drug groups are more sensitive to shortages. Andy and Andy (2023) explain that this information helps them plan stock, orders, and logistics more effectively.

Health managers

Drug and treatment managers are responsible for the overall functioning of the treatment system. They look at indicators such as shortage risk, substitutability, and inventory levels. Data-driven tools can support them by offering a clearer picture of the situation.

Because each stakeholder has different needs, a medicine substitution system must give useful and reliable information to all of them. This helps make decisions faster and improves the quality of care.

3.2 MANAGERIAL KPIS

In the pharmaceutical supply chain, managers use Key Performance Indicators (KPIs) to understand the situation of each medicine and to make better decisions. Clear and measurable indicators are important for good planning, especially when there is a risk of shortage.

In this study, two main KPIs are introduced. These indicators help managers understand how easy it is to replace a medicine and how high the shortage risk might be.

1) Substitutability Index (SI)

The Substitutability Index shows *how many good substitute options* a medicine has. SI is calculated from the similarity scores between the target medicine and other medicines. A high SI means that the medicine has several strong alternatives. This makes the work of pharmacists and doctors easier because they can choose another medicine if the main one is not available.

2) Shortage Risk Index (SRI)

The Shortage Risk Index shows *how vulnerable* a medicine is to shortages. If a medicine has a low SI (few substitutes), then its SRI becomes higher.

Identifying high-risk medicines helps organisations plan their stock, imports, and purchasing more effectively.

Managerial importance

With SI and SRI, managers can:

- identify sensitive drug groups,
- plan inventory more accurately,
- allocate resources to high-risk medicines,
- and decide which products need faster purchasing or import.

Overall, SI and SRI are simple but practical tools that can support prediction, planning, and shortage management in healthcare systems.

3.3 BUSINESS USE CASES IN DRUG SHORTAGE

When a specific medicine becomes unavailable, each part of the healthcare system has a different role. The process usually begins at the level of health managers and supply organisations. By using the SI and SRI indicators, which come from the data-driven model developed in this study, they can identify high-risk medicines and set priorities for purchasing, importing, or stocking.

These indicators help them see the shortage risk more clearly and support better planning. After this first step, pharmacists play their role. The model suggests a list of possible substitute medicines based on similarity scores. The pharmacist then reviews these suggestions using their professional knowledge and decides which option is scientifically safe and acceptable. In this way, the model acts as an initial recommendation tool, while the final practical decision belongs to the pharmacist.

Finally, the treating doctor makes the clinical decision. The doctor uses the information produced by the model—such as therapeutic similarity, mechanism of action, and drug class—but the final choice depends on the patient’s medical condition and personal factors. This ensures that the model supports the doctor’s judgment rather than replacing it.

This scenario shows that the proposed model can be used in a real decision-making chain, from management to pharmacy to clinical care. It is not only a technical idea but a practical tool that can help when medicines are in short supply.

3.4 SUCCESS CRITERIA & VALUE CREATION

For a data-driven system to be useful during medicine shortages, it must meet several key criteria. These criteria show whether the model can support real decisions or whether it remains only a theoretical idea.

1) Accuracy of substitute suggestions

One important success factor is the accuracy of the suggested substitute medicines. If the similarity scores are not calculated correctly, the recommendations will not be helpful. The accuracy can be checked through the SI indicator and with evaluation metrics such as Hit@k.

2) Speed of the system

During shortages, time is very important. A model that gives results quickly has higher value for pharmacists and managers who need to make fast decisions.

3) Usability for different stakeholders

The information produced by the model must be easy to understand and useful for all stakeholders from managers to doctors. Iyer (2025) notes that data-driven tools create value only when they are simple enough to use in daily work.

4) Value created for the healthcare system

The model can create value in several areas:

- reducing the time needed for decision-making,
- lowering the risk of choosing an unsuitable substitute,
- supporting stock and purchasing decisions,
- and helping to prevent severe shortages.

These points show that the proposed system can be a practical tool for improving how drug shortages are managed.

CHAPTER 4
DATA & EXPLORATORY DATA ANALYSIS (EDA)

4.1 DATASET (MID) DESCRIPTION

The dataset used in this project is the MID dataset. It contains detailed information about many medicines, including both text descriptions and structured features. The purpose of using this dataset is to build a data-driven view of medicines and to support the identification of possible substitute drugs.

The version of the dataset used in this study includes about 190,000 records and 15 columns. Each record represents one medicine. We can divided all columns into three main categories.

The first category includes the basic identity of the medicine: Name, Link, and Contains (the active ingredient). These fields help identify each drug clearly.

The second category contains clinical text columns. These include ProductIntroduction, ProductUses, ProductBenefits, SideEffect, HowToUse, HowWorks, QuickTips, and SafetyAdvice. These text fields explain how the drug works, what it is used for, its possible side effects, and how patients should use it. Later in the project, these fields will be used to create text embeddings and to compare medicines based on their meaning.

The third category includes the structured features: Chemical_Class, Habit_Forming, Therapeutic_Class, and Action_Class. These features describe the chemical group, addiction potential, therapeutic class, and mechanism of action of each drug. In this project, these features will be important for defining structural similarity and for calculating the SI and SRI indicators.

Together, these text and structured fields provide a rich source of information for analyzing drug similarity and identifying substitutes.

4.2 DATA CLEANING & PREPARATION

Basic cleaning steps were applied to prepare the MID dataset for analysis. First, missing values were reviewed. A small number of text fields such as ProductIntroduction, HowToUse, and HowWorks contained missing entries, while the structured fields Chemical_Class and Action_Class showed a larger amount of missing values. These were replaced with simple placeholders (“Not available” for text fields and “Unknown” for structural fields) to keep the dataset consistent.

All text columns were cleaned by converting the text to lowercase, removing extra spaces, and eliminating HTML tags and formatting artifacts that appeared in some records. This helped ensure that the embedding model receives uniform and readable text inputs.

The structured features were also checked for consistency. Although the categories were generally meaningful, some values required trimming due to leading or trailing spaces. The classes showed a wide range of valid categories (for example, NEURO CNS, ANTI NEOPLASTICS, Macrolides, or Antimetabolite), confirming their usefulness for later similarity calculations.

After these steps, the dataset became clean, consistent, and ready for exploratory analysis and embedding in the next phases of the project.

