DEVELOPING A SYSTEM ON THE EXPLORATION OF RELATIONSHIPS BETWEEN BIOMEDICAL ASSETS THROUGH BIOMEDICAL ARTICLES

M.Sc. THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
OSMAN ALTUNER
January 2025

DEVELOPING A SYSTEM ON THE EXPLORATION OF RELATIONSHIPS BETWEEN BIOMEDICAL ASSETS THROUGH BIOMEDICAL ARTICLES

M.Sc. THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
OSMAN ALTUNER
January 2025

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Osman Altuner
Signature:

REGULATORY COMPLIANCE

M.Sc. thesis titled **"DEVELOPING A SYSTEM ON THE EXPLORATION OF RELATIONSHIPS BETWEEN BIOMEDICAL ASSETS THROUGH BIOMEDICAL ARTICLES"** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor	Co Advisor
Osman Altuner	Assoc. Prof. Burcu Bakır Güngör	Asst. Prof. Mehmet Gökhan Bakal

Head of the Electrical and Computer Engineering Graduate Program

Asst. Prof. Samet GÜLER

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled “DEVELOPING A SYSTEM ON THE EXPLORATION OF RELATIONSHIPS BETWEEN BIOMEDICAL ASSETS THROUGH BIOMEDICAL ARTICLES” and prepared by Osman Altuner has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

..…/…../……..

JURY:

Advisor : Assoc. Prof. Burcu Bakır Güngör

Member : Assoc. Prof. Rifat Kurban

Member : Asst. Prof. Tayyip Özcan

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / …….. and numbered ……..…… .

…….. /…….. / ……..

(Date)

Graduate School Dean
Prof. İrfan ALAN

ABSTRACT

DEVELOPING A SYSTEM ON THE EXPLORATION OF RELATIONSHIPS BETWEEN BIOMEDICAL ASSETS THROUGH BIOMEDICAL ARTICLES

Osman Altuner
MSc. in Electrical and Computer Engineering
Advisor: Assoc. Prof. Burcu Bakır Güngör
Co-advisor: Asst. Prof. Mehmet Gökhan Bakal
January 2025

In today's world, digitalization is spreading rapidly. This proliferation, while making our lives easier on the one hand, also brings new challenges, such as the analysis and processing of large amounts of digital data on the other. This is particularly evident in the context of academic research. Academic research necessitates sophisticated evaluation procedures.

In this context, it is known that research conducted on diseases should be evaluated effectively. In conclusion, in this study, publications related to diseases were subjected to text analysis methods and then transformed into a network structure that allows the association of data with significant biomedical connections. It is aimed at examining the complex network structure of two biomedical entities that have important connections, such as therapeutic and causative.

In this case, it has been confirmed that the entity pairs obtained through manual search methods are real connections. This study has successfully resolved the often time-consuming manual search process for locating existing known biomedical entities. Additionally, through this method, there is potential to discover unknown or yet-to-be-explored possible new relationships (therapeutic, causative, etc.) via multiple binary connection patterns.

In conclusion, the combination of techniques such as graph analysis, knowledge discovery, and text mining is leading to the discovery of potentially significant new findings in biomedical research.

Keywords: text mining, knowledge discovery, graph analysis, neo4j

ÖZET

BİYOMEDİKAL VARLIKLAR ARASINDAKİ İLİŞKİLERİN BİYOMEDİKAL MAKALELER ARACILIĞIYLA KEŞFEDİLMESİNE DAİR BİR SİSTEM GELİŞTİRİLMESİ

Osman Altuner
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans
Tez Danışmanı: Doç. Dr. Burcu Bakır Güngör
İkinci Tez Danışmanı: Dr. Öğr. Üyesi Mehmet Gökhan Bakal
Ocak 2025

Günümüz dünyasında dijitalleşme hızla yayılmaktadır. Bu yayılma, bir yandan hayatımızı kolaylaştırırken diğer yandan büyük miktarda dijital verinin analizi ve işlenmesi gibi yeni zorlukları da beraberinde getirmektedir. Bu durum özellikle akademik araştırmalar bağlamında belirgindir. Akademik araştırmalar, gelişmiş değerlendirme süreçlerine ihtiyaç duymaktadır.

Bu bağlamda, hastalıklar üzerine yapılan araştırmaların etkili bir şekilde değerlendirilmesi gerektiği bilinmektedir. Bu çalışmada, hastalıklarla ilgili yayınlar metin analizi yöntemlerine tabi tutulmuş ve ardından verilerin önemli biyomedikal bağlantılarla ilişkilendirilmesini sağlayan bir ağ yapısına dönüştürülmüştür. Amaç, tedavi edici ve sebep verici gibi önemli bağlantılara sahip iki biyomedikal varlığın karmaşık ağ yapısını incelemektir. Bu durumda, manuel arama yöntemleriyle elde edilen varlık ikililerinin gerçek bağlantılar olduğu doğrulanmıştır.

Bu çalışma, mevcut bilinen biyomedikal varlıkların bulunmasında sıklıkla zaman alan manuel arama sürecini başarıyla çözmüştür. Ayrıca, bu yöntem sayesinde birden fazla ikili bağlantı örüntüsü aracılığıyla bilinmeyen veya henüz keşfedilmemiş olası yeni ilişkilerin (tedavi edici, sebep verici vb.) keşfedilme potansiyeli bulunmaktadır.

Sonuç olarak, çizge analizi, bilgi keşfi ve metin madenciliği gibi tekniklerin bir araya getirilmesi, biyomedikal araştırmalarda potansiyel olarak önemli yeni sonuçların keşfedilmesine yol açmaktadır.

Anahtar kelimeler: metin madenciliği, bilgi keşfi, çizge analizi, neo4j

Acknowledgements

Firstly, I would like to express my gratitude to my thesis advisor, Assoc. Prof. Burcu Bakır Güngör, for considering my requests, helping me choose my thesis topic, and for her patience and constructive attitude throughout the process.

I would like to express my sincere appreciation to my thesis co-advisor, Asst. Prof. Mehmet Gökhan Bakal. His support, patience, leadership in guiding me through challenges, and expertise greatly contributed to my master's process and the completion of this thesis.

I would also like to thank all the academic and administrative staff at Abdullah Gül University who contributed to my master's education.

I would like to extend my gratitude to my thesis defense committee members, Assoc. Prof. Rifat Kurban and Asst. Prof. Prof. Tayyip Özcan for their invaluable help and guidance during the evaluation of my thesis.

Moreover, I express my heartfelt thanks to my precious family, especially my mother Şengül Altuner, my father Adem Altuner, and my sister Asst. Prof. Gül Şebnem Altuner Çoban for their unwavering support, sacrifices, belief in my abilities, and encouragement throughout my educational journey.

Additionally, I would like to offer my special thanks to Ayça Aslı Minen, who has supported me in every aspect of life since the day she entered it, stood by me in difficult times, and is one of the architects of my success thanks to her encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interfaces
CUI	Concept Unique Identifier
NLM	United States National Library of Medicine
NLTK	Natural Language Toolkit
NLP	Natural Language Processing
SemMedDB	Semantic MEDLINE Database
SQL	Structured Query Language
UMLS	Unified Medical Language System
XML	Extensible Markup Language

To my family

Chapter 1

Introduction

The rapidly evolving technologies of today are fundamentally restructuring the life practices of individuals and institutions, as well as the ways in which they access and share information. One of the most important concepts in this transformation is digital transformation, which provides significant advantages in every aspect of life. The digitalization process offers great potential, especially in the healthcare sector. It creates new opportunities both in making healthcare services more efficient and in the collection and analysis of personal health data. These days, a lot of unstructured data is being created, particularly in the healthcare industry, even while the amount of text-based data generated in the digital environment is growing quickly [1]. Texts of various kinds, including scholarly articles, clinical patient records, web pages, online purchasing records, and social media posts, are included in this data. Smartwatches and other gadgets that offer personal health data also add to the daily generation of data [2]. It is challenging to assess the growing amount of text-based data, which emphasizes the value of computational techniques and automated analysis tools. As a result, text mining has emerged as a crucial instrument in the healthcare industry for deriving valuable information from unstructured texts. Effective analyses in disease diagnosis, treatment, and public health research are made possible by the application of computational techniques like text mining and machine learning in the healthcare industry [3, 4]. By making it possible to analyze health data more effectively, the fields of text mining and bioinformatics aid in the creation of novel therapeutic approaches and improved diagnostic strategies. In order to extract meaning from the vast volume of unstructured data in the healthcare industry, experts are turning to computational techniques like text mining.

Two particularly notable multidisciplinary areas in the healthcare industry are bioinformatics and text mining. While text mining finds valuable information by examining unstructured text data, bioinformatics makes it possible to analyze biological

and genetic data and turn it into scientific conclusions. Combining these two disciplines to analyze health data, for instance, improves decision support for disease diagnosis procedures and is crucial for treatment process optimization. In this regard, fields like text mining and bioinformatics enable a deeper examination of the available disease data and the identification of novel connections.

Studies of select conditions have explored this analysis leveraging open access materials in PubMed [5] , a digital library of life sciences literature. SemMedDB's triples informed scrutinizing sourced reports regarding each sickness using Neo4j's graph database abilities. Afterwards, prominently occurring phrases extracted from these relations served to compile target ailment network representations. Alternately, deeper investigations into particular disorders uncovered supplementary ties among qualifying publications through data-mined connections and relationships within the curated corpus [6]. These analyses have facilitated a better understanding of disease-related network structures and have enabled the discovery of new relationships in line with the research objectives. Information networks that can be used in the diagnosis and treatment of diseases have been supported by the analyses conducted, and the results obtained have confirmed existing relationships while also demonstrating the accuracy of the methods used.

The structure of the research is presented in the following manner: The second section examines pertinent literature regarding the subject matter, whereas the third section elaborates on the research design, methodologies, instruments, and techniques utilized. The fourth chapter showcases the results obtained from the implemented methods. In conclusion, the final chapter provides an in-depth overview of the research findings and suggests avenues for further investigation.

In conclusion, the combined use of computational methods and text mining in healthcare enables new discoveries in the field. As a result of this study, the potential of the combined use of these two disciplines in disease diagnosis and treatment processes has been better understood. In future studies, analyses conducted on larger disease cohorts will provide a broader perspective that enhances the effectiveness of healthcare services and contributes to achieving more advanced results in health research.

