FINDING POTENTIAL SERIOUS ADVERSE EVENTS OF DRUGS BY USING CLINICAL TRIAL DATA AND MACHINE LEARNING TOOLS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

VEYSEL BUĞRA DEMİR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOTECHNOLOGY

DECEMBER 2021

Approval of the thesis:

FINDING POTENTIAL SERIOUS ADVERSE EVENTS OF DRUGS BY USING CLINICAL TRIAL DATA AND MACHINE LEARNING TOOLS

submitted by **VEYSEL BUĞRA DEMİR** in partial fulfillment of the requirements for the degree of **Master of Science in Biotechnology, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Greaduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Güzin Candan Gültekin
Head of Department, **Biotechnology** ________________

Asst. Prof. Dr. Aybar Can Acar
Supervisor, **Biotechnology, METU** ________________

Prof. Dr. Tolga Can
Co-Supervisor, **Computer Engineering, METU** ________________

Examining Committee Members:

Assoc. Prof. Dr. Çağdaş Devrim Son
Biology, METU ________________

Asst. Prof. Dr. Aybar Can Acar
Biotechnology, METU ________________

Assoc. Prof. Dr. Tunca Doğan
Computer Engineering, Hacettepe University ________________

Date: 30.12.2021

I hereby declare that all information in this document has been obtained and presented 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 material and results that are not original to this work.

Name Last name : Veysel Buğra Demir

Signature :

ABSTRACT

FINDING POTENTIAL SERIOUS ADVERSE EVENTS OF DRUGS BY USING CLINICAL TRIAL DATA AND MACHINE LEARNING TOOLS

Demir, Veysel Buğra
Master of Science, Biotechnology
Supervisor : Asst. Prof. Dr. Aybar Can Acar
Co-Supervisor: Prof. Dr. Tolga Can

December 2021, 129 pages

Healthcare is improving day by day and these developments make healthcare more accessible and this leads to production of large amount of data. The interpretation of these data, making assumptions and revealing significant results by using data analysis methods are important here as in every field that produces big data. Analysed data that are collected during clinical trials have great effect in ensuring developments in healthcare. Adverse event reports are one of the important parts of the clinically studied drugs or treatments. Evaluating the safety of an anticancer treatment through serious adverse events in clinical practice can provide important information for future cancer treatments. In this study, we propose a method to discover the links between serious adverse events and drugs, and between serious adverse events themselves. Our hypothesis is that, it can be possible to estimate the drug specific serious adverse events that occur together by using the clinical trial data that are transformed into a table structure for turning data into information to provide significant insights. We used ClinicalTrials.gov to download the clinical trial results that reported serious adverse events and in particular studied the anticancer drugs Cytarabine, Sorafenib and Doxorubicin. We used the MeSH and the CTCAE thesaurus to assign unique IDs to the serious adverse events to handle the inconsistency in the reports of serious adverse events. t-SNE and DBSCAN combination was used to find similar serious adverse events. To cluster the serious adverse events based on the drugs, we used spectral co-clustering. With the help of the hierarchical structure of the thesaurus, the p-values of the root and parent events were calculated to find significant ones that are encountered more or less, relatively, in a specific drug. Most of the results of this study are compatible with the available sources in the literature and the approach provided could predict the serious adverse events that are specific to new treatment options.

Keywords: Serious Adverse Events, Clinical Trials, Anticancer Drugs, Data Analysis, Clustering Algorithms.

ÖZ

KLİNİK DENEY VERİLERİ VE MAKİNE ÖĞRENMESİ ARAÇLARI KULLANILARAK İLAÇLARIN POTANSİYEL CİDDİ ADVERS ETKİLERİNİN BULUNMASI

Demir, Veysel Buğra
Yüksek Lisans, Biyoteknoloji
Tez Yöneticisi: Dr. Öğr. Üyesi Aybar Can Acar
Ortak Tez Yöneticisi: Prof. Dr. Tolga Can

Aralık 2021, 129 sayfa

Sağlık hizmetleri her geçen gün gelişmektedir ve bu gelişmeler sağlık hizmetlerini daha erişilebilir kılmakta ve bu da büyük miktarda veri üretilmesine yol açmaktadır. Büyük veri üreten her alanda olduğu gibi burada da bu verilerin yorumlanması, varsayımlarda bulunulması ve veri analiz yöntemleri kullanılarak anlamlı sonuçların ortaya çıkarılması önemlidir. Klinik araştırmalar sırasında toplanan verilerin analizi, sağlık hizmetlerinde gelişmelerin sağlanmasında büyük etkiye sahiptir. Advers etki raporları, klinik olarak çalışılan ilaç veya tedavilerin önemli kısımlarından biridir. Klinik uygulamada, ciddi advers etkileri aracılığıyla bir antikanser tedavisinin güvenliğinin değerlendirilmesi, gelecekteki kanser tedavileri için önemli bilgiler sağlayabilir. Bu çalışmada, ciddi advers etkiler ile ilaçlar arasındaki bağlantıları ve ciddi advers etkilerin kendileri arasındaki bağlantıları keşfetmek için bir yöntem öneriyoruz. Hipotezimiz, önemli içgörüler sağlamak amacıyla verileri bilgiye çevirmek için bir tablo yapısına dönüştürülen klinik araştırma verilerinin kullanılmasıyla, birlikte ortaya çıkan ilaca özgü ciddi advers etkileri tahmin etmenin mümkün olabileceğidir. Ciddi advers etkileri bildiren ve özellikle Cytarabine, Sorafenib ve Doxorubicin antikanser ilaçlarını çalışan klinik araştırmaların sonuçlarını indirmek için ClinicalTrials.gov'u kullandık. Ciddi advers etki raporlarındaki tutarsızlığın üstesinden gelmek amacıyla ciddi advers etkilere benzersiz kimlikler atamak için MeSH ve CTCAE sözlükleri kullanıldı. Benzer ciddi advers etkileri bulmak için t-SNE ve DBSCAN kombinasyonu ve ilaçlara dayalı ciddi advers etkileri kümelemek için spektral birlikte kümeleme algoritması kullanıldı. Sözlüklerin hiyerarşik yapısı yardımıyla, belirli bir ilaçta nispeten az ya da çok karşılaşılan anlamlı olayları bulmak için kök ve ana olayların p-değerleri hesaplandı. Bu çalışmanın sonuçlarının çoğu, literatürdeki mevcut kaynaklarla uyumludur ve sağlanan yaklaşım, yeni tedavi seçeneklerine özgü ciddi advers etkileri öngörebilir.

Anahtar Kelimeler: Ciddi Advers Etkiler, Klinik Denemeler, Antikanser İlaçlar, Veri Analizi, Kümeleme Algoritmaları.