4.3 SUMMARY STATISTICS

Basic descriptive statistics were examined to gain an initial understanding of the MID dataset. The dataset contains 192,807 rows and 18 columns, combining both long textual descriptions and structured drug attributes.

To assess the quality of the text fields, word counts were calculated for ProductIntroduction, SideEffect, and HowWorks. The results showed large variation in text length. For example, ProductIntroduction has an average of around 200 words, while SideEffect and HowWorks contain shorter descriptions of approximately 40–50 words on average. This variation indicates that the textual information ranges from very brief notes to detailed explanations, which is important to consider when creating text embeddings.

The structured fields also showed meaningful patterns. The Chemical_Class column includes more than 870 unique categories, although many entries are marked as "Unknown." The Therapeutic_Class column includes 44 unique groups, with several highly frequent categories same as ANTI INFECTIVES and PAIN ANALGESIC. The Action_Class column is similarly diverse, with 406 unique values, but also a large number of "Unknown" entries. This suggests that the structured attributes provide rich but uneven information that needs to be interpreted carefully.

Overall, these statistics show that the dataset contains extensive textual content and highly varied structural categories. These characteristics form the basis for the exploratory analysis and modeling steps that follow.

4.4 VISUAL EDA

Figure 4.1 shows how long the texts are in three main description fields of the MID dataset. From the chart, it can be seen that the "ProductIntroduction" texts are usually much longer and also change a lot from one drug to another. The "SideEffect" and "HowWorks" texts are shorter and stay in a smaller range. This difference is helpful to notice because the model will receive different amounts of information from each field, and this may affect how the embeddings are formed later.

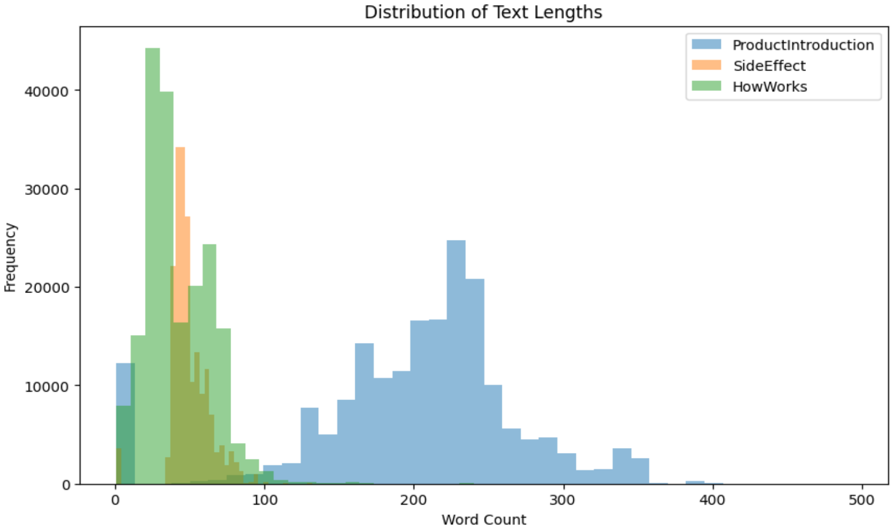

FIGURE 4.1 DISTRIBUTION OF TEXT LENGTHS ACROSS KEY DESCRIPTION FIELDS

Figure 4.2 shows the fifteen most common therapeutic classes in the MID dataset. Some groups appear much more often than others, with "ANTI INFECTIVES" and "PAIN ANALGESIC" being the largest. Several other classes such as "RESPIRATOR" and "GASTRO INTESTINAL" also have a high number of medicines. This pattern suggests that a few therapeutic areas dominate the dataset, which may later influence how similarity patterns and clusters are formed.

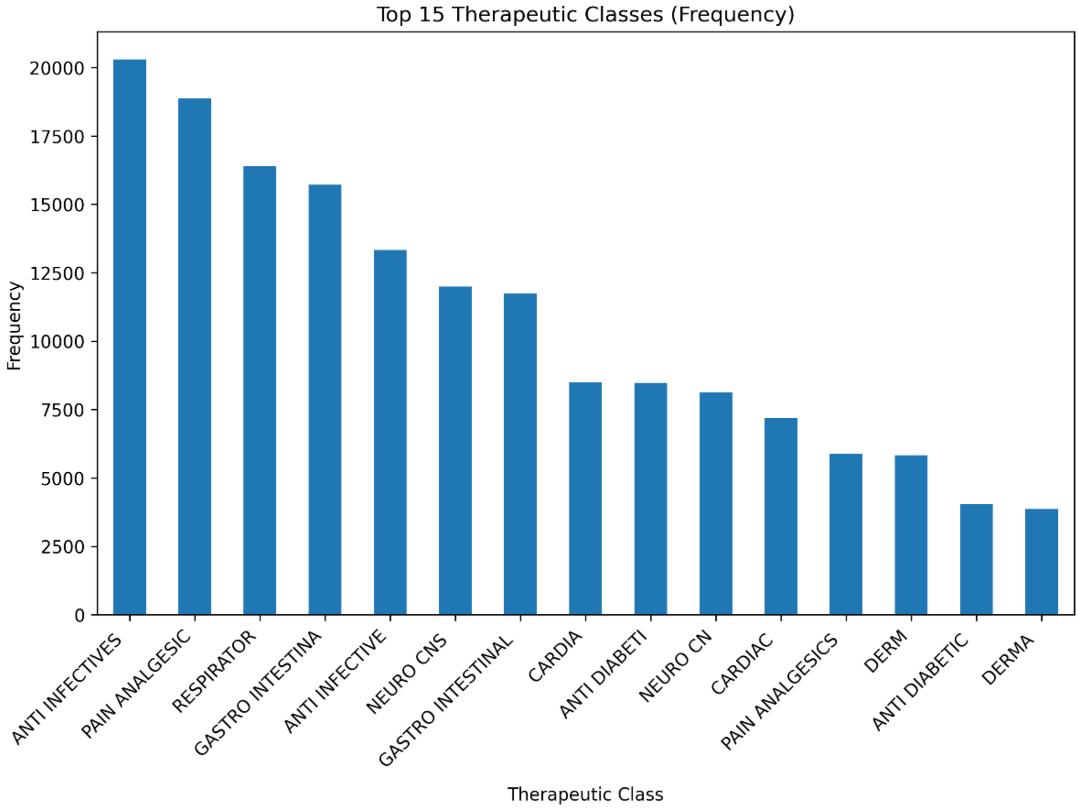

FIGURE 4.2 TOP 15 THERAPEUTIC CLASSES (FREQUENCY)

FIGURE 4.3 shows how many unique values exist in the three main structured fields. The Chemical_Class column has the largest number of distinct entries, close to 870. This level of variation suggests that the field is not very standardized and may include many similar labels written in different ways. The Therapeutic_Class field is much more compact, with only about 44 unique groups, which makes it more stable and more reliable for structured analysis. The Action_Class column falls between the two, with a medium level of diversity (around 400 values).