Chapter 2

Literature Review

When conducting a literature review regarding the existence of studies similar to the planned one, it was found that there are indeed similar studies. In their study, Altuner et al. (2024) analyze academic works from the PubMed database for biomedical knowledge discovery, utilizing the Neo4j graph database to identify specific terms related to target diseases and match them with CUI (Concept Unique Identifiers). Using text mining and graph-based analysis techniques, the study efficiently validated terms and relationships associated with target diseases, uncovering complex connections. The method used by Altuner et al. differs in that it quickly confirms common genetic links found in the literature and uncovers unknown or understudied possible connections. While prior investigations into extracting meaningful connections from expansive datasets were often laborious and slow, Altuner et al.'s novel technique demonstrates promising precision and swiftness in pinpointing notable correlations. Not only does their methodology present a useful substitution for hand-picked information discovery methods in biomedical analysis, but it also helps advance accessibility to relationships documented in biomedical publications via network-centric examinations. Furthermore, the study relieves some of the workload traditionally involved in manually scouring literature and synthesizing relationships, thereby serving as a helpful tool for researchers [7].

The creation and efficacy of biomedical relationship extraction (RE) models are examined by Zhang et al. (2023), with a focus on COVID-19. The authors stress the importance of RE models in biological text mining, especially for knowledge graph construction and information discovery. Using datasets containing 125 different types of connections, the study analyzed the performance of various models and found that pre-trained models such as PubMedBERT were very good at relation extraction tasks. To comprehend the biological significance and use of RE, the writers studied the COVID-19 literature. After that, they created a database of association graphs that showed present

and prospective drug routes. The findings highlighted biomedical RE model optimization, offering a potent instrument for comprehensive text mining and information discovery [8].

Effective biological information extraction necessitates investigating varied text mining algorithms, as Krallinger et al. demonstrate in 2009. Their work revealed how techniques can uncover protein interactions, metabolic processes, and signal transduction pathways. A range of technologies for capturing and categorizing biomedical data were probed in this inquiry. Resources leveraged to analyze life processes incorporated lexical and ontological sources as well as databases of literature. Case studies depicted fundamental approaches for correlating biology, like cancer-related proteins. The probe also discussed uses like extracting disease-gene links, identifying biomarkers, and mining mutation, SNP, and epigenetic data from available works. The examination underscored text mining systems' importance in cancer exploration and how scientists can better access knowledge employing tools supplementing regular literature review methods [9].

A thorough analysis of cutting-edge text mining techniques in bioinformatics was extensively appraised by Qi et al. in 2009. Their wide-ranging study delved deeply into several key subprocesses for example knowledge discovery from texts and information extraction, as well as relevant technologies like data mining and natural language processing. Besides emphasizing select applications involving named entity recognition and microarray data analysis, the investigation concentrated on important themes within bioinformatics like gene expression profiling, gene function determination, and systems-level comprehension of biology. The authors stressed the value of text mining methods for accelerating information retrieval and assisting with the discovery of novel relationships in life science research. In comparison with traditional approaches, this work increases efficiency and presents an intelligent strategy for handling the rapidly growing volumes of data in bioinformatics. Unlike prior examinations, Qi et al's extensive survey offered a comprehensive overview of the diverse analytical domains within the bioinformatics data landscape and proposed significant recommendations to enhance the systematic process of scientific discovery. By appraising contemporary text mining tools identified in the literature, this study highlighted how text analytics can reveal meaningful insights from biomedical knowledge sources [10].

The information density within biomedical literature is examined by Krallinger et al. in their 2005 study, exploring applications of text mining and natural language processing techniques within domains of biomedicine and molecular biology. Their research emphasized text mining's effectiveness in bioinformatics, particularly identifying biological entities, interpreting microarray data, and deriving protein interactions. The authors stressed the importance of knowledge discovery and information extraction tools for examining biological literature and biomedical databases. Through case studies, they exhibited how these technologies are used in processes including annotating gene and protein names, identifying linkages in literature, and constructing networks of protein interactions grounded in literature.

Despite the perplexing complexity of biomedical writing, text mining has proven an effective tactic for characterizing gene and protein roles, assisting with identifying chemical compounds tied to disease, and employing information extraction to uncover possible new therapeutic targets. By underscoring how platforms including BioCreative and TREC have increased text mining algorithm accuracy and efficacy, Krallinger highlighted the value of community-based evaluation systems for assessing these technologies, as described in their report [11].

Hoksza and Jelínek's 2015 investigation delved into both the troubles and rewards of leveraging Neo4j's graph database to uncover protein–protein interactions. The researchers aimed to pinpoint the amino acid series in protein structures contributing to interactions between proteins. They transformed the protein structural information within the Protein Data Bank into a graph database using Neo4j, where each protein was envisioned as an independent element inside the network. While running the interaction detection process through Neo4j, the academics discovered performance issues managing large informational constructs during queries.

In order to identify zones of amino acid engagement through a Conditional Random Field technique, the study further depended on additional knowledge within the graph database containing up-to-the-moment protein structure data. The precision of interaction prediction may be enhanced by this information-driven supplement to the CRF-based approach. In closing, it was established that whereas Neo4j can be helpful for specific small-scale graphical questions, it is pivotal to consider performance limitations when dealing with more substantial query networks [12].

In order to find new gene targets linked to complicated genetic illnesses like osteoporosis, Gajendran et al. (2007) created a text mining program to examine biomedical literature on PubMed. This program generates a ranking and network structure using an algorithm based on the co-occurrence principle of genes. The impact factors of the papers that cite the genes are used to assess their significance within the genomic network. Finding possible novel targets associated with osteoporosis and identifying genes like SMAD proteins that are essential for treatment are two benefits of this investigation. By highlighting genes that are often cited and providing a useful technique for finding novel target genes, Gajendran et al.'s methodology closes knowledge gaps in the literature in comparison to other studies. By offering high-accuracy data screening techniques, this study speeds up and improves the reliability of biological knowledge discovery in comparison to manual methods [13].

Chapter 3

Materials And Methods

3.1 The Concept of Data

Data serves as a depiction of a specific circumstance, occurrence, or information point. It embodies a representation of a condition, concept, or idea, and can manifest as a number, word, or visual image. The data that illustrate the qualitative or quantitative attributes of a variable or several variables consist of collections of numbers, words, or images that, in isolation, lack inherent meaning [14]. However, although the dictionary meaning of data is “fact,” it does not always represent concrete facts. Sometimes, data is not definitive or is used to describe things that have never happened, such as an idea.

Data is generally seen as simple facts that are structured to be converted into information. When information is evaluated or interpreted within a certain context or given meaning, it transforms into knowledge. Although there are various approaches on this subject, the general consensus is that data is less than information, and information is less than knowledge. Moreover, it is accepted that to obtain information, one must first have data, and knowledge only emerges with the existence of information [15].

In this context, while information exhibits a more independent and unique characteristic, knowledge is personalized and has a cognitive, human dimension. At the same time, information is defined as something that people collect, possess, transmit, include in databases, record, accumulate, count, and compare, and which is transmitted through electronic networks and information technologies. In contrast, knowledge is not suitable for actions like accumulating and counting; it requires assimilation and interpretation by the individual.

3.2 Database Models

Humankind started recording information long ago, and detailed database systems from ancient times were developed by government offices, libraries, hospitals, and various businesses. Some of the fundamental concepts and principles in the formation of these systems are still in use today. The separation of the database schema from the physical information record has become a standard principle of database systems. With the rise of computer usage in the 1980s and the commercialization of relational databases, SQL (Structured Query Language) became a global standard. During this period, the concept of object-oriented databases also emerged. In the object-oriented database model, primarily used in object-oriented fields, object databases exist alongside more extensive database management systems, unlike relational databases. The emergence of object-oriented databases made it possible to reduce the number of relationships by creating objects. However, the biggest challenge with object-oriented databases was converting existing databases into object-oriented ones. These problems were addressed with certain programming languages and APIs (Application Programming Interfaces), but this reduced the flexibility of the database.

As a solution to this issue, the object-relational database model, combining the advantages of relational and object-oriented databases, was developed in the early 1990s. The rise of the internet in the mid-1990s made remote access to existing data via computer systems possible. Over the closing stages of the twentieth century, significant developments in digital technologies transpired, including the conception of XML. Designed as a meta-language with defined rules for encoding documents comprehensible to both humans and machines, XML debuted in 1997. The application of XML to database processing offered solutions to long-standing issues through its simplicity, universality, and utility across online platforms. With each new year, advances in database technologies continued at a steady clip, constantly rendering more dynamic programs on which we had grown reliant. However, human steering or outside influences sometimes remained indispensable in refining information and determining outcomes. Those initial steps down this nascent path commenced in the infant years of the new millennium, with the journey still ongoing even today. Now standard practices like excavating insights from accumulated masses of facts and aggregating organized collections have taken root in contemporary times. Open-source NoSQL, which emerged

in 1998 and is not tied to a SQL interface, disregarded the core principles of traditional relational databases such as atomicity, consistency, isolation, and durability (ACID), leading to the rise of non-relational distributed data stores throughout the 2000s [14].

3.3 Database Management Systems

Databases serve as structured systems for securely and efficiently organizing, storing, accessing, modifying and deleting information in an organized manner. This ordered arrangement guarantees that data is kept up-to-date, intact and full, notwithstanding its volume, extent or perplexity. Additionally, databases guarantee that data is effectively leveraged to build overall productivity by facilitating querying, documenting and investigating information. Key parts incorporate information records and administrations, complex data arrangements, overseers, security layers, reinforcement and recovery instruments just as enhancement components. Furthermore, databases ensure that data remains easily retrievable while playing an integral role in strategic decision making through reporting and analytics. The complexity and scale of today's digital landscape underscores the importance of robust, carefully designed database infrastructure to get the most insight and value from increasingly vast stores of disparate data.

SQL (Structured Query Language) is commonly used to query, insert, update, and delete data in databases. Other query languages include PL/SQL (Procedural Language), TCL (Transaction Control Language), and Transact-SQL. Some of the most popular databases today include Oracle Database, Informix, Microsoft Access, dBASE, Firebird, Microsoft SQL Server, MySQL, Microsoft Visual FoxPro, and OpenLink Virtuoso. Databases are used across a wide range of industries, including e-commerce, banking, education, government, healthcare, industry, and private services. They are categorized based on their data storage and access methods.

While all relational database management systems (RDBMS) are optimized for storing structured and organized data, they are neither the best nor the most appropriate solution for storing graph-structured data. To address this challenge, NoSQL databases such as key-value, column-family, or document databases can be used. However, the most suitable solution for modeling data in a graph structure is graph databases, which are

specifically designed for this purpose. Among graph databases, Neo4j stands out for its strength and high performance. Neo4j, which is a graph database based on the Java programming language, can be used either as an embedded or server database. The following table illustrates Neo4j's place among NoSQL databases.