To My Family

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Asst. Prof. Dr. Aybar Can Acar, for his guidance, unending support, encouragement and patience. Without his invaluable knowledge and recommendations, this thesis would not be possible. I also would like to thank my co-advisor Prof. Dr. Tolga Can, for his advices and constructive criticisms. I sincerely appreciate their support and input.

I would like to thank Assoc. Prof. Dr. Nurcan Tunçbağ for the courses that helped me in this study and Assoc. Prof. Dr. Can Özen for his encouragement and support in choosing the right path for my graduate studies.

I am also grateful to the members of the thesis examining committee for accepting to be in the committee and also for their questions and precious comments during the evaluation of my study.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

AML	Acute Myeloblastic Leukemia
ANC	Absolute Neutrophil Count
ARA-C	Arabinosylcytosine
CSV	Comma Separated Values
CTCAE	Common Terminology Criteria for Adverse Events
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNA	Deoxyribonucleic Acid
EHR	Electronic Health Record
ERK	Extracellular Signal-Regulated Kinase
FDAMA	Food and Drug Administration Modernization
INR	International Normalized Ratio
MAPK	Mitogen Activated Protein Kinase
MedDRA	Medical Dictionary for Regulatory Activities
MEK	Mitogen-Activated Protein Kinase Kinase
MeSH	Medical Subject Headings
NCI	National Cancer Institute
NIH	National Institutes of Health
NLM	National Library of Medicine
NLTK	Natural Language Toolkit
NOS	Not Otherwise Specified
PDGFR	Platelet Derived Growth Factor Receptor
PRR	Propotional Reporting Ratio
RAF	Rapidly Accelerated Fibrosarcoma
RNA	Ribonucleic Acid
SOC	System Organ Class
t-SNE	t-Distributed Stochastic Neighbor Embedding

UMLS	Unified Medical Language System
VEGFR	Vascular Endothelial Growth Factor Receptor
WHO	World Health Organization
XML	Extensible Markup Language

LIST OF SYMBOLS

SYMBOLS

ε	Radius
χ^2	Chi-square Test
α	Incidence Rate

CHAPTER 1

INTRODUCTION

1.1 Motivation

Healthcare is improving day by day due to developments in science and technology. These developments make healthcare more accessible and this leads to the use of healthcare services by more people, and more people mean large amounts of data. On average, a patient generates 80 MB of data per year [1] and it is easily predictable that this will increase in the future. The interpretation of these data, making assumptions and revealing significant results by using data analysis methods are important here as in every field that produces big data.

Analysing the data that is provided from healthcare systems has the potential to highly influence medical research, service planning and health policy [2]. These analyses produce better estimates, improve patient care and treatment decisions [3]. As a result of analyzing the raw form of data, certain patterns can be found and high amounts of data turn into actionable knowledge and provide insights to decision makers [4]. These insights either make decisions more precise or provide unique approaches. Converting analyzed data into a graph structure provides even more insight about the data. Due to the improvements it offers and the hidden patterns, correlations and trends it uncovers, data analysis and data visualization are important in the field of healthcare.

One of the major data sources for healthcare is clinical data. Clinical data make great contributions to measuring and monitoring outcomes, improving and developing pre-existing methods and accelerating innovation and exploration. Thus it has great potential in the fields of healthcare, health research and medical research [5]. Clinical data have many types and one of them is clinical trial data [6] which is also the main data source of this study.

Clinical trials are important for the development of treatments and thanks to these we have the high standards of medical care [7]. Analysed data that are collected during clinical trials have great effect in ensuring developments in healthcare [8]. Clinical trials informs clinicians and decision makers about the efficacy and safety of treatments [9]. Furthermore, clinical trials can provide information about the adverse effects of medical interventions or treatments by controlling the variable that may effect the results of the study [10].

Adverse event reports are one of the important parts of the clinically studied drugs or treatments. Measuring and monitoring safety of the study is standardized using adverse event and serious adverse event reporting protocols [11]. An *adverse effect* is an unexpected outcome, which is an unfavorable and dentrimental change in the health of a participant, that happens during or within a certain amount of time after a drug treatment or any medical intervention with at least a reasonable possibility of causal relationship. On the other hand, an *adverse event* is more or less the same thing but it is not necessarily caused by the intervention or treatment being studied meaning that causality is unknown [12], [13]. Serious adverse events are any adverse events that has serious medical consequences [12] and they are also the main data source from collected clinical trial reports for this study. A serious adverse event can result in; death, life threathening condition, inpatient hopitalization, prolongation of hospitalization, permanent or significant disability or incapacity, congenital anomaly and birth defect [14].

Any medical events can be considered as serious adverse events if they put participant in danger, require medical or surgical interventions to protect them from above-listed consequences even though they do not result in death, they do not require hospitalization or they are not life-threatening [13].

If the casuality is unknown, adverse event can be a general term for adverse effect, adverse reaction or adverse drug reaction [12]. Due to the use of cancer drugs, adverse drug reactions are an important cause of morbidity and mortality. Serious adverse drug reactions are frequent outcomes in oncology practices where cancer drugs are widely used [15] since they are essentially designed to be toxic [16]. It is particularly important to consider adverse drug reactions associated with cancer drugs because these drugs are notably likely to cause these kind of reactions since they are cytotoxic and therefore often damage normal cells in addition to malignant cells [16]. So, overall occurrence of serious adverse events is high and some of them are specific to chemotherapy. For example the global incidence rate of serious adverse events is 44.5% in a study that consists of 1000 patients [17]. While some of adverse events are specific to molecules that make the drug toxic to certain organs, some of them are common to all types of chemotherapy drugs because of their action mechanisms [17].

Evaluating the safety of an anticancer treatment through adverse drug events in clinical practice can provide important information for future cancer treatments [18]. The World Health Organization (WHO) recognized the importance of having an efficient way to monitor adverse drug events. This is the basis of the International Drug Monitoring Program [19]. Also, since even the preventable adverse events are estimated to cost aroud $3.5 billion annually just in the United Sates alone, adverse drug events constitute one of the most substantial types of healthcare adverse events [20]. For these reasons monitoring, analyzing and investigating the nature and frequence of serious adverse drug events are worth to be considered.

Based on the importance of clinical trial studies and serious adverse events arising from anticancer drugs, this study tries to establish a connection between drugs, treatments and adverse events and develop certain insights based on this connection by using the adverse event data that are collected from various clinical trial studies. Methods of data analysis and data visualization is used for their importance and effectiveness in the delivering insights from raw data.

1.2 Aim of the Thesis

The aim of this thesis is to discover links between drugs and serious adverse events and finding serious adverse events that trigger each other or occur together by measuring the similarities of serious adverse events.

Our hypothesis is that, with the proposed methods in this study it can be possible to estimate the drug specific serious adverse events that occur together by using the clinical trial data that are transformed into a table which is then data mined to provide significant insights. We expect that the results of this study will be compatible with the available sources in the literature and the approach provided could predict the serious adverse events that are specific to new treatment options.