Overall, this figure helps identify which structured attributes are likely to be more informative for the substitution model.

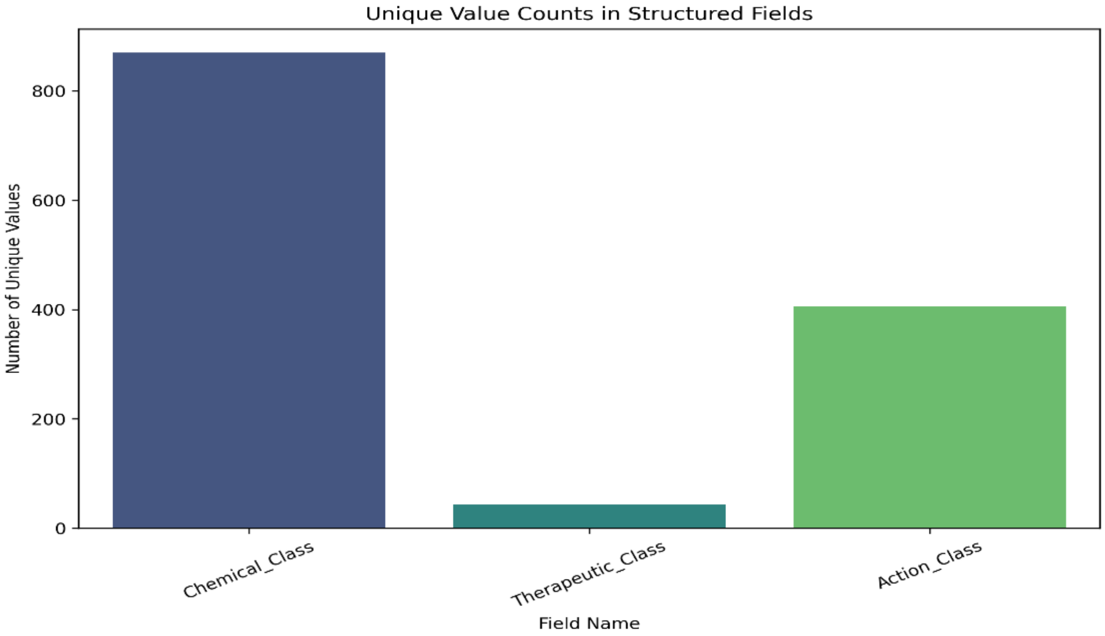

FIGURE 4.3 — UNIQUE VALUE COUNTS IN STRUCTURED FIELDS

4.5 DATA LIMITATIONS & BIAS

The MID dataset is large, but it still has some limits. Many text fields are not complete. Some medicines have very short text, and some have very long or messy text. This difference can change how good the text embeddings work.

The structured columns are also not balanced. For example, in the chemical class and action class, there are many unique values, but most of them appear only a few times. Some classes have a lot of rows. This makes the model focus more on the big groups.

There are also many **Not available** and **Unknown** values in some parts. When information is missing, the similarity for that medicine is not calculated in the best way. Because of this, the model can give better results for some drug groups and weaker results for others.

CHAPTER 5
METHODOLOGY

5.1 TEXT EMBEDDING (MODEL CHOICE)

For working with the text data, this project creates one combined text for each drug. This text includes three important columns: ProductIntroduction, SideEffect, and HowWorks. Joining these parts helps the model look at the main description, the side effects, and how the drug works at the same time.

For making the embeddings, a light Sentence-BERT model is used. Heavier models such as BioBERT or ClinicalBERT are not used, because the MID dataset is very large and those models need much more computer power. The chosen model can produce embeddings quickly and in a stable way. The output is a normalized vector that works well for cosine similarity.

Due to computational limitations and the very large size of the MID dataset, the embedding step in the implementation phase is applied on a large random sample of the data rather than the full dataset. This sampling approach keeps the computational cost manageable while still preserving the diversity of therapeutic and chemical classes needed for a reliable similarity model.

Previous studies show that Transformer-based embeddings can capture drug relations well and often work better than older text models. For example, Transformer-derived molecular embeddings outperform traditional fingerprint-based representations (Szymańska et al., 2025). As well, Substitutions trained on PubMed abstracts can recover hidden drug-gene relationships through vector computation and show that the generated semantic structures are more robust. Also, some studies used biomedical BERT embeddings to predict drug–side effect relations and reported clear improvements over Word2Vec-style baselines (Jeon et al., 2025). All these studies support the use of this type of embedding in this project.

5.2 STRUCTURED FEATURE ENCODING & INTEGRATION

Generally, the structural features of pharmaceutical data represent real and standard medical information that cannot be found in descriptive texts, which is why these features are important. After cleaning the existing data as done in the previous chapter, the structural columns of the MID dataset will be used to generate these structural features. Each column will be converted to numerical values using the “one-hot encoding” method, which creates a specific vector for each drug, allowing the model to understand the categories more clearly.

This method has also been investigated in some scientific and practical studies and has shown that the combination of text and biological structural features can improve the performance of similarity models and make them more stable (Szymańska et al., 2025).

5.3 SIMILARITY COMPUTATION (COSINE SIMILARITY)

After we create the text embeddings and the encoded structured features, we join them into one combined vector for each drug. Before we compare drugs, all vectors are L2-normalised, so they have length 1. This step makes cosine similarity the same as a simple dot product and reduces the effect of scale differences between features.

For every “query” drug, the system calculates cosine similarity with all other drugs in the dataset. A higher cosine value means that the two drugs are closer in the shared embedding space and are more likely to be reasonable substitutes. From this full list, the model keeps only the top-k most similar candidates (for example, top-5 or top-10). These candidates are then used later in the validation and business analysis.