3.3.1 NoSQL databases

It is a system composed of data that does not use the SQL language and does not contain relationships with each other. For this reason, it is called NoSQL. The impact on the emergence of the NoSQL model is the rapid development of technology and the enormous amounts of unstructured data generated on the internet. Managing this large amount of data with relational database management systems causes performance issues. At this point, very high performances are being achieved with non-relational database models [16]. NoSQL systems guarantee the BASE (basically available, soft state, eventually consistent) rules as an alternative to ACID. Fundamentally, availability is the constant availability of the database according to the CAP (consistency, availability, partition tolerance) theorem. The safe state means that the system's state can change even without any input. Final consistency is the ability of the system to become more consistent over time. Thanks to these features, even if there is a problem with one of the components of the distributed system, the system continues to operate [17]. NoSQL, which stands for Non-Relational SQL, refers to a technique for data storage and management. Unlike relational databases that organize data into tables, NoSQL databases utilize various data modeling approaches [18]. A key benefit of NoSQL databases is their exceptional scalability and availability, making them particularly suitable for real-time web applications and the realm of big data.

Among the other advantages of NoSQL databases are flexibility, high performance, and functionality.

• **Flexibility:** Flexible schemas offered by NoSQL typically facilitate quicker and more iterative program development. This adaptable data type makes NoSQL databases suitable for both unstructured and semistructured data..

• **Scalability:** NoSQL databases are made to be scalable through the use of distributed hardware clusters rather than the addition of costly, permanent servers.

• **High performance:** In comparison to relational databases, NoSQL databases are specialized for particular data models and access patterns, which allow for improved performance [19].

• **Availability:** Data from several servers, data storage facilities, or cloud sources is automatically replicated using NoSQL databases. Consequently, users' latency time is reduced, irrespective of their location [20].

• **Functionality:** NoSQL databases are intended for distributed data stores with very high storage requirements. This characteristic indicates that NoSQL databases are well-suited for applications that handle large volumes of data, such as real-time web applications, e-commerce platforms, online gaming, the Internet of Things (IoT), social media networks, and online advertising [20].

NoSQL databases can be categorized into four distinct types: key-value stores, document stores, wide-column stores, and graph databases.

3.3.1.1 NoSQL database types

Key-value databases are a high-performance, scalable, and distributed data storage system where data is stored as key-value pairs. Values can include numerical and complex structures [JavaScript Object Notation (JSON), list, BLOB (Binary Large Object), etc.]. The values in this database cannot be retrieved through a query. Only the keys are in a queryable state.

3.3.1.2 Document databases

Document databases, also referred to as document or document databases, are used to store, manage, and retrieve semi-structured data. Documents are stored in XML, JSON, and BSON formats. Among the most commonly used document databases are RethinkDB, Amazon DocumentDB, MongoDB, Couchbase, and Firebase.

3.3.1.3 Wide column databases

Wide column databases stores and manages data in a columnar format. Wide-column databases are commonly used in applications that require a column format to store schemaless data, and the main purpose of these databases is to keep query times short. Wide-column database systems significantly enhance disk input/output performance.

3.3.1.4 Graph databases

In graph databases, the graph structure known from graph theory is used to store data. In this structure, there are peaks, edges, and features. Data is stored at the vertices, and data relationships are represented by the edges between the vertices. The graph database is optimized to overcome the cost of merging multiple tables in SQL systems. It can operate on many servers in a distributed manner. Graph databases are applied to social networks, reservation systems, and fraud detection solutions.

The advantages of graph databases are listed below:

• In information retrieval, the graph database utilizes the adjacency data of one or more vertices. This situation means it is much more optimized compared to relational databases.

• It is quite intuitive because it uses graph theory.

• They are very agile in development because they easily adapt to the addition or deletion of information over time.

• They allow the addition of new data types.

• It generally provides suitability for interconnected data present in the real world.

A graph database implements CRUD (Create, Read, Update, and Delete) operations and the concept of index-free adjacency. Non-contiguous adjacency is critical for high-performance transitions. In this feature, each peak contains a direct reference to its neighbors. It is less costly than universal indexes and is known as a "micro index" for other peaks. This means that the query time is unaffected by the graph's overall size and is only related to the time required to search the graph. Additionally, it demonstrates how the database's connected peaks consistently point to one another. It has a very good capacity to handle big volumes of data and responds to queries quite quickly. Tables are not used to store data in graph databases. Each vertex or edge is therefore immediately connected to every other vertex; there is no merging step. The data on a graph is arranged in peaks that are related by several relationships. In modern applications, graph-like structures have become important again due to their processing capabilities and are referred to as "the future of database management systems." [21].

3.3.1.4.1 Neo4j graph database

It is a NoSQL graph database that was released as open source by a company named Neo4J Technology in 2007 and is still being developed. This database, which uses a graph structure to store data and the relationships between data, is frequently used in areas where interconnected data such as networks, maps, and social networks are stored and processed. In Neo4j, each piece of data is stored in the form of a node, edge, or property/attribute. Each peak and edge can have as many features as desired. Tops and edges can be labeled, and this labeling makes them more useful in search constraints.

The Neo4j database adopts a graph approach, which is beneficial for both traversal performance and transaction runtime. Under the database, the Community edition and the Enterprise edition are available. However, it has tools named Neo4j Desktop, Browser, Neo4j Operation Manager (NOM), Data Importer, and Neo4j Bloom. Neo4j has a cloud-ready architecture that scales based on data needs, minimizes infrastructure costs, and maximizes performance in connected data sets. Thanks to self-managing clustering, it benefits from infrastructure flexibility with less manual intervention and scales data horizontally. Additionally, it keeps the query at a simple level, maintaining performance. This way, it can scale very large lines across multiple databases. Each indexing action had to be carried out directly and by hand when the idea of indexing for Neo4j was initially presented. To limit user growth and enhance the indexing module, Neo4j has created an automatic indexing function that uses a global feature key to automatically integrate changes made to the data in the index. The accuracy of query answers is decreased by the general feature key, which reaches the same feature from multiple nodes and has no semantic link [17].

As a graph database, Neo4j fundamentally consists of a triple structure comprising nodes, relationships that connect these nodes, and properties belonging to both nodes and relationships.

Nodes: The basic components that make up a graph are nodes and edges. In Neo4j, both nodes and relationships can have properties. In Neo4j, a basic graph structure can be represented that includes a node represented by just one property. When an entire network is considered as a large and interconnected graph, a server node can be expressed as a point or as corners between server relationships or edges. In Neo4j, similarly, nodes are the vertices between edges that can hold data expressible as key-value pairs [22].

Relationships: The relationships between nodes are key components of a graph database and allow for finding the relevant data. Relationships, like knots, can also have characteristics. A relationship connects two nodes defined as the start node and the end node. Since relationships are always directional, they can be defined as incoming or outgoing relationships for a node. This definition facilitates navigation on the graph. In other words, there is no need to add the same relationship in the opposite direction to the graph for a relationship. To model the degree of the relationship (for example, the more messages exchanged between two people, the higher the degree of the relationship), a weight is assigned to each node [23].

Features: Relationships and nodes can both have attributes.The properties are key-value pairs when the key is a string. A basic data type or an array of basic data types can be used as property values.

Paths: Relationships between interconnected nodes form paths. Each path has a beginning (starting node), an end (ending node), and a length (the number of relationships between nodes) [23]. In other words, a database path generally refers to one or more nodes connected by relationships, obtained as a result of a query or traversal.

Traversal: Specific to the graph model, traversal is a fundamental activity for retrieving data within the graph. The main feature of the circulations within the graph is that these circulations can be restricted. In this context, querying data using circulation, unlike join operations on relational data, only considers the necessary data without the need for unnecessary grouping operations on the entire data set [24].

Each node and edge in the graph maintains a "mini-index" for the items it is related to, which is one of the key distinctions between traversals and SQL searches. Additionally, traversals are bounded operations rather than global adjacency indexes.

This means that the size of the graph will not create a performance constraint on traversal and that costly grouping operations occurring in SQL JOIN statements will not be needed. However, although global indexes exist in Neo4j, they are only used for determining the starting point of traversal [25].

Neo4j can use a query language called Cypher to perform operations in the database.

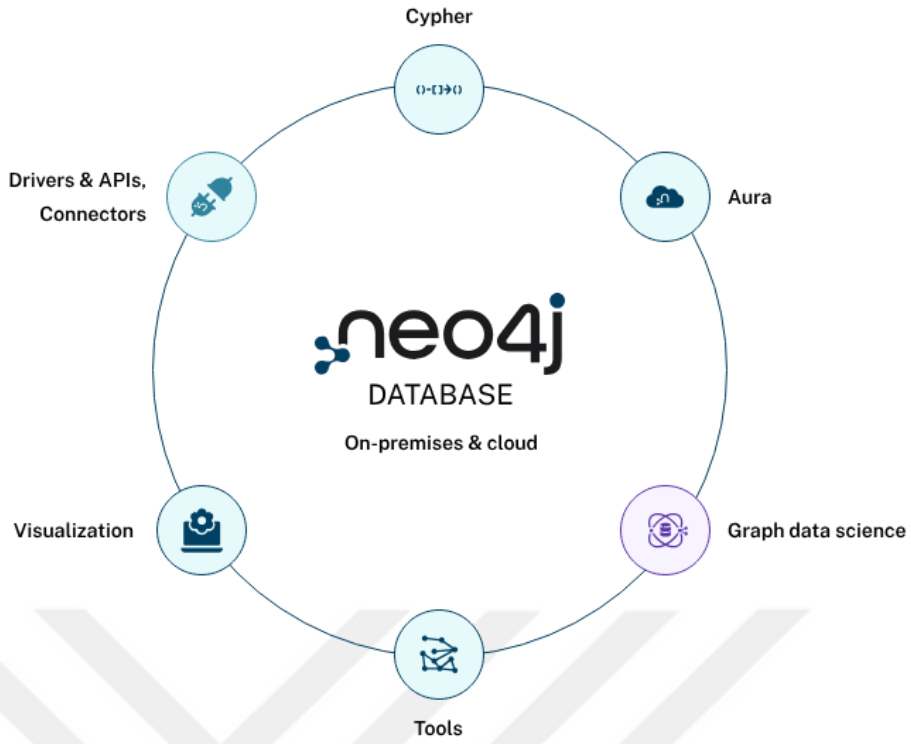

Figure 3.1 The Neo4j graph database ecosystem [26]

3.3.1.4.2 Cypher query language

Cypher is a graph query language that enables fast graph updating and querying without requiring the creation of graph traversal code by hand. It enables meaningful and intuitive querying for both developers and professional users who want to perform real-time queries on a graph database. Designed to be human-centric and natural, Cypher's philosophy is to simplify simple tasks and make complex ones possible. It is structured based on English grammar and basic iconography, making it easy to learn and understand.

Cypher is a declarative language that emphasizes what to query rather than how to execute the query within the graph, in contrast to scripting languages like Gremlin or command-centric languages like Java. This makes Cypher more intuitive, as users can describe the data, they want to retrieve without worrying about the underlying traversal mechanisms. Cypher, developed by Neo Technology for Neo4j, has gained popularity owing to its resemblance to other query languages and its intuitive syntax.