1.3 Outline of the thesis

The clinical trial results were downloaded from ClinicalTrials.gov and they were transformed into a table. They were analyzed to turn data into knowledge, to find certain patterns and to provide significant insights.

In Chapter 2, background and related work is presented. The methods and tools used in this study are explained and previous studies are reviewed.

In Chapter 3, the methodology is explained in detail. The retrieval and transformation processes of the data and how the algorithms were applied to the data are explained. The pipeline chart is also provided to make the following steps easier to understand.

In Chapter 4, the results of the applied algorithms are provided. It contains visualization of the data along with the p-value calculations. The results of these analyses are also available as a Jupyter Notebook at:

github.com/VBugraDemir/ClinicalTrials/blob/main/Thesis-Results.ipynb

In Chapter 5, the discussion and conclusion are given along with the potential future works.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Tools and Methods Used

2.1.1 Source/Database

Clinicaltrials.gov is a resource which is web-based and publicly accessible registry of clinical trials and it was established by National Library of Medicine (NLM) at the National Institutes of Health (NIH) in 2000 for the purpose of having an online information source that makes researching easy by using clinical data among all registered clinical trials [21]. It was created due to the need for a publicly accessible online information source that arose as a result of the Food and Drug Administration Modernization Act (FDAMA) of 1997. FDAMA made necessary to establish a registry of clinical trials for federally, privately and publicly funded trials to test efficacy of new or experimental drugs for deadly or life-threatening conditions [22].

Clinicaltrials.gov provides information which includes all recruiting, active, suspended, terminated, withdrawn and completed clinical studies and their results if it is available [23]. Currently it contains more than 370.000 research studies from 220 countries and it is the most robust one among the other clinical trial registries [24]. The database has two main components; clinical study registry and results database [25]. The results database of the clinicaltrials.gov was launched in 2008 which made the submission of the basic results mandatory, at least 1 year after the completion date of the study. Early on when the result database was first released, the submission of adverse event information was optional but it has been required since 2009 [26]. Result databases provide systematic reporting of registered clinical trials in a complete, structured and timely manner [27]. While registry data elements are downloadable and available in list or spreadsheet format such as rows are records and data elements are columns, results data elements are only available in XML since data structure of result data tables has high degree of variability [25].

Clinicaltrials.gov provides valuable, easy to access information to patients, clinicians, researchers, healthcare professionals and the public on a wide variety of diseases and conditions [22]. It creates a considerable opportunity to understand clinical research better, increases public awareness of clinical trials and also improves monitoring of research [24]. ClinicalTrials.gov encourages reporting of results and recording of biomedical and health-related studies on humans [28]. With the clinical trials registered regardless of the country in which the clinical study was conducted, Clinicaltrials.gov looks for cooperation with other countries and tries to establish a universal trial registration system [29].

The increased use of Clinicaltrials.gov may lead to facilitation of the systematic evaluation of clinical studies in order to establish a knowledge base that provides information about medical practice and prevention [30]. It becomes clear that the database is a unique resource for those researching clinical studies. Between 2010 and 2015, 404 research articles containing; adverse events, ethics and analyses of studies on specific conditions were published using the data provided from Clinicaltrials.gov [25]. For proper analyses, it is important to understand the structure of the database, its evolution over time, the organization of records and the factors that shape the content of the database [25].

2.1.2 Ontologies (Dictionaries) Used in This Study

2.1.2.1 Medical Subject Headings (MeSH)

Medical Subject Headings (MeSH) thesaurus is a *controlled vocabulary* [31], since it is a set of consistent terms used in a particular field of knowledge [32]. MeSH contains more than 200.000 terms [33] that are organized into hierarchies [34] which is called MeSH Tree Structures [33]. It was developed [35] and has been maintained by the National Library of Medicine (NLM) since 1960 [36]. As literature evolves and changes, NLM adds new terms or modifies existing ones [37]. Mesh was needed due to the necessity of indexing and cataloging of books, journal articles, research publications or any information related to the life sciences such as biomedical and health [38]. It is translated into many different languages and widely used around the world [35].

MeSH contains all the concepts that appear in the medical literature, hence searches are more focused and unambiguous when main headings and subheadings of MeSH are used [39]. Also, for each of the headings there are appropriate substitutes called entry terms that are used as pointers to the main headings which leads to association of different terms that indicates the same heading [35]. Concepts in the medical literature are represented by descriptors in MeSH and each descriptor has a MeSH heading name and unique ID along with other identifying details. For instance, stomach neoplasms has a unique ID of D013274 as a descriptor [40] and its hierarchical structure can be seen in Figure 2.1. So the concepts in MeSH can be used as keywords and are helpful in the information retrieval process [41]. That makes MeSH a powerful tool for performing a targeted search which gives narrower results [39] and it provides organized medical knowledge and information [42].

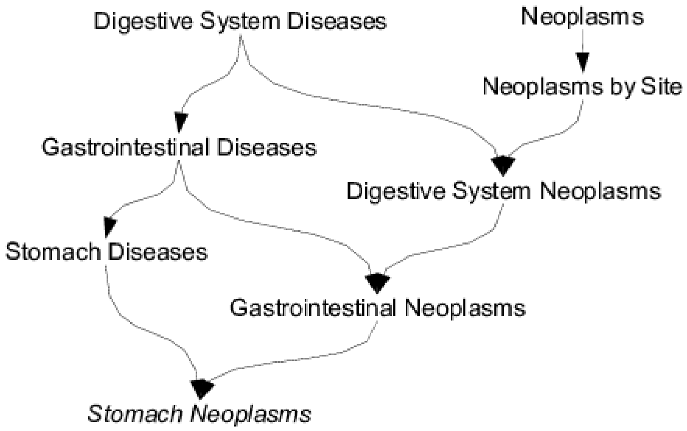

Digestive System Diseases
Neoplasms
Neoplasms by Site
Gastrointestinal Diseases
Digestive System Neoplasms
Stomach Diseases
Gastrointestinal Neoplasms
Stomach Neoplasms

Figure 2.1 Tree structure of stomach neoplasms [43].

The words used in the literature search involving medical topics vary from searcher to searcher, if they are not made uniform [44]. As a result of this, the desired results may not be obtained so the research may slow down and the accumulation of knowledge may be interrupted. MeSH ensures the language used by authors is compatible with a standard and that a searcher can find relevant studies using the MeSH terms since the author and searcher do not differ in how they describe a concept [37]. MeSH provides official words to represent any health-related concepts. The terms used for labeling an article should be selected from the official MeSH list by the indexer. For example if an article is about *heart attack*, the official MeSH term *myocardial infarction* should be used as label instead [45]. This regulation results in more efficient searching, saves time and effort and improves precision because after discovering the correct MeSH, the potential variants in spelling and the trends in terminology are not needed to be considered anymore [45].