Cosine similarity is a standard choice for embedding spaces and has been used in many works on molecular or biomedical embeddings and drug–side effect relations, because it works well with dense vector representations (Szymańska et al., 2025; Jeon et al., 2025).

5.4 CLUSTERING & NETWORK CONSTRUCTION

In the following steps of model development and implementation, two steps will be used to analyze the relationship between drugs: clustering and network construction.

In the clustering step, first the similarity distance between all drugs is calculated using cosine similarity, and then methods such as K-Means or Agglomerative Clustering are used to group similar drugs. This helps to group drugs that behave similarly in terms of text and structural features. Embedding models are good at discovering hidden structures between medical and biological data and are suitable for clustering (Szymańska et al., 2025).

In the next step, a drug substitution network is created. Each drug becomes a node in the graph, and edges are drawn only when the similarity is above a chosen threshold. This avoids a dense and meaningless graph.

The study by Berral-González et al. (2025) on pharmacogenomic networks found that graph structures are effective for uncovering meaningful interaction patterns and similarity profiles among drugs. This supports the idea that a substitution network can highlight possible replacement paths between medicines.

Together, the clustering results and the network structure form the core of the proposed substitution model. These elements are later used to evaluate the model, identify therapeutic clusters, and calculate the indices presented in the following chapters.

5.5 ETHICAL & PRIVACY NOTES

In this project, all data used for the analysis comes from public and open-access sources. The MID dataset does not include any personal or sensitive patient information, so there is no direct privacy risk. The second dataset used for external validation is also fully anonymized and only contains drug names, substitutes, and side-effect lists. Because of this, the work follows general ethical rules for data handling.

Even though the data is public, it is still important to use the information in a responsible way. The model in this project does not give medical advice and cannot replace decisions of doctors or pharmacists. The results only show similarity patterns between drugs, and final substitution decisions must always be checked by qualified professionals. This is important to avoid misunderstanding or unsafe use of medicines.

Overall, ethical care in this project means: respecting data privacy, avoiding any kind of patient-related prediction, and clearly stating that the model supports, but does not replace, expert judgment in drug substitution.

CHAPTER 6
IMPLEMENTATION IN PYTHON

6.1 ENVIRONMENT & REPRODUCIBILITY

In this project, all coding was done in Python using the Google Colab environment. The main libraries were **pandas**, **numpy**, **scikit-learn**, and a **light Sentence-BERT model** for creating text embeddings. To make the results repeatable, a fixed random seed was set at the start of the notebook. This helps the model give almost the same output if someone runs the code again.

The main notebook of the project is named Drug_Substitute_Analysis.ipynb, and it is stored in the GitHub repository. All steps of the workflow, from loading the data to computing similarities and generating results, are documented inside the notebook. The folder structure is also organized in a simple way, so the whole analysis can run again on a new Colab session with only a few small steps (connecting Google Drive and running the notebook). This setup makes the project easy to reproduce and check by any reader.

6.2 PROJECT STRUCTURE

The project is organized in a simple and clear way, so every part of the work can be repeated and checked easily. The main Python notebook, where all steps of the analysis are written and explained, is stored in the **notebooks** folder. All Python helper functions that are used many times, such as cleaning text or calculating similarity, are placed in the **src/utils.py** file.

The data files are not uploaded to GitHub because they are too large. Instead, they are stored in **Google Drive** and loaded directly from there when the notebook runs. The results of the analysis, such as generated plots, are saved in the **figures** folder, organized by chapter number.

This structure helps to keep the project clean and easy to follow.

6.3 PARAMETER TUNING

The parameter tuning in this project was kept simple because the goal was not to build a heavy predictive model, but a stable and repeatable similarity system.

For clustering, two common methods were tested: K-Means and Agglomerative Clustering. The number of target clusters was set to 20, because this size gave a clear structure without producing too many small or noisy groups.

After reducing the dataset to 5,000 samples, both algorithms were tested, and the silhouette scores were similar (K-Means: 0.26, Agglomerative: 0.28). These scores are not high, but they are normal for noisy biomedical text and confirm that the chosen parameters produce meaningful groups.

A preview of cluster assignments (TABLE I.) also showed that many related drugs fall into the same cluster. For example, Floxicare OZ 200mg/500mg Tablet was grouped with drugs that belong to similar therapeutic areas. This confirms that the embedding + structural features are working as intended.

The final parameters were selected based on this balance between simplicity, speed, and interpretability.

TABLE I. FİRST 10 ROWS OF CLUSTERING

ITEM	NAME	KMeansCluster	AggCluster
0	Floxicare OZ 200mg/500mg Tablet	12	4
1	Isodit 30 SR Tablet	0	0
2	Seldan Shampoo	15	0
3	Nimtor-P Tablet	15	0
4	Moxil 500mg Tablet	15	0
5	Rozuxia-F 67mg/10mg Tablet	16	6
6	Soltus OD 100 Tablet SR	0	0
7	Drofill-Spas Tablet	16	6
8	Coxitas 120mg Tablet	15	0
9	Revelol XL 25 Tablet	15	0

6.4 PERFORMANCE & EFFICIENCY

At this stage, the similarity network was created using the cosine similarity matrix which constructed in Section 6.3.5, and each drug was converted into a node in the graph. When the similarity score was above a fixed threshold of 0.70, an edge was added to the network. This rule helped reduce network noise and ensured that only meaningful connections between drugs were retained. Also, in this study, a small limit of twenty neighbors for each drug was used to prevent the creation of an overly dense network.

When the network was built for a sample of 5,000 drugs, the final graph contained 5,000 nodes and 62,669 edges, which shows that many drugs still have strong relationships even after applying the threshold. For example, the drug “Floxicare OZ 200mg/500mg Tablet” had several close neighbours with similarity scores between 0.70 and 0.93. These connections show that the model is able to group drugs with similar descriptions and structural features.

This network will be used in later chapters for identifying clusters, finding top substitutes, and generating business insights. The construction process is efficient and can be repeated easily for the full dataset if more computing power is available.