Cypher uses round brackets to denote nodes, arrows to signify relationships, and square brackets to contain relationship details. This ASCII-art-inspired grammar renders

graph structures visually distinct and comprehensible. Moreover, as to other programming languages, comments in Cypher are denoted by prefixing the line with "//". Cypher offers an accessible and fast method for interacting with graph databases, enabling users to query intricate graph topologies effortlessly.

Example Cypher language queries are as follows:

- MATCH (subject:Node {id:'C5433293'})-[r:RELATION {type: 'AFFECTS'}]->(object:Node {id: 'C0027361'})

  RETURN subject, r, object;

This query identifies the node with the identifier 'C5433293' and the node with the identifier 'C0027361', verifies their connection via an AFFECTS relationship, and returns both nodes along with the corresponding relationship.

- MATCH (subject:Node {id: 'C0011860'})-[r:RELATION {type: 'PREDISPOSES'}]->(object:Node)

  RETURN subject, r, object;

This query discovers a node in the database with the id 'C0011860', determines the PREDISPOSES relationship with another node, and retrieves both nodes along with their relationship. This inquiry examines the connection between the node 'C0011860' and another node using the PREDISPOSES relationship, returning both nodes and their interrelation.

- MATCH (subject:Node {id: 'C0311277'})-[r:RELATION {type: 'COEXISTS_WITH '}]->(object:Node)
- RETURN subject, r, object;

This query searches for a node with the ID value of 'C0311277'. If there exists a relationship of type 'COEXISTS_WITH' (RELATION) between this node and another node, the query returns both the nodes (subject and object) as well as the relationship between them. In essence, it investigates how the node with the ID 'C0311277' is connected to another node via the 'COEXISTS_WITH' relationship, and retrieves both nodes along with the relational link between them.

3.4 Used Tools

The technologies applied in our thesis, the tools and development processes of our application, and the methods we employ in detail are all detailed in this chapter under the title "used tools".

3.4.1 PubMed Biomedical Database

PubMed is an online resource that provides a vast array of literature in the fields of biomedical sciences, life sciences, and related areas. It provides access to research articles. Founded in 1996 as a part of the United States National Library of Medicine (NLM), it has grown into a vital resource for biomedical researchers, clinicians, and academics. PubMed serves as a vital tool for academic research, particularly in the fields of biomedical and life sciences, and is widely used by researchers in areas such as biology, chemistry, genetics, and pharmacology.

PubMed provides access to millions of abstracts and details from a diverse spectrum of scholarly journals focused on scientific study. A substantial proportion of the publications housed in the database originate from MEDLINE, a meticulously curated resource cataloging medical and biological literature. However, PubMed's scope extends beyond what is indexed in MEDLINE to incorporate additional biomedical literature. The database offers researchers a bountiful source for exploring interconnected relationships between publications, authors, and keywords.

PubMed represents an openly accessible database for any individual, regardless of association with scientific research or casual interest. It facilitates broad dissemination of biomedical study across global boundaries and allows for the expedited transmission of scientific understanding. The search functionality delivers potent filtering and query options, permitting searches to be limited to precise dates, publications, authors, or terms. PubMed plays a vital role in rigorous scholarly investigation.

This academic work presents an aggregation of publication data for specified disorders of interest compiled utilizing the Pubmed Python API.

3.4.2 Python Programming Language

Python is a high-level programming language that can also support multiple different programming paradigms, designed mainly around object-oriented design and introduced to the world by Guido van Rossum in the 1990s. Python designed by a van Rossum has very simple and readable syntax which allows you to use it correctly. It was designed prioritizing human readability and human usability, that is the reason it has grown to be a most commonly abused item for website developing & scientific study consisting of properly organized communication tends to make existence easy for programmers who can easily learn how to deal with elements as well as how manages extravagant projects.

Python is very powerful within itself, but the standard library and third-party packages you have for Python extend this even further. Python's popularity is evidenced by packages for web development (Django, Flask), machine learning platforms (TensorFlow and PyTorch), and scientific computing tools such as NumPy, Pandas and SciPy. With an engaged user base and extensive documentation been written, it became a favorite for researchers and data-analysts trying to find solutions to their complicated problems.

The real versatility of Python comes to when you can run it on any operating system, meaning that developers only need flexibility in terms of the environment used. In Python code, object-oriented and functional approaches can merge easily. Its flexibility helps in rapid experimentation and testing for the data science and AI work in particular, providing a quick way to move things along.

Beyond the analysis methods and natural language processing techniques described here, there were many other data preprocessing steps utilized in this thesis including extensive use of the Pymed API, text mining for citation behavior, stopword filtering using NLTK, access with UMLS or MeSH terms as keywords for automated deduction that were filtered to single occurrences or aggregations on findings.

3.4.3 Natural Language Toolkit

NLTK is indeed a robust Python library utilized for examining natural language. It affords specialists and builders an assortment of linguistic strategies, like written substance investigation, phrase handling, grouping, morphological investigation, and syntactic examination. NLTK skillfully deals with both elementary and state-of-the-art dialect preparing assignments, consequently upgrading knowing and utilizing in the NLP field through furnishing important instructional assets. NLTK is very regarded for its wide range of utilizations in tutoring and examine, particularly in the orders of lingual examination and common language preparing.

The doctoral prospectus actualized separating calculations in Python, drawing upon the lexicon acquired from NLTK's rundown of stops.

3.4.4 Unified Medical Language System

The Unified Medical Language System (UMLS) is an immense framework developed by the National Library of Medicine to standardize disparate biomedical terminology from various sources. By connecting different health data through a shared vocabulary and organization, it facilitates the amalgamation of clinical information in a complex manner. Composed of three core components that explain biomedical semantic associations, the enormous Metathesaurus accumulates millions of short and long terms alongside their interconnections drawn from domains, while the extensive Semantic Network categorizes and precisely defines conceptual linkages. As a powerful instrument, the UMLS benefits both researchers and practitioners with sophisticated information mining, retrieval and analysis of electronic health records, enabling the derivation of accurate conclusions. A Concept Unique Identifier (CUI) singularly identifies each medical notion or phrase throughout the system in an intricate way. This figure denotes interconnected themes between ideas originating in diverse places, thereby permitting intricate linking of related notions.

The doctoral dissertation leveraged the intricate UMLS application programming interface to precisely map biomedical entities to their respective Concept Unique Identifiers.

3.4.5 Semantic MEDLINE Database

SemMedDB, a vast biomedical database created by the National Library of Medicine, systematically collects relationships and semantic details extracted from peer-reviewed articles cataloged in MEDLINE. By utilizing natural language processing techniques, SemRep is able to mine connections between diverse concepts mentioned across abundant summaries - linking, for instance, pharmaceuticals to afflictions or biological mechanisms to conditions.

This comprehensive resource finds extensive usage in numerous fields of biomedicine, assisting with hypothesis formulation, simplifying information access, and supporting natural language research by imposing organization on unstructured life science texts. The database lays bare the intricately interwoven network of associations interlaced throughout the literature, better prioritizing applicable findings and easing navigation. Complex connections are uncovered between varied topics, presenting fresh perspectives for inquisitive researchers to explore while maintaining the same word count as the original text.

The dissertation presented results firmly established upon the relationship types cataloged within SemMedDB, drawing connections between concepts to ground hypotheses.

3.4.6 MySQL Database Management System

Relational databases SQL queries seamlessly offer structured ways to organize second and retrieve entities. MySQL uses SQL to handle related tabular data of a wide variety of types. MySQL has logically linked tables where data exists (and its relations maintain consistency and accuracy). Some tables contain high-level information and some contain details with complex relationships. Queries against this mesh deepen insight with either filtered complex relationships or aggregated simple facts. In all cases, the ability to represent a wide variety of real-life entities and forces through minimum-brick normalization lends itself to flexible but disciplined management of transforming throughputs. MySQL is a highly utilized, powerful, and cross-platform database, but to take full advantage of its power, it is necessary to assemble complex SQL queries to find the connections. MySQL is the ubiquitous engine powering portals, marketplaces and any

app that makes serious data demands because of its flaming speed and granite-hard reliability in the face of massive workloads. MySQL supports everything from smaller hobby projects to larger cross-continent enterprises and works on all major environments (Linux, Windows, macOS). Yet in order to make sense of it, we must carefully harness its query language to impose order on the chaos of relations and form links between what seem to be non-entities..

In the thesis, SQL queries have been created in MySQL to elucidate significant linkages.

3.4.7 Neo4j Graph Database Management System

Neo4j is an open-source graph database management system designed for the storage and management of graph data. In contrast to conventional relational databases, Neo4j organizes data as nodes and the interconnections among these nodes. This framework provides efficient functionality, particularly when intricate data interactions require modeling and querying.

Neo4j has a specific query language, known as Cypher, built for graph databases. Cypher enables users to efficiently compose data queries and do significant analytics. Neo4j is widely utilized in applications such as social network analysis, recommendation systems, fraud detection, and bioinformatics because of its inherent ability to represent complicated data connections while delivering exceptional performance. Its adaptability and proficient analytical tools render it appropriate for both minor and extensive undertakings.

In the thesis, significant and meaningful relationships were identified by queries formulated in the Cypher language on Neo4j.

3.5 Proposed Approach

3.5.1 Article Data Collection

In this study, publications in the medical field (articles, conference proceedings, etc.) available in the PubMed repository were used as the primary data elements. To derive the dataset, publicly accessible medical studies in the PubMed digital repository were retrieved by searching for target disease names as keywords. The diseases of interest in this study include “Type 2 Diabetes,” “COVID-19,” “Obesity,” “Inflammatory Bowel Disease,” and “Colorectal Cancer.” For each of these target diseases, studies with non-empty abstract and title information were extracted using Python and the Pymed API. This process was constrained to publications from the last two years (730 days) and three years (1,095 days), prioritizing recent information in the literature.

Due to the Pymed API’s limitation of processing up to 9,999 entries per query, the data collection process was carried out incrementally in monthly batches. Consequently, the number of articles retrieved for each disease varies according to publication frequency. The final counts of publications collected for each disease are as follows:

- Obesity: 90,485 publications from the last three years
- Type 2 Diabetes: 59,676 publications from the last three years
- COVID-19: 113,431 publications from the last two years
- Colorectal Cancer: 60,176 publications from the last three years
- Inflammatory Bowel Disease: 27,303 publications from the last three years

To create a homogeneous structure and facilitate further processing, the extracted abstracts were consolidated into a single `.json` file. This structured dataset is expected to support subsequent steps in analyzing and interpreting current trends in medical research related to the selected diseases.