As we want to discover links between adverse events, drugs and clinical trial studies in this study, we need to standardize adverse event terms since some of these terms change report to report and are not compatible with each other even when they represent the same condition. This stems from the lack of conformity to a standard in the clinical trial reports. To overcome this issue, MeSH was chosen to be the backbone used in assigning the same unique ID to different terms that represent the same condition while different unique ID's are assigned to different conditions. For example, in order to make the analyses reliable, avoid confusion and loss of information; "Head pain", "Pain: Head" and "Headache" all should have the same unique ID which is D006261 that represents the MeSH heading of "Headache" in this case. So, MeSH has served as a good starting point to catch the synonyms with its entry terms, and to create the adverse event knowledge base.

2.1.2.2 Common Terminology Criteria for Adverse Events (CTCAE)

Due to the toxicity of anticancer drugs, adverse events can occur in patients who receive the treatment [46]. Adverse events are increasing in the treatments of patients [47] with the introduction of new anticancer agents and this makes them a growing concern. Therefore, adverse events are an essential part of the patient care and require close collaboration and partnership that lead to uniform cataloging and grading of adverse events for efficient communication and better interventions [48].

In order to standardize the reporting and documentation of adverse events, the US National Cancer Institute (NCI) published the Common Terminology Criteria for Adverse Events (CTCAE v3.0) in 2003 [49]. It was the first standardized and complete descriptive terminology [50] that provides standardized definition criteria for identifying and grading adverse events [51]. The risks caused by the treatments has been understood more clearly and the comparison of different anticancer drugs and treatment methods has been made more easily with the widespread use of the CTCAE [50]. The CTCAE also used in clinical studies to evaluate adverse events with a grading system that makes possible to classify them [52] based on their severity [53]. Also, adverse events are collected and reported based on the CTCAE in cancer clinical trials [54]. Thus, it assists in the evaluation of new anticancer drugs or treatment methods for patient safety [55].

All the adverse event terms in the CTCAE have been mapped to the Medical Dictionary for Regulatory Activities (MedDRA) codes [56] and grouped by System Organ Class (SOC) which one of the hierarchy levels in MeDRA [51]. MedDRA is a medical terminology that is used to support health monitoring internationally [51]. The version used in this study (CTCAE v5.0) contains 837 [57] adverse event terms with their grading and unique MedDRA codes. For most of the adverse events, grades consist of Grade 1 through Grade 5 as mild and extremely severe respectively with their clinical definitions [57]. All the available information of hypotension in the CTCAE can be seen in Table 2.1.

Table 2.1 CTCAE grading for hypotension [58].

MedDRA Code	10021097
MedDRA SOC	Vascular disorders
CTCAE term	Hypotension
Grade 1	Asymptomatic, intervention not indicated
Grade 2	Non-urgent medical intervention indicated
Grade 3	Medical intervention indicated; hospitalization indicated
Grade 4	Life-threatening consequences and urgent intervention indicated
Grade 5	Death
Definition	A disorder characterized by a blood pressure that is below the normal expected for an individual in a given environment.

As a result of not conforming to a certain standard in the clinical trial reports, some of the clinicians stick to the MeSH terms, some to the terms in CTCAE, and some of them report adverse events in their own way. Therefore, in addition to MeSH, another terminology library was needed in the association of adverse event terms with their unique ID's when MeSH is insufficient. CTCAE is a good secondary source for this study, because it includes unique ID's for each term, is an adverse event focused terminology and contains some of the frequently used terms in the adverse event reporting that are not available in MeSH.

2.1.3 Anticancer Drugs Investigated in This Study

2.1.3.1 Sorafenib

Sorafenib is a small molecule that was designed as an oral inhibitor of a few kinases involved in tumor growth and tumor angiogenesis [59]. Sorafenib is approved for the treatment of thyroid [60], advanced renal cell carcinoma (kidney cancer) [61] and hepatocellular carcinoma [62] which is the most common type of primary liver cancers [63]. It delays the progression of tumor and prolongs survival in patients with carcinoma [62].

Although, initial intention was to develop a rapidly accelerated fibrosarcoma (RAF) inhibitor, subsequent studies of Sorafenib showed that it is actually a multikinase inhibitor for variety of tyrosine kinase receptors [61] such as vascular endothelial growth factor receptor (VEGFR) and platelet derived growth factor receptor (PDGFR) [62].

In the development of carcinoma, tumor angiogenesis plays an important role because it is the essential mechanism for tumor growth and metastasis [61]. Thus, inhibition of angiogenesis is a convenient target [61], [62]. VEGFR is involved in formation of blood vessels, endothelial cell proliferation and survival [64]. As a result of tumor angiogenesis; tumor cell proliferation, differentiation and survival is upregulated through the mitogen activated protein kinase (MAPK)/extracellular signal-regulated kinases (ERK) signal transduction pathway [62]. MAPK/ERK pathway (Figure 2.2) includes the kinase, MAPK (MEK) [65] and is usually activated in numerous cancers [66]. Sorafenib inhibits the MAPK/ERK pathway by biding to RAF, VEGFR and PDFGR [67]. Thus, growth and cell population of tumor are decreased by preventing the activation of the pathway [68].

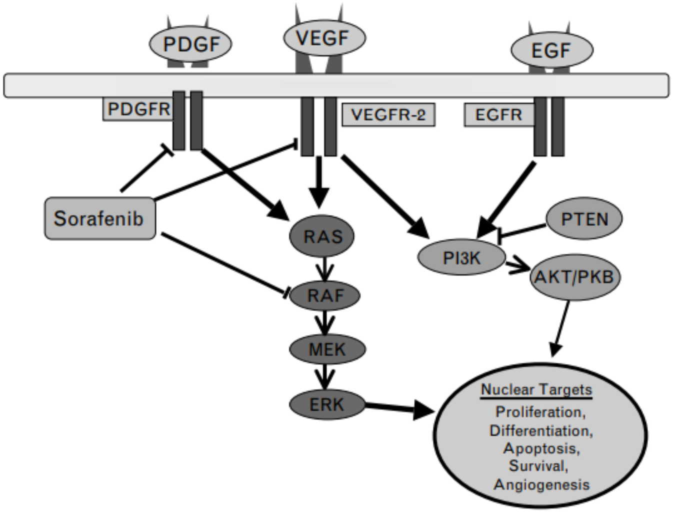


Figure 2.2. MAPK/ERK pathway and its inhinition by Sorafenib [61].

Discovery of pathways such as MAPK/ERK cascade has led to development of molecularly targeted therapies [62]. Targeted therapies aim at abnormal molecular pathways involve in growth or proliferation of tumor, unlike cytotoxic chemotherapy [69]. Molecular targeted therapies increase the efficiency of cancer therapy and increase the survival rate with less side effects than traditional cytotoxic chemotherapies [69], [70]. Thus, therapy with Sorafenib is generally well tolerated in patients with manageable adverse events. Diarrhea and hand-foot skin reaction are the most common adverse events caused by Sorafenib [62].