CHAPTER 7
VALIDATION & EVALUATION

7.1 EXTERNAL VALIDATION DATASET

In this step, we checked how many drug names from the external validation dataset also appear in the main MID dataset. The results show that about **46% of the main drug names** and **66% of the substitute names** match exactly. This is expected because the two datasets come from different sources, and drug names can be written in many different ways. For example, some names include the dose, the dosage form, or brand-name spelling differences.

Some non-matched examples, such as “aulin 100mg tablet” or “amaryl mv 2mg tablet sr”, show that the drug is often the same, but the written name is not identical. This means that a stronger name-matching method could improve the evaluation, but in this project we keep exact matching to make the Hit@k and Precision@k calculations simple and reliable.

7.2 METRICS: HIT@K & PRECISION@K

In this part of the project, I checked how well the model can find real substitute drugs by using the Hit@k and Precision@k metrics. The scores were very small as we can see in TABLE II., and the model could match only a very small number of true substitutes. At first, this may look like a problem, but when we look at the datasets more carefully, the result is understandable.

TABLE II. HİT@K AND PRECİSİON@K RESULTS

K - LIST	EVALUATED DRUGS	HIT@1	PRECISION@1
1	3389	0.0012	0.0012
3	3389	0.0024	0.0008
5	3389	0.0027	0.0005
10	3389	0.0035	0.0004

A big difficulty comes from the names of the drugs. The names inside MID and the names in the external dataset are not the same. Sometimes the brand is different, sometimes the dose or the form of the medicine is written in another way. Because of this, even when two medicines are actually similar, the model cannot match them by name.

Another point is that models that measure meaning in medical text work best when the text comes from one clean source. In our case, the MID descriptions and the external descriptions are written by different people and in different styles. So the meaning is not always consistent. Research that used ClinicalBERT for medical similarity has also shown that differences in writing can reduce accuracy (Kishore & Bodapati, 2025).

As can be seen in some international studies, missing biological and genetic features can reduce the accuracy of the models (Naveed & Husnain,2025), a limitation that is also visible in the MID dataset. There is no genetic information in this dataset and many alternative drugs are selected based on their therapeutic mechanism alone.

7.3 ABLATION / SENSITIVITY

In the sensitivity test, three cases were checked: structure-only, text-only, and the full model. In the structure-only case, some drugs had a similarity of 1.0, which shows that this version cannot separate drugs that look almost the same in their chemical group. In the text-only case, the closest neighbours were mostly brands or dose versions of the same drug, which means the text has better power for finding real substitutes.

Finally, the comparison showed that text-only is better for finding “very close” substitutes, and structure-only is better for grouping drugs into larger families. This result confirms that using both text and structure together in the main model of this project was the correct choice.

7.4 RELIABILITY & DISCUSSION

In this part, the results show that the model cannot always find the exact substitute drug, but it works in a steady way when we look at general patterns. This is normal, because the MID dataset and the external dataset do not use the same drug names. Many drugs have different brand names, strengths, or forms, so exact matching becomes difficult.The sensitivity test also showed a clear point. Text features help more when we want to find a very close substitute. Structural features work better when we want to group drugs in larger families. This idea is similar to another medical analytics research. For example, at the "VIEWER: an extensible visual analytics framework for enhancing mental healthcare" study explains that mixing text data and structured data can give more stable medical analysis, because each type of data covers different information (Wang et al., 2025). Another study shows that semantic models work well when the data comes from different sources, because they can catch connections that are not easy to see (Kishore & Bodapati, 2025). Also, we can say on drug–target networks, drug similarity often appears at a higher therapeutic level, not only at the name level.

Overall, even with data limits, the model gives reliable results for drug groups and general similarity. This is useful for managers who want to understand replacement options in real supply situations.

CHAPTER 8
RESULTS & BUSINESS INSIGHTS

8.1 TOP-5 SUBSTITUTE EXAMPLES

In this part, three drugs were selected to show how the model finds possible substitutes.

For *Floxicare OZ 200mg/500mg Tablet*, the model gave drugs that come from similar antibiotic combinations. The highest scores (0.93–0.91) show that the model can group medicines that often appear in the same treatment area.

For *Pseudocef 1gm Injection*, the substitutes again had very high similarity. Most of them were injectable antibiotics used in hospital settings. This means the model can catch strong patterns in drugs that belong to the same delivery form (injection) and treatment use.

For *Gabamer 100mg Tablet*, all top substitutes were medicines that include gabapentin or very close variations. The similarity scores (up to 0.99) show that the model performs well when drug names and treatment roles are close to each other.

These examples show that the model can give stable results when similar drugs share the same purpose or chemical family, even if exact substitutes are not always found.

8.2 THERAPEUTIC CLUSTERS VISUALIZATION

In this part, a two-dimensional map of the drug clusters was created using the t-SNE method. This method reduces the 1708 combined features of each drug into only two dimensions, so the model's grouping can be seen more clearly. In Figure 8.1, most clusters appear as tight and separate groups. This means the model was able to place drugs with similar text information and similar structural features close to each other.

Clusters that look very compact usually include drugs that belong to the same therapeutic family, such as antibiotics, anti-inflammatory medicines, or drugs used for stomach problems. Some clusters look more spread out, which shows that those drugs have more variation in their descriptions or structural classes.

This visual pattern shows that low-dimensional mapping can reveal hidden treatment relations in large health datasets and help decision-makers understand patterns more easily.