3.5.2 Merging of Abstracts and Creation of Meaningful Word Lists

The abstracts retrieved through Python were combined into a single ".json" file. From these consolidated abstracts, meaningless words (e.g., "the," "and," "a," "an," "what," "which," etc.) were filtered out using the Natural Language Toolkit (NLTK) text mining library. To further refine the dataset, the 500 most frequently occurring words with at least three characters were identified using a custom Python script. These meaningful, high-frequency terms were saved in a new ".json" file, creating a list of 500 context-relevant words or terms that reflect the thematic focus of the dataset.

3.5.3 Obtaining CUIs Using the UMLS API and Removal of Duplicate Records

Using the UMLS (Unified Medical Language System) Application Programming Interface (API), we successfully obtained the Concept Unique Identifier (CUI) for each of the 500 relevant words associated with each disease. The Python programming language was employed to identify and eliminate duplicate records within the dataset. Additionally, the frequency of each term was preserved as a separate value to track its occurrence.

The complex network of interconnecting health concepts were analyzed through leveraging the expansive catalog of the Unified Medical Language System. The UMLS maintains a vast collection of over a million biomedical terms spanning numerous coding systems, ontologies, and classifications from diverse sources such as published literature, nosologies of diseases, pharmaceutical nomenclature, diagnostic laboratory tests, and additional health domains [27]. This aggregation amalgamates heterogenous lexicons into a cohesive structure where each entry possesses a distinctive Concept Unique Identifier facilitating mapping of equivalent notions across disparate terminologies.

To uncover relationships between entities within the dataset, the most pertinent keywords associated with each condition were matched to their corresponding CUIs using the UMLS API. This approach proved integral for finding conceptual connections, generating singular CUIs for every keyword to facilitate discovery and investigation of the conceptual interdependencies inherent within the data.

3.5.4 Running Queries on MySQL and Filtering of Relationship Types from SemMedDB

Out of the 60 relationship types available in SemMedDB, 30 were selected based on relevance. The data was filtered to include only these relationships, resulting in a new filtered_predication_table.csv file. A query was written to run on the .csv file filtered in the previous step, using MySQL. This query generated results containing the SUBJECT_CUI, PREDICATE, and OBJECT_CUI columns, with rows representing relationships that occurred 3, 5, 10, and 20 times. These results were saved as separate .csv files with frequency values included.

SemMedDB is a comprehensive repository that catalogs subject-predicate/object connections derived from biomedical research, specifically through the analysis of titles and abstracts, formatted as semantic triplets [28]. This resource is accessible to the public and is maintained by the National Library of Medicine (NLM) in the United States. It employs SemRep, a rule-based natural language processing (NLP) tool, to extract "semantic predictions" from biomedical literature [29]. The database is created by applying the SemRep tool to all biomedical literature indexed in the PubMed search engine, with the relationships identified during this process being classified as semantic triplets. The extraction process involves standardizing the textual references of entities (subjects and objects) into distinct UMLS Metathesaurus concepts and aligning predicates/relations with those defined in the UMLS semantic network..

Since the records in SemMedDB may be extracted from multiple academic studies, there may be duplicate entries. To address this, a preprocessing and cleaning step was performed to remove duplicate records (due to their presence in multiple studies) and organize the dataset. As a result, each record was preserved as a distinct unit, and duplicates were filtered out while their frequency of occurrence was retained as a separate value. In the next filtering step, 30 relationship types, as listed in Table 3.1, were selected from the over 60 unique relationship types found in SemMedDB, and a .csv file was created containing only these relationships. This filtering step helped eliminate potentially misleading records that could introduce noise, ensuring that only the most relevant relationship types were included in the dataset.

Table 3.1 Relationships defined for graph structure

PROCESS_OF	COEXISTS_WITH	PREVENTS
ISA	DIAGNOSES	INHIBITS
CAUSES	PREDISPOSES	higher_than
LOCATION_OF	STIMULATES	PART_OF
METHOD_OF	CONVERTS_TO	lower_than
PRODUCES	ASSOCIATED_WITH	DISRUPTS
COMPLICATES	OCCURS_IN	MEASURES
AFFECTS	ADMINISTERED_TO	AUGMENTS
TREATS	MEASUREMENT_OF	PRECEDES
USES	INTERACTS_WITH	same_as

3.5.5 Data Preprocessing and Identifying Relationship Types

Preprocessing of the resulting .csv files was performed using Python. During this step, semicolons (;) were replaced with commas (,), and the SUBJECT_CUI and OBJECT_CUI columns were filtered to retain only rows where terms followed the CUI format, starting with the letter "C." Rows containing multiple CUI codes were removed, and frequency values were listed in a separate column. New .csv files with the updated format were created.

The frequency_3 .csv file was imported into Neo4j. Using Neo4j queries, meaningful relationship types for the five selected diseases were identified by modifying the search direction to ensure proper relationships between SUBJECT and OBJECT. After querying, separate .csv files were generated for each relationship, and these files were integrated into a combined Excel sheet for each disease.

3.5.6 Adding Semantic Information and Filtering for Meaningful Relationships

To improve the readability of the SUBJECT (CUI), RELATION (Selected Relationships), and OBJECT (CUI) columns in the combined Excel sheets, a Python script was written. This script added new columns for Subject_Name,

Subject_SemanticTypes, Object_Name, and Object_SemanticTypes. Separate sem_type_info Excel tables were created for each disease based on these enhancements.

Upon examining the created Excel tables, meaningful and concrete relationships were identified. Subsequently, filtering was applied for the relationship types TREATS, PREVENTS, CAUSES, PRECEDES, COEXISTS_WITH, COMPLICATES, DIAGNOSES, DISRUPTS, ISA, MEASURES, OCCURS_IN, AUGMENTS, ASSOCIATED_WITH, AFFECTS, INTERACTS_WITH, and STIMULATES in each Excel table, revealing significant associations.

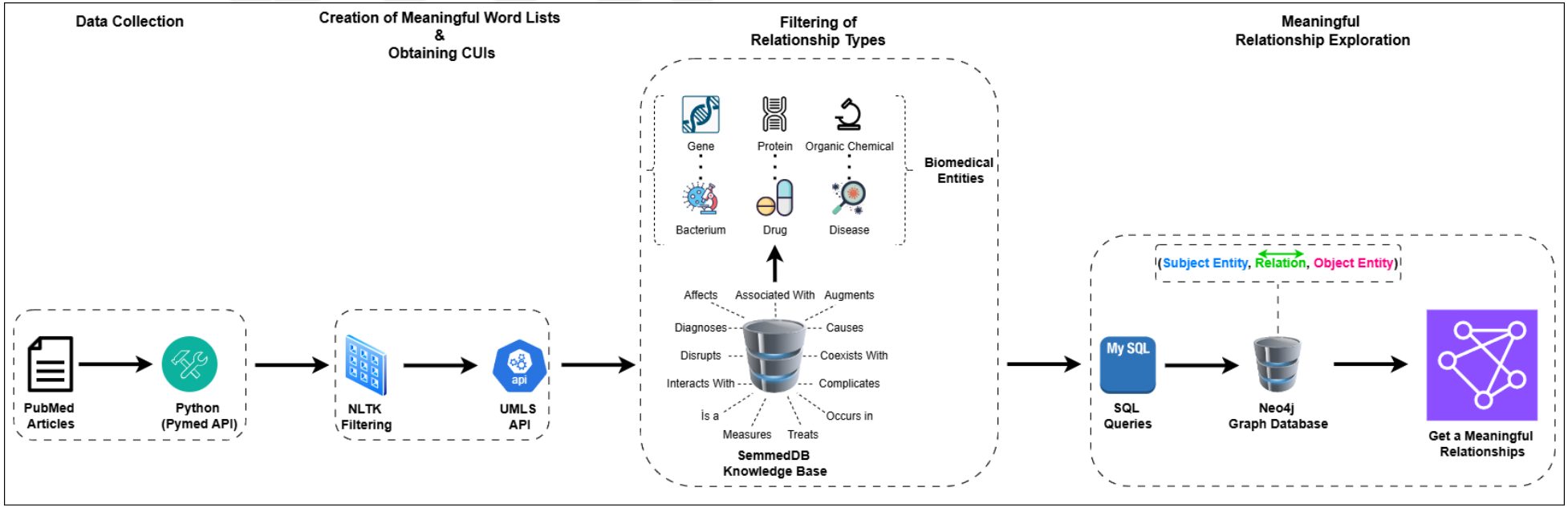

Figure 3.2 Overall execution flow

Chapter 4

Results

Following the completion of all these studies, the results obtained were compiled. Separate tables were generated for each disease. Since the results obtained with a frequency value of 3 encompassed all other frequency values (5, 10, and 20), we proceeded with the tables generated using a frequency value of 3. Meaningful outputs could not be obtained only for Colorectal Cancer (CUI value C0346629). Out of the five diseases targeted at the beginning of the study, results were achieved for four: Obesity, COVID-19, Inflammatory Bowel Disease, and Type 2 Diabetes. In this chapter, the significant relationships identified for these diseases were manually filtered by us. After this filtering, the results for each disease were ranked from highest to lowest frequency, with the top 25 results shared.

To facilitate understanding of the table, each column header can be explained as follows: The **SUBJECT_CUI** column provides the CUI value of the entity used as the subject; the **Subject_Name** column lists the name corresponding to the subject CUI; the **RELATION** column denotes the type of relationship; **Object_Name** lists the name corresponding to the object CUI; **OBJECT_CUI** gives the CUI value of the entity used as the object; and the **FREQUENCY** column indicates the number of unique PubMed abstracts in which this relationship appeared within the specified periods.