2.1.3.2 Cytarabine

Cytarabine, also known as arabinosylcytosine (ARA-C) [71], is an antineoplastic and antimetabolite agent [72]. It has also antiviral and powerful immunosuppressant properties [72], [73]. Since Cytarabine is a cell cycle-specific agent, it kills cells that synthesize deoxyribonucleic acid (DNA) and inhibits the synthesis of DNA [74]. It is mostly used in the treatment and management of lymphoma and leukemia [72]. It is especially an important agent to consider in the treatment of acute myeloid leukemia [71]. Cytotoxic concentrations around the cancer cell should be maintained for many days for Cytarabine to be most effective [75]. Thus, continuous infusion of Cytarabine is more toxic than discontinuous dosage, still low dose Cytarabine is well tolerated [76].

Cytarabine is converted to the active metabolite 5’ triphosphate ester in cell [77], then it incorporates into DNA strands to stop the synthesis by blocking the rotation of the DNA molecules with the sugar part within it [71], induces miscoding, interrupt the chain elongation and inhibits DNA polymerase [78]. The DNA replication is terminated during the cell cycle and this makes Cytarabine a drug specifically suited for cells which rapidly divide like in the case of cancer [71]. Some toxic effect can be expected on rapidly dividing normal cells like corneal epithelium because of the ceased replication and repair of DNA due to the inhibition of DNA polymerase by Cytarabine [71], [79].

Gastrointestinal toxicity is the major problem in both standard and high dose Cytarabine therapy. Since the gastrointestinal lining is composed of rapidly dividing tissues, it is highly sensitive to Cytarabine [76]. Vomiting and nausea are really common in the therapy with Cytarabine [73] and almost all patients treated with Cytarabine experience them [76]. Stomatitis (or oral mucositis) can be seen especially with the high dose regimens within a few days because of the susceptibility to infection due to the destruction of the oral mucosal barrier [76], [73]. Diarrhea also occurs with the high dose Cytarabine therapy with intestinal injury, since Cytarabine is toxic to the intestinal stem cells and inhibits epithelial regeneration [76]. Headache, fever and skin rash are other common reactions that may occur with the use of Cytarabine [76], [80].

2.1.3.3 Doxorubicin

Doxorubicin is a wide-spectrum antitumor antibiotic which is from the anthracycline group of chemotherapeutic agents [81]. Anthracyclines are used in the treatment of various cancer types and they are extracted from the Streptomyces peucetius bacterium [82], [83]. Because Doxorubicin is an anthracycline, it is widely used against solid tumors [81]. Unlike other anthracyclines, Doxorubicin has a broader spectrum of activity [84], so it is also used in the treatment of soft tissue and bone sarcomas as well. Lymphoblastic leukemia, acute myeloblastic leukemia (AML), small cell lung cancer and cancer of breast, ovary, bladder and thyroid are other types of cancers that Doxorubicin is used to treat [82].

Due to the Doxorubicin’s ability to inhibit DNA and cause cell apoptosis, it has antineoplastic activity [81]. Doxorubicin intercalates into DNA and inhibits the repair activity of topoisomerase II that results in breakage of DNA and inhibition of both DNA and ribonucleic acid (RNA) synthesis. Doxorubicin also generates free radicals such as reactive oxygen species that cause oxidative damage to membranes, proteins and DNA which especially further limits DNA synthesis [82], [85]. DNA damage, membrane damage and oxidative stress due to the reactive oxygen species trigger apoptotic pathways and lead cells to death [85].

Although Doxorubicin and all anthracyclines in general are very effective, their use is limited by their cardiotoxicity [81] which occurs with cumulative doses and can lead to congestive heart failure [84]. This limits the long term use of Doxorubicin [82]. Bone marrow suppression, nausea, vomiting, diarrhea and fatigue are other adverse events that occur due to use of Doxorubicin [84]. Doxorubicin also induces immunogenic cell death and that leads to systemic immunosuppression [86].

2.1.4 Clustering Methods Used

2.1.4.1 T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-Distributed Stochastic Neighbor Embedding or t-SNE is a machine learning algorithm that is widely used to explore high-dimensional data [87]. It is suitable for use in the visualization of high-dimensional data. It is a dimensionality reduction technique that embeds high-dimensional data in a low dimensional space such as two or three dimensions [88]. The reason it has become popular is that t-SNE can create two dimensional maps from data with thousands of dimensions [87]. To create that map, t-SNE assigns each data point to a location in two or three-dimensional space [88]. Patterns on the map represent the natural clustering of the data based on the similarity of data points [89].

In order to reduce the dimensionality properly, methods preserve the significant structures of the high-dimensional data in the low-dimensional map as much as possible. Several methods are proposed for this and classical ones are linear techniques that keep the dissimilar data points far apart and preserve the global structure of the data thus fail to preserve the similarities within the clusters. Keeping very similar data points together is not possible with linear methods but t-SNE is a nonlinear dimensionality reduction method which aims to preserve the local structure of data by aggregating similar data points (Figure 2.3). It visualizes the resulting similarity information by preserving most of the local structure of the high-dimensional data with well separated clusters. It also reveals global structure to some extent with the clusters at several scales [90]. t-SNE performs well as an embedding method [91] since local structures are preserved in small neighborhoods [92].

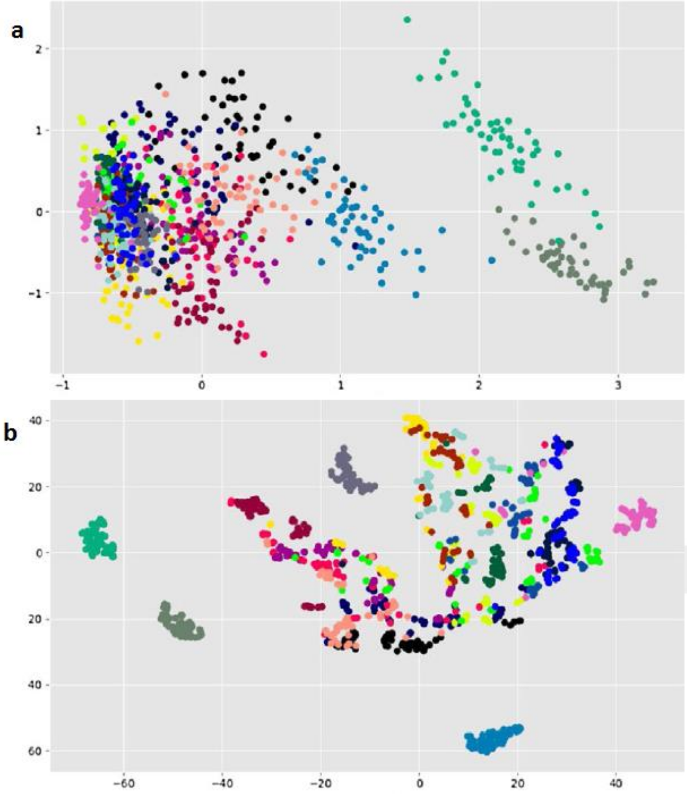


Figure 2.3. Visualization of dimension reduction results of a linear method (a) and t-SNE (b) on the same data [93].