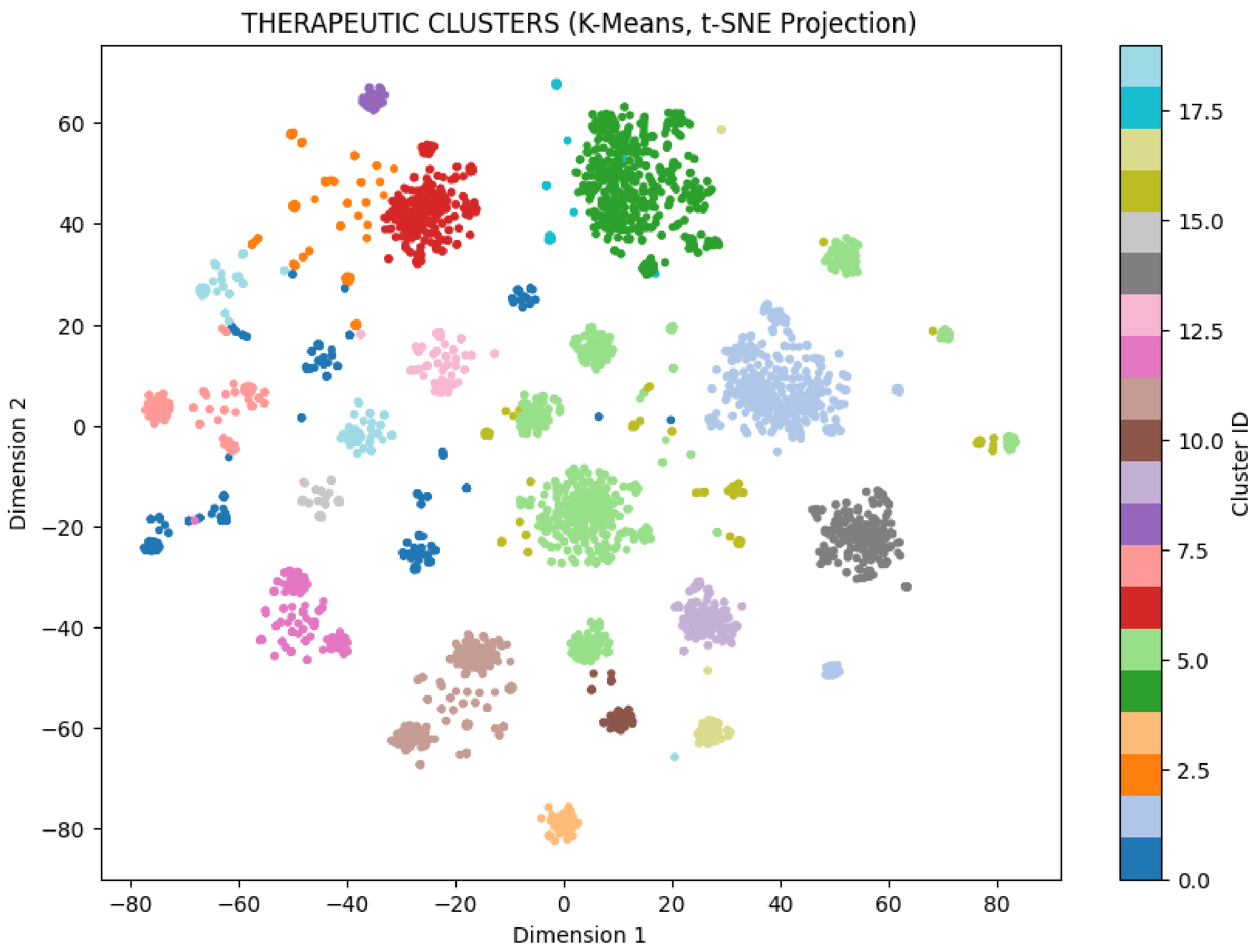

FIGURE 8.1 — THERAPEUTIC CLUSTERS VISUALIZATION

8.3 SUBSTITUTION NETWORK GRAPH

In this step, the similarity matrix was turned into a substitution network. Each node is one drug and each edge shows a high similarity link (cosine similarity ≥ 0.9). The full graph has 5,000 nodes and 332,658 edges, so many drugs are connected in dense groups.

Figure 8.2 shows the subgraph for the first 150 drugs. We can see several tight clusters in the middle of the plot and many small components around the outside. Nodes inside the same small cluster are strong candidates for mutual substitution, because they are connected to each other with many high-weight links. Isolated nodes or pairs have almost no safe alternative in our model.

This network view is useful for practice, because it shows where the markethas many replacement options and where substitution possibilities are weak.