Table 4.2 Obesity results

SUBJECT_CUI	Subject_Name	RELATION	Object_Name	OBJECT_CUI	FREQUENCY
C0020538	Hypertensive disease	COEXISTS_WITH	Obesity, Abdominal	C0311277	44
C0021655	Insulin Resistance	COEXISTS_WITH	Obesity, Abdominal	C0311277	35
C0311277	Obesity, Abdominal	COEXISTS_WITH	Insulin Resistance	C0021655	32
C0311277	Obesity, Abdominal	COEXISTS_WITH	Polycystic Ovary Syndrome	C0032460	18
C0311277	Obesity, Abdominal	COEXISTS_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	17
C0242339	Dyslipidemias	COEXISTS_WITH	Obesity, Abdominal	C0311277	15
C0311277	Obesity, Abdominal	COEXISTS_WITH	Cushing Syndrome	C0010481	14
C0311277	Obesity, Abdominal	COEXISTS_WITH	Hypertensive disease	C0020538	14
C0011847	Diabetes	COEXISTS_WITH	Obesity, Abdominal	C0311277	14
C0011860	Diabetes Mellitus, Non-Insulin-Dependent	COEXISTS_WITH	Obesity, Abdominal	C0311277	14
C0394664	Acupuncture Therapy	TREATS	Obesity, Abdominal	C0311277	9
C0022661	Kidney Failure, Chronic	COEXISTS_WITH	Obesity, Abdominal	C0311277	9
C0311277	Obesity, Abdominal	COEXISTS_WITH	Sleep Apnea, Obstructive	C0520679	8
C0017710	Glucocorticoids	CAUSES	Obesity, Abdominal	C0311277	8
C2744535	CD69 protein, human	AFFECTS	Obesity, Abdominal	C0311277	8
C0311277	Obesity, Abdominal	COEXISTS_WITH	Dyslipidemias	C0242339	7
C0006560	C-reactive protein	AFFECTS	Obesity, Abdominal	C0311277	7
C0037511	sodium glutamate	CAUSES	Obesity, Abdominal	C0311277	5
C0963992	resistin	AFFECTS	Obesity, Abdominal	C0311277	5
C0311277	Obesity, Abdominal	COEXISTS_WITH	Hyperglycemia	C0020456	4
C0311277	Obesity, Abdominal	COEXISTS_WITH	Asthma	C0004096	4
C0011849	Diabetes Mellitus	COEXISTS_WITH	Obesity, Abdominal	C0311277	4
C0020459	Hyperinsulinism	COEXISTS_WITH	Obesity, Abdominal	C0311277	4
C0020456	Hyperglycemia	COEXISTS_WITH	Obesity, Abdominal	C0311277	4
C0311277	Obesity, Abdominal	COEXISTS_WITH	Impaired glucose tolerance (disorder)	C0271650	3

A literature review was conducted to verify the **“hypertensive disease” COEXISTS_WITH “Obesity, Abdominal”** relationship, with results obtained for obesity appearing in 44 unique publications. Following the studies, numerous works were found that clearly confirm this relationship. For instance, in the study by Jiang et al. (2016), it was mentioned that cardiovascular diseases, which are the leading cause of death globally and include hypertension and diabetes, are the primary diseases associated with obesity [30]. This study also highlighted the relationship between diabetes arising from insulin resistance and obesity, as presented in the results table. Additionally, in the work by Seravelle et al. (2024), it was demonstrated that obesity is a worldwide pandemic that contributes to cardiovascular morbidity, leading to increased hospital admissions, events, and costs within healthcare systems. Several pathophysiological mechanisms are present in both obesity and obesity-induced hypertension [31].

A literature review was performed to confirm the relationship between **"polycystic ovary syndrome" COEXISTS_WITH "Obesity, Abdominal"** based on findings from 18 different articles. Additional research uncovered a significant amount of evidence that firmly backs this connection. According to Gambineri et al. (2002), almost 50% of women with PCOS are classified as overweight or obese, with a predominant abdominal phenotype observed. It was also mentioned that obesity could contribute to the development of the condition in individuals who are at risk, and that losing weight may lead to improvements in hormones, metabolism, and clinical symptoms. The research indicated that adding insulin-sensitizing medications and/or anti-androgens to weight-loss programs might improve both clinical and hormonal results, emphasizing the importance of obesity in the development of polycystic ovary syndrome [32]. Alves et al. (2009) demonstrated that abdominal obesity and hyperandrogenism both play a role in dyslipidemia and other metabolic traits that could negatively impact the long-term health of women with polycystic ovary syndrome [33].

A literature review has been conducted to validate the association of **"Acupuncture Therapy" TREATS "Obesity, Abdominal"** revealing obesity outcomes in nine separate articles. The research indicated that other studies unequivocally corroborate this association. A study by Zhang et al. (2018) indicated that the integration of acupuncture and associated therapies is a particularly effective method for decreasing weight and body mass index [34].

Table 4.3 COVID-19 results

SUBJECT_CUI	Subject_Name	RELATION	Object_Name	OBJECT_CUI	FREQUENCY
C0003467	Anxiety	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	1611
C0011581	Depressive disorder	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	1284
C0042210	Vaccines	TREATS	COVID-19 Virus Disease	C5203670	834
C4726677	remdesivir	TREATS	COVID-19 Virus Disease	C5203670	823
C0020336	hydroxychloroquine	TREATS	COVID-19 Virus Disease	C5203670	763
C1609165	tocilizumab	TREATS	COVID-19 Virus Disease	C5203670	535
C2744535	CD69 protein, human	AFFECTS	COVID-19 Virus Disease	C5203670	496
C0001617	Adrenal Cortex Hormones	TREATS	COVID-19 Virus Disease	C5203670	390
C2744535	CD69 protein, human	ASSOCIATED_WITH	COVID-19 Virus Disease	C5203670	315
C0008269	chloroquine	TREATS	COVID-19 Virus Disease	C5203670	287
C0011777	dexamethasone	TREATS	COVID-19 Virus Disease	C5203670	262
C0052796	azithromycin	TREATS	COVID-19 Virus Disease	C5203670	248
C0022322	ivermectin	TREATS	COVID-19 Virus Disease	C5203670	196
C1138226	favipiravir	TREATS	COVID-19 Virus Disease	C5203670	190
C0939237	lopinavir / ritonavir	TREATS	COVID-19 Virus Disease	C5203670	181
C0022660	Kidney Failure, Acute	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	163
C0014695	ergocalciferol	ASSOCIATED_WITH	COVID-19 Virus Disease	C5203670	163
C0497327	Dementia	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	160
C0040053	Thrombosis	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	160
C5435024	molnupiravir	TREATS	COVID-19 Virus Disease	C5203670	156
C0292818	ritonavir	TREATS	COVID-19 Virus Disease	C5203670	143
C0021760	interleukin-6	ASSOCIATED_WITH	COVID-19 Virus Disease	C5203670	139
C0960880	angiotensin converting enzyme 2	ASSOCIATED_WITH	COVID-19 Virus Disease	C5203670	134
C0038454	Cerebrovascular accident	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	133
C0011854	Diabetes Mellitus, Insulin-Dependent	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	117

A literature review was conducted to confirm the relationship between **"Anxiety" COEXISTS_WITH "COVID-19 Virus Disease"** uncovering relevant findings from 1611 unique articles related to COVID-19. A lot of papers were found that clearly support this connection. A study by Kan et al. (2021) suggested that identifying population segments more vulnerable to mental disorders during the COVID-19 crisis could lead to the implementation of targeted mental health services and supportive interventions, offering appropriate assistance during the pandemic [35]. Sher et al. (2020) highlighted the concern of rising suicide rates amid the outbreak of the highly contagious coronavirus disease that emerged in China at the end of 2019 (COVID-19). The COVID-19 pandemic has been linked to increased levels of anxiety, depression, distress, sleep issues, and suicidal thoughts. Additionally, the study highlighted that identifying and addressing insomnia is especially crucial during stressful times such as the COVID-19 pandemic, as it may greatly lower suicide rates [36].

For the purpose of confirming the relationship between **"remdesivir" TREATS "COVID-19 Virus Disease"** a comprehensive examination of the relevant literature was carried out. The findings concerning COVID-19 were discovered in 823 distinct articles. There are a great number of research that have been found that unequivocally demonstrate this relationship. In a research investigation conducted by Beigel et al. (2020), findings demonstrated that remdesivir significantly outperformed a placebo in reducing recovery times for hospitalized adults suffering from lower respiratory tract infections related to COVID-19 [37]. Similarly, a study by Gottlieb et al. (2021) indicated that for non-hospitalized patients at an elevated risk of COVID-19 complications, a three-day regimen of remdesivir was associated with a favorable safety profile and led to an 87% reduction in the likelihood of hospitalization or death when compared to a placebo group [38].

To verify the relationship **"interleukin-6" ASSOCIATED_WITH "COVID-19 Virus Disease"**, a review of the literature was conducted, with findings related to COVID-19 identified in 139 unique publications. Numerous studies were found that clearly confirm this relationship. For example, in the study by Coomes et al. (2020), it was reported that serum interleukin-6 levels are significantly elevated in cases of complicated COVID-19, with higher interleukin-6 levels strongly associated with adverse clinical outcomes. The findings underscore the need for ongoing controlled clinical trials

to explore the role of immunomodulation, particularly through interleukin-6 inhibition, in the treatment of severe COVID-19 [39]. In a related investigation conducted by Grifoni et al. (2020), it was observed that addressing the cytokine storm induced by SARS-CoV-2 with anti-interleukin-6 therapies reinforces the theory that such a strategy could serve as a viable treatment alternative. This approach, when combined with supportive care measures, has the potential to enhance patient outcomes in cases of COVID-19 [40].

Table 4.4 Inflammatory Bowel Disease (IBD) results

SUBJECT_CUI	Subject_Name	RELATION	Object_Name	OBJECT_CUI	FREQUENCY
C0079189	cytokine	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	212
C0666743	infliximab	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	187
C0079225	Dextran Sulfate Sodium (DSS)	CAUSES	Inflammatory Bowel Diseases	C0021390	153
C0002871	Anemia	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	153
C0950624	Leukocyte L1 Antigen Complex	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	139
C0006826	Malignant Neoplasms	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	135
C0334044	Dysplasia	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	130
C2985398	Intestinal Microbiome	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	129
C0021390	Inflammatory Bowel Diseases	COEXISTS_WITH	Primary sclerosing cholangitis	C0566602	129
C0021390	Inflammatory Bowel Diseases	COEXISTS_WITH	Crohn Disease	C0010346	126
C0021390	Inflammatory Bowel Diseases	COEXISTS_WITH	Ulcerative Colitis	C0009324	101
C0009402	Colorectal Carcinoma	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	84
C0042866	vitamin D	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	76
C2744535	CD69 protein, human	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	74
C0343386	Clostridium difficile infection	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	70
C1448177	TNF protein, human	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	69
C0021390	Inflammatory Bowel Diseases	AFFECTS	Pathogenesis	C0699748	65
C0021390	Inflammatory Bowel Diseases	COEXISTS_WITH	Arthritis	C0003864	65
C0162316	Iron deficiency anemia	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	63
C0103647	Antineutrophil Cytoplasmic Antibodies	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	60
C0021390	Inflammatory Bowel Diseases	COEXISTS_WITH	COVID-19 Virus Disease	C5203670	60
C0011991	Diarrhea	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	59
C3658208	Dysbiosis	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	57
C0016059	Fibrosis	COEXISTS_WITH	Inflammatory Bowel Diseases	C0021390	57
C0127615	mesalamine	ASSOCIATED_WITH	Inflammatory Bowel Diseases	C0021390	53

In order to confirm the relationship between **"infliximab" ASSOCIATED_WITH "Inflammatory Bowel Diseases"** a comprehensive literature review was undertaken, revealing findings pertinent to Inflammatory Bowel Diseases across 187 distinct publications. A large body of research has been identified that substantiates this relationship. For instance, in the study by Papamichael et al. (2019), it was explicitly stated that the existing evidence demonstrates that infliximab is highly effective in the treatment of IBD [41]. In a separate investigation carried out by Danese (2008), it was proposed that when evaluating the effects of infliximab across various pathogenic levels, it seems to act as a comprehensive anti-inflammatory agent within the intricate network of mucosal cell interactions in both types of inflammatory bowel disease (IBD) [42].