t-SNE uses feature extraction to reduce dimensionality [94]. Feature extraction is a method that removes redundant data while keeping the necessary data. Thus it creates more manageable data from the initial raw data [95]. t-SNE defines similarity between the high and low-dimensional spaces by describing symmetric joint-probability distributions for both. The distributions measure the pairwise similarities between data points which are modeled based on the conditional probabilities for each data point, in order to pick another data point as its neighbor [96]. So, pairwise similarity creates affinity between data points (Figure 2.4). Data points with high affinity create close neighbors while distant data points have near zero attraction [97]. Thus, similar data points have a higher probability of being selected than dissimilar data points [98]. These pairwise similarities are based on the probability density under the Gaussian distribution in high-dimensional space and Student's t-distribution in the low-dimensional space [99]. To place the points in two dimension after defining the distributions [97], t-SNE minimizes a cost function which is the divergence between the two distributions [96] thus local structure of data points is preserved in both high and low-dimensional spaces [99].

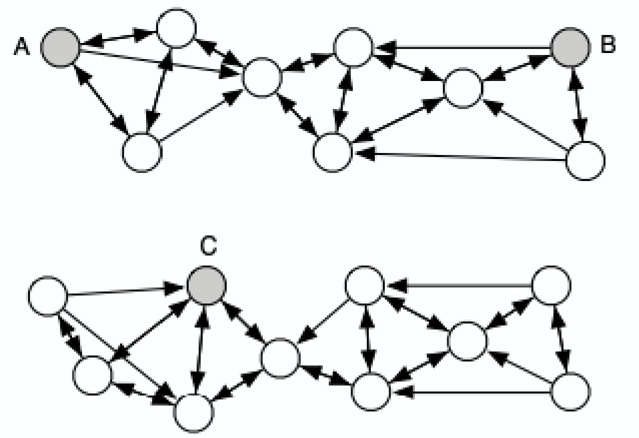


Figure 2.4. Clusters created based on the affinities between data points. The pairwise affinity is larger between A and B than between A and C, therefore A is in the same cluster with B rather than C [90].

t-SNE is one of the widely used dimensionality reduction techniques and has many applications. Overall, it can be considered as a visualization, pattern recognition, classification or compression method for big data sets with high dimensonality [89].

2.1.4.2 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-based spatial clustering of applications with noise (DBSCAN) is an unsupervised density-based clustering algorithm [100]. DBSCAN defines clusters as regions of dense data points that are surrounded and separated from each other by lower density regions [101]. It automatically discovers arbitrarily shaped clusters with varying densities and it can be efficiently used with large data sets [102].

Since widely used model-based clustering algorithms optimize the fit between data and a mathematical model [103], they assume that data are identically distributed and this makes it hard to identify arbitrary shaped clusters [104]. Unlike model-based clustering algorithms, DBSCAN and all density-based clustering algorithms can discover clusters with any arbitrary shape and size (Figure 2.5), even if there are noise and outliers [102], [103].

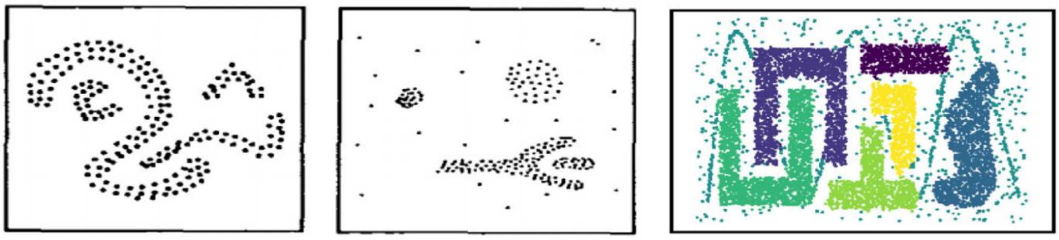


Figure 2.5. Some arbitrary shapes that can not be clustered by model-based clusters unlike DBSCAN [105], [106].

DBSCAN creates clusters automatically based on a dataset and two given parameters, so it does not need the number of clusters to be stated initially [102]. These two given inputs are radius (*Eps* or *ε*) and minimum number of objects (*MinPts*) [103]. Eps is the maximum accepted distance between two potential neighbors [107]. The objects within a radius ε of a given object constitute a neighborhood which is called the *ε-neighborhood* [108], [109]. And an ε-neighborhood has to contain at least MinPts objects to become a *core object* [108]. Points that are not a core point but belong to the ε-neighborhood of any core point, are defined as *border points* (Figure 2.6). Hence a border point is *density connected* [110]. On the other hand, points are considered as noise if they are neither a core point nor connected to any core point [108].

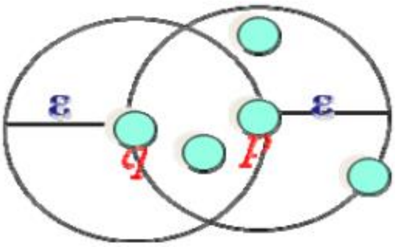


Figure 2.6. ε-neighborhood of p and q. MinPts is 4 and p is the core point. Because q is in the ε-neighborhood of p, q is directly density reachable from p [108].

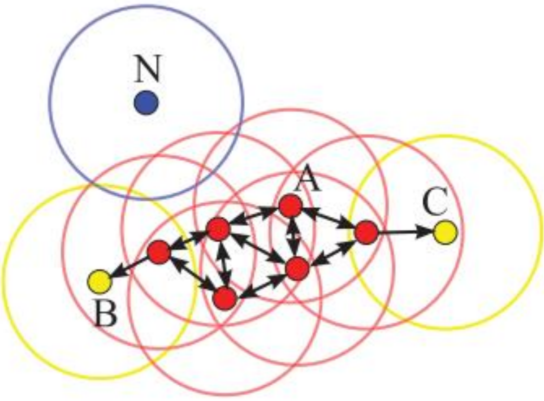


Figure 2.7. DBSCAN cluster illustration [111].

Figure 2.7 shows an example of a single cluster that is created with DBSCAN and formed by neighbor points. The MinPts is 4 and the circles indicate the radius. A and all the other red points are core points since there are at least 4 points (the points themselves are included) around them within a radius of Eps. Arrows indicate that they are all reachable from one another (direct density reachability). Although B and C are not core points they are in the cluster since they are density connected to point A and other core points. N is a noise point because it is not density reachable [111].

DBSCAN iteratively checks for ε-neighborhoods of each point in the data set and creates clusters. Finally, the process is terminated when there are no new points to add to any cluster [109].