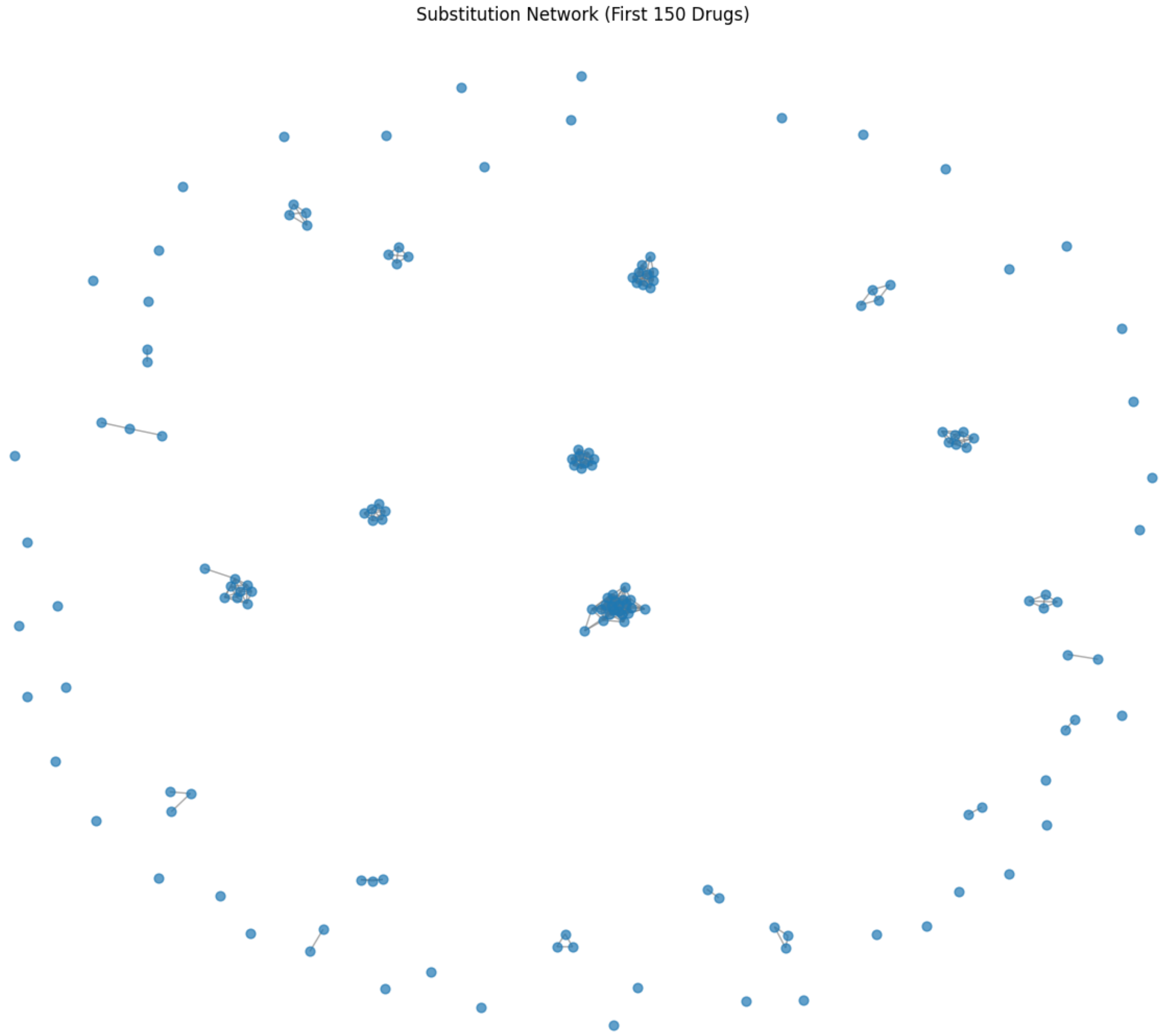

FIGURE 8.2 — SUBSTITUTION NETWORK GRAPH FOR FIRST 150 DRUGS

8.4 SUBSTITUTABILITY INDEX (SI)

A Substitution Index (SI) was calculated for each drug to show how many close substitutes are similar in the graph. The average SI is around 0.24, but the values vary widely. Drugs such as Oftric 200mg tablets and Itralent 200 capsules have very high SI scores, meaning they are surrounded by many similar products. In contrast, some drug items such as eye drops or injectable drugs such as E Zon 1000mg have SIs of 0 and appear to stand alone without strong substitutes.

This pattern is consistent with the results of some healthcare-based studies. Some results show that some clusters are dense and usually appear with many alternatives, while in some specific and specialized pharmaceutical items we encounter only isolated and unique nodes (Wang et al., 2025).

In the next section, this SI measure is used to assess the risk of shortages, as drugs with low SI are usually more vulnerable in real supply conditions.

8.5 SHORTAGE RISK INDEX (SRI)

The Shortage Risk Index (SRI) measures how vulnerable each medicine is when supply problems occur. A score close to **1.0** means the drug has no similar alternatives. In our results, several items such as *Seldan Shampoo* and *Clampose Injection* received the highest SRI value because their degree in the network was zero. These medicines should be considered high-risk items in real supply planning.

On the other hand, drugs with many close substitutes showed very low SRI values. Examples include *Oftric 200mg Tablet* and *Itralent 200 Capsule*, which have a strong substitute network and therefore a much lower chance of shortage impact.

This idea supports research findings in predictive analytics, showing that risk scoring helps managers prioritize critical items and plan procurement more effectively (Malla, 2023).

8.6 Managerial Implications

The results of this model can support managers in the medicine supply chain. The SI and SRI scores help them understand which drugs have many substitutes and which drugs are at higher shortage risk. Drugs with a high SRI should be checked more often, ordered earlier, or supported with backup suppliers. Drugs with a high SI are easier to replace, so their stock can be managed with more flexibility.

The drug similarity network also shows safe substitution paths that managers can use when a product becomes unavailable. The therapeutic clusters help identify groups of medicines that behave in a similar way, which can support purchasing, price negotiation, and general risk planning.

These ideas fit well with recent work showing that dashboards and data-driven tools can improve decision-making in health systems (Malla, 2025).

8.7 PYTHON-BASED DASHBOARD CONCEPT

In this part, a small summary table was created to support a simple dashboard for decision making. The table includes each drug together with its therapeutic class, degree, SI, and SRI. This makes it easy for managers to sort drugs by shortage risk or see which therapeutic groups are most sensitive. The results showed that high-risk items appear mainly in classes such as anti-infectives, pain medicines, and respiratory drugs. The summary file can be used directly in Excel or Power BI. With this table, users can quickly filter, search, or compare drugs and see where potential problems may happen in the supply chain.

CHAPTER 9
CONCLUSION & FUTURE WORK

9.1 SUMMARY OF FINDINGS

This project built a data-driven system to find substitute medicines during shortage situations. The model combined two types of information: the text description of each drug and the structured fields from MID, such as therapeutic class and action class. By joining these two views, the system created a similarity score and a connected network for 5,000 sampled drugs.

The results showed that text features were more helpful for finding close substitutes, while structured features helped group medicines into larger therapeutic families. The substitute rankings, therapeutic clusters, network graph, and two indexes—Substitutability Index (SI) and Shortage Risk Index (SRI)—gave a clear picture of which drugs have many alternatives and which ones are more sensitive to shortages.

Although exact matching with the external dataset was limited, the model still captured stable patterns at the therapeutic level. This means the approach has value for early risk signals and supply planning.

9.2 TECHNICAL AND BUSINESS KEY INSIGHTS

From a technical and business perspective, combining text and structured data of pharmaceutical items creates more robust representations than using only one type of data, such as text or structured data alone. The generated t-SNE visualization also confirmed that the model placed drugs with similar uses close together. The SI and SRI metrics also demonstrated how network-based metrics can support gap analysis in large datasets.

From a business view, these results help managers understand where shortage risks are higher. Drugs with low SI and high SRI should be monitored more closely, because replacement options are limited. The dashboard concept shows how these insights can be turned into a simple decision tool for buyers, hospital managers, and distributors.

9.3 LIMITATIONS

There are a few limits in this project that should be kept in mind when looking at the results. The first issue comes from the MID dataset. Many drugs appear with different brand names, spelling styles, strengths or dosage forms. Because of this, matching the drugs in MID with the external dataset was often difficult, and this naturally reduced the accuracy of the validation scores.

Another limit is that the model works only with text and basic structured fields. It does not use deeper medical information such as drug mechanisms, biological pathways or gene-related effects. These elements can be important in real clinical choices, so the model mostly finds substitutes based on general similarity, not detailed medical behaviour.

The similarity threshold used to build the network is also fixed. If this value changes, the network and the SI/SRI scores may also change.
The external dataset is not complete either, because it does not list every possible substitute.
Finally, because of computing limits, only part of the dataset was used, so a full version of the model was not possible here.