A literature review was undertaken to assess the causal link between **"anemia" CAUSES "Inflammatory Bowel Diseases"** identifying relevant findings in 153 distinct publications. Numerous studies were found that clearly support this association. For instance, in the study by Gomollón et al. (2009), it was noted that anemia is quite prevalent in IBD. Within the context of the relationship indicated in the table as **"Inflammatory Bowel Diseases" COEXISTS_WITH "Crohn Disease"** which appears in 126 unique publications, it was highlighted that in Crohn's disease, iron deficiency, deficiencies in vitamin B12 and/or folic acid, malabsorption, inadequate nutrition, inflammation, bowel resections, and drug effects can all contribute to a multifactorial and complex problem, making it a challenging clinical issue. The recent findings related to the timing of the study have highlighted a significant concern for healthcare providers: anemia could be a persistent or at least recurring problem in individuals with inflammatory bowel disease (IBD). This suggests the necessity for ongoing monitoring of patients after the conclusion of their treatment, alongside the need for routine assessments of anemia and iron deficiency in standard medical evaluations [43]. In a related investigation, Oustamanolakis et al. (2011) identified a key challenge for IBD specialists: the integration of traditional anemia indicators with advanced diagnostic parameters to optimize the detection of anemia in IBD patients. It was emphasized that anemia serves as a clinical manifestation of IBD [44].

Table 4.5 Type 2 Diabetes (T2D) results

SUBJECT_CUI	Subject_Name	RELATION	Object_Name	OBJECT_CUI	FREQUENCY
C0025598	metformin	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	2906
C0021655	Insulin Resistance	COEXISTS_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	1368
C0243192	agonists	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	1057
C0038766	Sulfonylurea Compounds	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	810
C0020456	Hyperglycemia	COEXISTS_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	706
C0752046	Single Nucleotide Polymorphism	ASSOCIATED_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	689
C1456408	liraglutide	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	649
C5392125	Glycemic Control	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	634
C0011860	Diabetes Mellitus, Non-Insulin-Dependent	COEXISTS_WITH	Obesity	C0028754	574
C0038432	streptozocin	CAUSES	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	560
C0071097	pioglitazone	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	527
C0167117	exenatide	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	420
C2917359	GLP-1 Receptor Agonist	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	369
C0289313	rosiglitazone	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	279
C0061323	glimepiride	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	278
C0389071	Adiponectin	ASSOCIATED_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	276
C0017628	glyburide	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	274
C0022658	Kidney Diseases	COEXISTS_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	256
C1570906	vildagliptin	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	244
C0011884	Diabetic Retinopathy	COEXISTS_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	218
C0050393	acarbose	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	218
C1562292	Incretins	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	212
C3490348	empagliflozin	TREATS	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	197
C1420644	TCF7L2 gene	ASSOCIATED_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	194
C0063684	Amylin	ASSOCIATED_WITH	Diabetes Mellitus, Non-Insulin-Dependent	C0011860	189

A review of the literature was carried out to confirm the relationship **"metformin" TREATS "Diabetes Mellitus, Non-Insulin Dependent"** with relevant findings on Non-Insulin Dependent Diabetes Mellitus appearing in 2,906 unique publications. Many studies were identified that provide strong evidence supporting this relationship. The research conducted by DeFronzo et al. in 1995 clearly showed that both metformin alone and its combination with sulfonylureas were generally well tolerated among patients. Furthermore, these approaches significantly enhanced glycemic control and improved lipid levels in people with non-insulin-dependent diabetes mellitus who did not achieve adequate management through diet or sulfonylurea alone [45]. In addition, an investigation by Saenz et al. in 2005 proposed that metformin may serve as the preferred initial treatment for overweight or obese individuals diagnosed with type 2 diabetes mellitus owing to its potential to mitigate the risk of specific vascular complications and reduce mortality rates. Moreover, their findings highlighted that metformin alone or along with other drugs considerably enhanced blood sugar management and lipid profile in non-insulin-dependent diabetics that failed to gain suitable regulation through diet or sulfonylurea treatment exclusively [46].

A literature review was performed to confirm the association between **"Adiponectin" ASSOCIATED_WITH "Diabetes Mellitus, Non-Insulin Dependent"** identifying relevant findings in 276 unique publications. Numerous studies were found that strongly support this association. For instance, in the study by Xita et al. (2012), it was explicitly stated that adiponectin plays a significant role in insulin resistance and the development of type 2 diabetes. Insights into the signaling mechanisms involved in adiponectin action were suggested to provide benefits for understanding the pathophysiology and treatment of diabetes. Furthermore, it was highlighted that compounds capable of enhancing or imitating the activation of adiponectin receptors and/or the subsequent signaling pathways may offer a viable therapeutic approach for both the prevention and management of diabetes [47]. The research by Howlader et al. (2021) indicated that the consistent association observed in diverse populations, along with a dose-response relationship and corroborative evidence from mechanistic investigations, positions adiponectin as a significant target for mitigating the risk of diabetes [48].

Chapter 5

Conclusions and Future Prospects

5.1 Conclusions

In this thesis, the process of transforming academic studies related to the predefined target diseases Obesity, COVID-19, Inflammatory Bowel Disease, and Type 2 Diabetes into a graph structure and then finding meaningful relationships through queries in the graph structure for these target diseases has been described. As shown in the Results chapter, when conducting a literature review on the meaningful relationships found, multiple academic studies that clearly support these relationships have been identified. In addition to the 30 examples shared for each disease in the Results chapter, other results obtained through similar queries can also be validated through publications found in PubMed and other academic databases. In summary, these results demonstrate the accuracy of the proposed methodology.

5.2 Societal Impact and Contribution to Global Sustainability

This thesis combines text mining methodologies with bioinformatics techniques, aiming to produce a multidisciplinary study that integrates innovative approaches in both fields. The research aligns with the research focus areas of AGU (Abdullah Gül University), specifically in the domains of "Health and Medical Biotechnology" and "Innovation and Entrepreneurship." The study will involve analyzing disease-related content within abstracts of previously published research articles using text mining tools. The goal is to uncover multiple relationships, such as disease-pathway, disease-virus species, disease-bacteria species, and disease-gene associations.

Aligned with the United Nations Sustainable Development Goals, particularly the "Good Health and Well-being" goal, this work will focus on advancing the understanding of disease progression, improving treatment methods, and addressing gaps in current research. Additionally, under the "Partnerships for the Goals" category, this study encourages collaborative efforts across different academic disciplines and research domains, fostering a shared motivation to conduct impactful research. By merging text mining and bioinformatics, this thesis will generate insights that contribute to disease understanding and treatment development, potentially offering guidance for future medical breakthroughs.

In this study, we explore the integration of different literature-based information retrieval methods, specifically text-to-graph transformation processes that identify relevant graphs and association queries to extract novel and statistically significant associations for target diseases. The focus of these results from experimental studies has been into the target disease categories such as Obesity, COVID-19, Inflammatory Bowel Disease and Type 2 Diabetes as we extensively explore here. The accuracy of the derived associations shows the effectiveness of the proposed strategy, thus validating its aptitude for creating valuable associations. In addition, it is more effective and efficient than manual searches through the PubMed web interface as it detects relationships among all conditions.

This study combines contemporary text mining methodologies with bioinformatics, augmenting our understanding of the molecular underpinnings of diseases and simultaneously advancing global sustainability by addressing health-related challenges. The findings and methodologies of this study could greatly benefit society by improving diagnostic and treatment approaches and promoting international cooperation to tackle global health issues.

5.3 Future Prospects

In this study, meaningful relationship inference for target diseases was made through graph querying processes using the text-to-graph transformation operation, one of the literature-based information retrieval methods. Experimental study results have been reported, especially for target disease groups such as Obesity, COVID-19, Inflammatory Bowel Disease, and Type 2 Diabetes. The accuracy of the obtained relationships has proven the success of the implemented method. Additionally, the proposed method finds relationships much more effectively than queries conducted manually through the PubMed web interface for each disease. The following goals are planned as relevant areas of work in the future.

- Queries consisting of singular binary connections can be expanded to discover possible unknown new relationships with more sophisticated queries.
- Every relationship in the UMLS semantic network comes with a specific set of constraints related to domain semantic types. According to the expert opinion of NLM, it has been established which types of entities can take on the roles of subject and object for each relationship. Will identify new candidate drugs prior to clinical trials by using CUIs that align with the results of their relationships.
- The proposed method can be re-tested by retrieving publications containing specific terms or terms related to much more specialized data retrieval methods through the PubMed API, including studies published up to a certain year.
- Similar analyses will also be conducted for different diseases.

BIBLIOGRAPHY

[1] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the Future*, vol. 2007, no. 2012, pp. 1–16, 2012.

[2] V. Vijayan, J. P. Connolly, J. Condell, N. McKelvey, and P. Gardiner, "Review of wearable devices and data collection considerations for connected health," *Sensors*, vol. 21, no. 16, p. 5589, Aug. 2021.

[3] Z. Lv and L. Qiao, "Analysis of healthcare big data," *Future Generation Computer Systems*, vol. 109, pp. 103–110, Sep. 2020.

[4] B. Kolukisa, B. K. Dedeturk, B. A. Dedeturk, A. Gulsen, and G. Bakal, "A comparative analysis on medical article classification using text mining & machine learning algorithms," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 2021, pp. 360–365.

[5] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, p. baq036, Jan. 2011.

[6] J. Guia, V. G. Soares, and J. Bernardino, "Graph Databases: Neo4j Analysis," in *ICEIS*, 2017, vol. 1, pp. 351–356.

[7] O. Altuner, B. Bakir-Gungor, and G. Bakal, "Graph-based Biomedical Knowledge Discovery," in *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, Mersin, Turkey, 2024, pp. 1–4, doi: 10.1109/SIU61531.2024.10600774.

[8] Z. Zhang, M. Fang, R. Wu, H. Zong, H. Huang, Y. Tong, Y. Xie, S. Cheng, Z. Wei, M. Crabbe, X. Zhang, and Y. Wang, "Large-scale biomedical relation extraction across diverse relation types: model development and usability study on COVID-19," *J. Med. Internet Res.*, vol. 25, e48115, 2023, doi: 10.2196/48115.