2.1.4.3 Spectral Co-Clustering

Clustering is a substantial data exploration technique and it is fundamental in data mining [112], [113]. It is an unsupervised approach for data that has no predefined labels [114]. Clustering is used to group together similar objects by collecting closely related entities in the data set and assigning them to the same group [113], [115], whereas objects from different groups are dissimilar [116]. This reveals a new set of categories that leads to discover of hidden structure in sample data [117], [118]. Thus, with the use of clustering techniques, significant patterns can be found from data sets and hidden information can be obtained from raw data [116].

Unlike clustering which independently groups similar rows or columns in a data set [119], co-clustering looks for row and column blocks or submatrices that are intercorrelated [120]. To achieve intercorrelation, row cluster prototypes, that are created with co-clustering, incorporate column clustering information, and vice versa [115], [120]. So briefly, co-clustering is simultenous clustering of both rows and columns [121]. Utilizing the duality between rows and columns improves clustering of both objects and features [122], and this makes co-clustering a powerful data analysis technique [121]. Co-clustering has been received lots of attention and studied in several fields such as; text mining, bioinformatics, natural language processing and recommendation systems [119], [120].

There are many co-clustering formulations and the spectral co-clustering model is one of them [121]. Spectral co-clustering converts the co-clustering problem into a partitioning problem on a bipartite graph [123] such that the nodes correspond to rows and columns (Figure 2.8) in the input data martix [121]. Also, each cell of the input data matrix corresponds to an edge (and weight) in this bipartite graph [123].

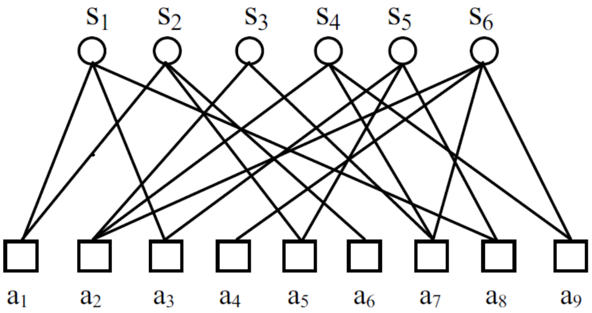


Figure 2.8. Example of bipartite graph with two kind of vertices (square and circular nodes). With partition of the graph, co-clustering of the data is achieved [124].

Spectral co-clustering is thus based on spectral graph theory. The weighted graph is constructed from the input data matrix as explained above. Therefore, each node of this graph represents a pattern and each weighted edge represents the similarity between two patterns. After that, the clustering problem turns into a *graph cut* problem (Figure 2.9) which can be solved with spectral graph theory [125]. The partitioning is an optimization procedure [126] where a minimum number of cuts are performed such that the weights across the cuts (i.e. edge weights of the nodes in different sub-graphs [120]) are also minimized [123].

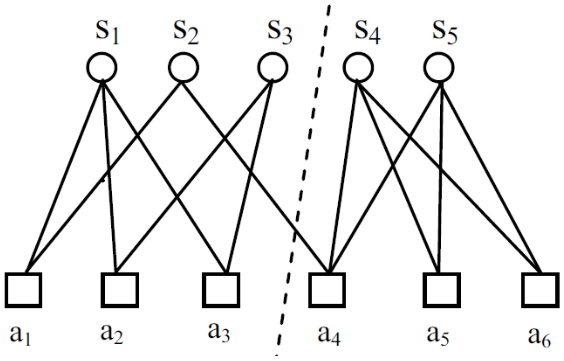


Figure 2.9. Example of an optimized cuts to partition the graph [124]

Spectral co-clustering is suitable for providing robust estimations from correlated data points and revealing structure that are difficult to observe due to noise or sparsity [126].

2.2 Literature Review

C. Federer et al. developed a database that links clinical trials, drugs, and adverse events. They retrieved clinical trials with adverse events results from the ClinicalTrials.gov and performed big data mining along with pattern analysis and data visualization on the published results. Their motivation is very similar to ours, since they thought that studying clinical trial data, and more spesifically adverse events, would provide new insights and reveal relationships between drugs and adverse events. They used propotional reporting ratio (PRR) to evaluate the adverse events. With PRR they basically compared the frequency of adverse events that are reported for the patients who take a certain drug and the frequency of the same adverse event that is reported in other drugs. They concluded that a PRR greater than one means that the drug of interest has a higher frequency of the AE than others. They used 10,786 trials with results (at least one reported adverse event). They limited their study by the FDA drugs by using an FDA-approved drug list to extract drug-AE relationships. After that, they grouped the adverse events into 26 unique adverse events categories that comply with the Common Terminology Criteria for Adverse Events (CTCAE v4.0). But they did not explain how did they do that since they also complained about the lack of standards in clinical trial data elements, different ontologies that were used in reporting adverse events and typos in data entry. These are the same problems we faced too while doing this study. The 10 most common adverse events they found in the clinical trials were headache, nausea, dizziness, vomiting, fatigue, constipation, diarrhea, back pain, nasopharyngitis (common cold), and cough. Also they found that the kinase inhibitors have higher numbers of adverse events, compared with other drugs and suggested that kinase inhibitors have more “off-targets” [127].

Jake Luo and Ron A. Cisler, integrated multiple adverse event reports from clinical trials to compare adverse events across different cancer drugs for better understanding of adverse events of cancer drugs. They looked for significant adverse events of drugs which is a similar thing we did in this study even though the approaches are different. They extracted 12,922 distinct adverse events from 1,602 cancer trials that were collected from ClinicalTrials.gov and selected 30 common cancer drugs. Like we did in this study, they standardized the adverse even terms but by using a different thesaurus. They mapped the extracted data elements to the Unified Medical Language System (UMLS) to standardize terminologies in the reports. They ranked all the adverse events based on their prevalence in the trials. Nausea was the top adverse event with a very high prevalence at 82.77%, followed by fatigue at 77.34%, vomiting at 75.97%, constipation at 72%, and cough at 63%. These are very similar to what C. Federer et al. found. The incidence rate of the adverse events was also calculated and alopecia (hair loss) was found to have the highest incidence rate at 26.43%. They suggested that the incidence rates accross different drugs can be significant and they looked for statistically associated drugs and adverse events. They grouped together trials that tested the same drug and compared them. For significant adverse events across different drugs they used Grubbs' test to evaluate the significance. They found several drug-adverse event associations such as; axitinib with hypertension, imatinib with muscle spasm, vorinostat with deep vein thrombosis and afatinib with paronychia [128].