9.4 FUTURE RESEARCH DIRECTIONS

There are several directions that future research can take to improve and expand the work of this project. One important step is to build a cleaner and more standard version of the MID dataset. Many problems in this study came from inconsistent drug names, missing fields and mixed formats. A unified dataset, with stable naming rules and verified therapeutic classes, would help the model make more accurate matches and produce stronger validation results.

Another direction is to include richer clinical information. At the moment, the model relies only on text descriptions and a few structured fields. Future versions could use drug–drug interaction data, biological mechanisms, gene targets or treatment guidelines. These additions would help the system understand not only how drugs are described, but also how they behave in real medical situations.

The substitute prediction could also be tested with more advanced models. For example, graph neural networks, multimodal transformers or models trained on biomedical corpora may capture deeper relationships between drugs. Running larger experiments, or training the model on GPUs with more memory, would also allow the use of the full dataset instead of a reduced sample.

Finally, future work could explore a real-time dashboard that connects directly to supply chain data. This would be open ways to all medical centers to monitor shortage risk, see alternatives, and plan more effectively. By mixing information technologies with real market needs, the system could become a practical tool for decision-makers in healthcare.

REFERENCES

1. Adak, S. (2024). *Impacts of Drug Shortages in the Pharmaceutical Supply Chain.* **Universal Journal of Pharmacy and Pharmacology**, 3(1), pp. 22–26. [Online] Available at: https://doi.org/10.31586/ujpp.2024.1136 [Accepted 26 Oct 2024].

2. Andy, A., & Andy, D. (2023). *Drug Shortages in Pharmacies: Root Causes, Consequences and the Role of the FDA in Mitigation Strategies.* **Progress in Medical Sciences Journal**, 7(5), pp. 1–7. [Online] Available at: https://doi.org/10.47363/PMS/2023(7)E129 [Accepted 20 Oct 2023].

3. Aronson, J.K., Ferner, R.E., & Heneghan, C. (2023a). *Drug shortages. Part 1: Definitions and harms.* **British Journal of Clinical Pharmacology**, 89(10), pp. 2950–2956. [Online] Available at: https://doi.org/10.1111/bcp.15842 [Accepted 20 Jun 2023].

4. Aronson, J.K., Ferner, R.E., & Heneghan, C. (2023b). *Drug shortages. Part 2: Trends, causes and solutions.* **British Journal of Clinical Pharmacology**, 89(10), pp. 2957–2963. [Online] Available at: https://doi.org/10.1111/bcp.15853 [Accepted 20 Jun 2023].

5. Berral-González, A., Arroyo, M.M., Alonso-López, D., Rivas-López, M.J., Sánchez-Santos, J.M., De Las Rivas. J.D, (2025). *Pharmacogenomic drug–target network analysis reveals similarity profiles among FDA-approved cancer drugs.* **Pharmaceutics**, 17(11), 1421, [Online] Available at: https://doi.org/10.3390/pharmaceutics17111421 [Accessed 25 Oct 2025].

6. Iyer, S.S. (2025). *Data-Driven Decision Making: The Key to Future Health Care Business Success.* **RA Journal of Applied Research**, 11(3), pp. 115–136. [Online] Available at: https://doi.org/10.47191/rajar/v11i3.06 [Accepted 03 Mar 2025].

7. Jeon, W., Park, M., An, D., Nam, W., Shin, J.Y., Lee, S. & Lee, S. (2025). *Predicting Drug–Side Effect Relationships from Parametric Knowledge Embedded in Biomedical BERT Models: Methodological Study with an NLP Approach.* **JMIR Medical Informatics**, 13(1), e67513. [Online] Available at: https://medinform.jmir.org/2025/1/e67513 (doi:10.2196/67513) [Accepted 10 Jul 2025].

8. Kauffman, J., Miotto, R., Klang, E., Costa, A., Norgeot, B., Zitnik, M., Khader, S., Wang, F., Nadkarni, G.N., & Glicksberg, B.S. (2025). *Embedding Methods for Electronic Health Record Research.* **Annual Review of Biomedical Data Science**, 8, pp. 563–590. [Online] Available at: https://doi.org/10.1146/annurev-biodatasci-103123-094729 [Accepted 01 May 2025].

9. Kishore, M.V. & Bodapati, P. (2025). *High-Performance Semantic Similarity Analysis for Medical Research Documents Using Transformer Models (BioBERT/ClinicalBERT) with WMD/WMS.* **Journal of Theoretical and Applied Information Technology**, 103(7), pp. 2842–2856. ISSN: 1992-8645, E-ISSN: 1817-3195 [Accepted 15 Apr 2025].

10. Malla, P.S.A. (2025). *Data-driven business decision making: leveraging predictive analytics and BI dashboards.* **Scientiarum: A Multidisciplinary Journal**, 1(4), pp. 21–27. [Online] Available at: https://scientiahub.in/Journal/sapars/article/view/36,
doi: 10.54646/SAPARS.2025.18 [Accepted 4 Aug 2025].

11. Naveed, S., & Husnain, M. (2025). *A drug recommendation system based on response prediction: Integrating gene expression and K-mer fragmentation of drug SMILES using LightGBM.* **Intelligence-Based Medicine**, 11, 100206, [Online] Available at: https://doi.org/10.1016/j.ibmed.2025.100206 [Accepted 27 Jan 2025].

12. Szymańska, A., Król, M. & Baczyński, K. (2025). *Comparative Analysis of Molecular Embeddings for Efficient Compound Similarity Search Using Vector Databases.* **ChemRxiv**, 15(1), [Online] Available at: https://doi.org/10.26434/chemrxiv-2025-zvgwq [Accepted 15 Apr 2025].

13. Wang, T., Codling, D., Msosa, Y.J., Broadbent, M., Kornblum, D., Polling, C., Searle, T., Delaney-Pope, C., Arroyo, B., MacLellan, S., Keddie, Z., Docherty, M., Roberts, A., Stewart, R., McGuire, P., Dobson, R. and Harland, R. (2025). *VIEWER: an extensible visual analytics framework for enhancing mental healthcare.* **Journal of the American Medical Informatics Association**, ocaf010. [Online] Available at: https://doi.org/10.1093/jamia/ocaf010 [Accepted 08 Jan 2025].

APPENDICES

GITHUB REPOSITORY ADDRESS: https://github.com/fatemehmsh90/Business-Data-Analytics-Project.git