[9] M. Krallinger, F. Leitner, and A. Valencia, "Analysis of biological processes and diseases using text mining approaches," in *Bioinformatics Methods in Clinical Research*, 1st ed., M. Ochs, Ed. New York: Springer, 2009, pp. 341–382, doi: 10.1007/978-1-60327-194-3_16.

[10] Y. Qi, Y. Zhang, and M. Song, "Text mining for bioinformatics: State of the art review," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, Beijing, China, 2009, pp. 88–91, doi: 10.1109/ICCSIT.2009.5234922.

[11] M. Krallinger, R. Erhardt, and A. Valencia, "Text-mining approaches in molecular biology and biomedicine," *Drug Discovery Today*, vol. 10, no. 6, pp. 439–445, Jun. 2005, doi: 10.1016/S1359-6446(05)03376-3.

[12] D. Hoksza and J. Jelinek, "Using Neo4j for mining protein graphs: A case study," in *Proc. 2015 26th Int. Workshop Database Expert Syst. Appl. (DEXA)*, 2015, pp. 333–337, doi: 10.1109/DEXA.2015.59.

[13] V. K. Gajendran, J.-R. Lin, and D. P. Fyhrie, “An application of bioinformatics and text mining to the discovery of novel genes related to bone biology,” *Bone*, vol. 40, no. 5, pp. 1378–1388, 2007, doi: 10.1016/j.bone.2006.12.067.

[14] K. L. Berg, T. Seymour, and R. Goel, “History of Databases,” *Int. J. Manag. Inf. Syst.*, vol. 17, no. 1, pp. 29–35, Jan. 2013.

[15] S. Ahsan and A. Shah, “Data, Information, Knowledge, Wisdom: A Doubly Linked Chain?” in *Proc. 2006 Int. Conf. Inf. Knowl. Eng.*, Las Vegas, NV, USA, Jun. 2006, p. 273.

[16] S. George, “NOSQL - NOTONLY SQL,” *Int. J. Enterp. Comput. Bus. Syst.*, vol. 2, no. 2, 2013.

[17] S. Eken, F. Kaya, S. Aslan, and K. Arslan, "Document-Based NoSQL Databases: A Comparison of Horizontal Scalability of MongoDB and CouchDB," presented at the 7th *Eng. Technol. Symp.*, Ankara, Turkey, 2014.

[18] G. Yılan, “NoSQL Nedir?” *Medium*, 2020. [Online]. Available: https://gamzeyilan1.medium.com/nosql-veritabanları-c94d1f23617c. [Accessed: Nov. 1, 2024]. (in Turkish)

[19] Amazon, “Why should you use a NoSQL database?” *AWS*, 2023. [Online]. Available: https://aws.amazon.com/tr/nosql/. [Accessed: Nov. 2, 2024].

[20] Oracle, “Advantages of NoSQL databases,” 2023. [Online]. Available: https://www.oracle.com/tr/database/nosql/what-is-nosql/. [Accessed: Nov. 5, 2024].

[21] B. Chicho and A. O. Mohammed, “An empirical comparison of Neo4j and TigerGraph databases for network centrality,” *Sci. J. Univ. Zakho*, vol. 11, no. 2, 2013.

[22] E. Redmond and J. R. Wilson, *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*, 1st ed. America: The Pragmatic Bookshelf, 2012.

[23] L. Warchal, “Using Neo4j Graph Database in Social Network Analysis,” *Studying Computer Science*, no. 2A(105), pp. 271–279, Dec. 2012.

[24] J. Partner, A. Vukotic, and N. Watt, *Neo4j in Action*, 1st ed. New York: Manning Publications, 2012.

[25] J. J. Miller, "Graph Database Applications and Concepts with Neo4j," in *Proceedings of the Southern Association for Information Systems Conference*, USA, Mar. 2013, pp. 141–147.

[26] Neo4j, "What's Neo4j?", Neo4j Documentation. [Online]. Available: https://neo4j.com/docs/getting-started/whats-neo4j/. [Accessed: Nov. 11, 2024].

[27] O. Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology,” *Nucleic Acids Research*, vol. 32, suppl. 1, pp. D267–D270, 2004.

[28] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, “SemMedDB: a PubMed-scale repository of biomedical semantic predications,” *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012.

[29] H. Kilicoglu, G. Rosemblat, M. Fiszman, and D. Shin, “Broad-coverage biomedical relation extraction with SemRep,” *BMC Bioinformatics*, vol. 21, pp. 1–28, 2020.

[30] S.-Z. Jiang, W. Lu, X.-F. Zong, H.-Y. Ruan, and Y. Liu, "Obesity and hypertension," *Experimental and Therapeutic Medicine*, vol. 12, no. 4, pp. 2395–2399, Sep. 2016, doi: 10.3892/etm.2016.3667.

[31] G. Seravalle and G. Grassi, "Obesity and Hypertension," in *Springer eBooks*, 2024, pp. 65–79, doi: 10.1007/978-3-031-62491-9_5.

[32] A. Gambineri, C. Pelusi, V. Vicennati, et al., "Obesity and the polycystic ovary syndrome," *International Journal of Obesity*, vol. 26, pp. 883–896, 2002, doi: 10.1038/sj.ijo.0801994.

[33] A. Couto Alves, B. Valcarcel, V. P. Mäkinen, et al., "Metabolic profiling of polycystic ovary syndrome reveals interactions with abdominal obesity," *International Journal of Obesity*, vol. 41, pp. 1331–1340, 2017, doi: 10.1038/ijo.2017.126.

[34] Y. Zhang, J. Li, G. Mo, J. Liu, H. Yang, X. Chen, and W. Huang, "Acupuncture and Related Therapies for Obesity: A Network Meta-Analysis," *Evidence-Based Complementary and Alternative Medicine*, vol. 2018, pp. 1–20, 2018, doi: 10.1155/2018/9569685.

[35] F. P. Kan, S. Raoofi, S. Rafiei, S. Khani, H. Hosseinifard, F. Tajik, N. Raoofi, S. Ahmadi, S. Aghalou, F. Torabi, A. Dehnad, S. Rezaei, Z. Hosseinipalangi, and A. Ghashghaee, "A systematic review of the prevalence of anxiety among the general population during the COVID-19 pandemic," *Journal of Affective Disorders*, vol. 293, pp. 391–398, 2021, doi: 10.1016/j.jad.2021.06.073.

[36] L. Sher, "COVID-19, anxiety, sleep disturbances and suicide," *Sleep Medicine*, vol. 70, p. 124, Jun. 2020, doi: 10.1016/j.sleep.2020.04.019.

[37] J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, and H. C. Lane, "Remdesivir for the Treatment of Covid-19 — Final Report," *New England Journal of Medicine*, vol. 383, no. 19, pp. 1813–1826, 2020, doi: 10.1056/nejmoa2007764.

[38] R. L. Gottlieb, C. E. Vaca, R. Paredes, J. Mera, B. J. Webb, G. Perez, G. Oguchi, P. Ryan, B. U. Nielsen, M. Brown, A. Hidalgo, Y. Sachdeva, S. Mittal, O. Osiyemi, J. Skarbinski, K. Juneja, R. H. Hyland, A. Osinusi, S. Chen, G. Camus, M. Abdelghany, S. Davies, N. Behenna-Renton, F. Duff, F. M. Marty, M. J. Katz, A. A. Ginde, S. M. Brown, J. T. Schiffer, and J. A. Hill, "Early Remdesivir to Prevent Progression to Severe Covid-19 in Outpatients," *New England Journal of Medicine*, vol. 386, no. 4, pp. 305–315, 2022, doi: 10.1056/NEJMoa2116846.

[39] E. A. Coomes and H. Haghbayan, "Interleukin-6 in Covid-19: A systematic review and meta-analysis," *Reviews in Medical Virology*, vol. 30, no. 6, pp. 1–9, 2020, doi: 10.1002/rmv.2141.

[40] E. Grifoni, A. Valoriani, F. Cei, R. Lamanna, A. M. G. Gelli, B. Ciambotti, and L. Masotti, "Interleukin-6 as prognosticator in patients with COVID-19," *Journal of Infection*, 2020, doi: 10.1016/j.jinf.2020.06.008.

[41] K. Papamichael, S. Lin, M. Moore, G. Papaioannou, L. Sattler, and A. S. Cheifetz, "Infliximab in inflammatory bowel disease," *Therapeutic Advances in Chronic Disease*, vol. 10, 2019, doi: 10.1177/2040622319838443.

[42] S. Danese, “Mechanisms of action of infliximab in inflammatory bowel disease: an anti-inflammatory multitasker,” *Digestive and Liver Disease*, vol. 40, Suppl. 2, pp. S225–S228, 2008, doi: 10.1016/S1590-8658(08)60530-7.

[43] F. Gomollón and J. P. Gisbert, “Anemia and inflammatory bowel diseases,” *World Journal of Gastroenterology*, vol. 15, no. 37, pp. 4659–4665, Oct. 2009, doi: 10.3748/wjg.15.4659.

[44] P. Oustamanolakis, I. E. Koutroubakis, and E. A. Kouroumalis, “Diagnosing anemia in inflammatory bowel disease: Beyond the established markers,” *Journal of Crohn's and Colitis*, vol. 5, no. 5, pp. 381–391, Oct. 2011, doi: 10.1016/j.crohns.2011.03.010.

[45] R. A. DeFronzo and A. M. Goodman, “Efficacy of Metformin in Patients with Non-Insulin-Dependent Diabetes Mellitus,” *New England Journal of Medicine*, vol. 333, no. 9, pp. 541–549, 1995, doi: 10.1056/NEJM199508313330902.

[46] A. Saenz, I. Fernandez-Esteban, A. Mataix, M. Ausejo Segura, M. Roqué i Figuls, and D. Moher, “Metformin monotherapy for type 2 diabetes mellitus,” *Cochrane Database of Systematic Reviews*, no. 3, 2005, doi: 10.1002/14651858.CD002966.

[47] N. Xita and A. Tsatsoulis, “Adiponectin in Diabetes Mellitus,” *Current Medicinal Chemistry*, vol. 19, no. 32, pp. 5451–5458, 2012, doi: 10.2174/092986712803833182.

[48] M. Howlader, M. I. Sultana, F. Akter, and M. M. Hossain, “Adiponectin gene polymorphisms associated with diabetes mellitus: A descriptive review,” *Heliyon*, vol. 7, no. 8, p. e07851, 2021, doi: 10.1016/j.heliyon.2021.e07851.

CURRICULUM VITAE

2016 – 2021	B.Sc., Electircal and Electronis Engineering, Erciyes University, Kayseri, TURKEY
2021 – Present	M.Sc., Electrical and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

(C1) O. Altuner, B. Bakir-Gungor, and G. Bakal, "Graph-based Biomedical Knowledge Discovery," in *Proceedings of the 32nd Signal Processing and Communications Applications Conference (SIU)*, Mersin, Turkey, 2024.