Zitnik et al. developed Decagon, a model to predict side effects of drug combinations. Use of drug combinations or polypharmacy is a common application for patients with complex diseases or co-occurring conditions. However, this increases the risk of adverse side effects for the patient because of drug-drug interactions. Since these interactions are not that common, the knowledge about them is limited and it remains a challenge. Thus, they developed Decagon to model and discover polypharmacy side effects. The method creates a multimodal graph of protein–protein interactions, drug–protein target interactions and the polypharmacy side effects, which are represented as edges between drug–drug interactions. To predict the side effects; they used co-prescribed drugs that have common target proteins for drug-target protein information and they also used protein-protein interactions to model characteristics of drugs that have common side effects. By that, drug combination interaction and the exact type of side effect can be predicted [129].

CHAPTER 3

MATERIALS AND METHODS

3.1 Overview of the Methods

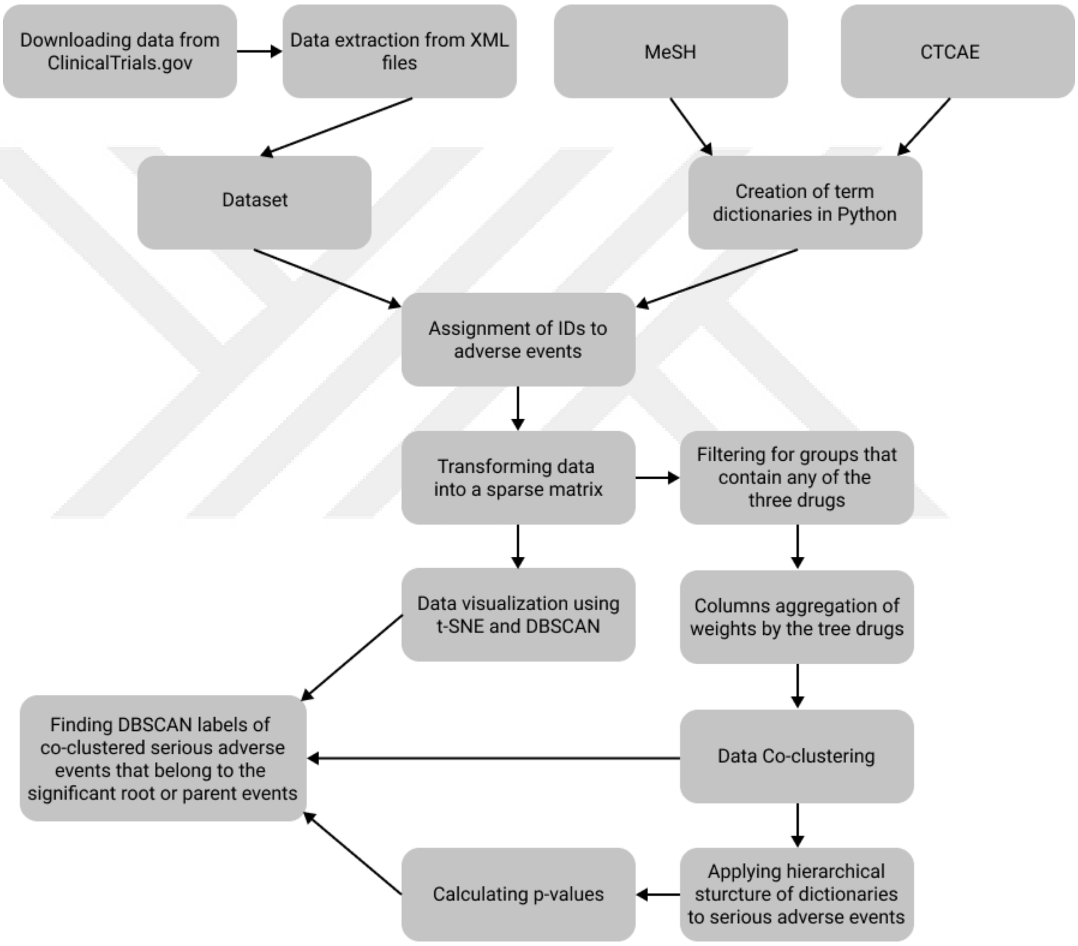
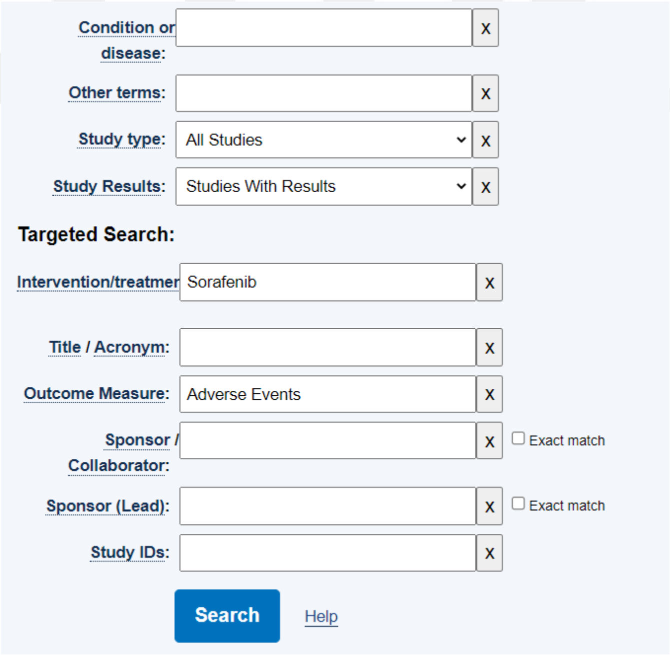

Downloading data from ClinicalTrials.gov
Data extraction from XML files
MeSH
CTCAE
Dataset
Creation of term dictionaries in Python
Assignment of IDs to adverse events
Transforming data into a sparse matrix
Filtering for groups that contain any of the three drugs
Data visualization using t-SNE and DBSCAN
Columns aggregation of weights by the tree drugs
Finding DBSCAN labels of co-clustered serious adverse events that belong to the significant root or parent events
Data Co-clustering
Calculating p-values
Applying hierarchical sturcture of dictionaries to serious adverse events

Figure 3.1. The pipeline of the methodology.

3.2 Dataset

3.2.1 Downloading Data From Clinicaltrials.gov

Clinical trials involving the use of Sorafenib, Cytarabine and Doxorubicin anticancer drugs were retrieved using the advanced search feature of Clinicaltrials.gov (Figure 3.2). The names of the cancer drugs were entered in the "Intervention/treatment" field under the "Targeted Search" section respectively for each drug. "Adverse events" was also entered in the "Outcome Measure" field under the same section since assessing serious adverse events based on clinical studies is the aim of this study. The searches were narrowed down by selecting the "Studies with Results" option in the "Study Results" field because studies without results could not provide any information for further data analyses. As a result of the search, 71 studies were found for Sorafenib, 177 for Doxorubucin and 99 for Cytarabine (Figure 3.3).



Condition or disease:

Other terms:

Study type: All Studies

Study Results: Studies With Results

Targeted Search:

Intervention/treatmen Sorafenib

Title / Acronym:

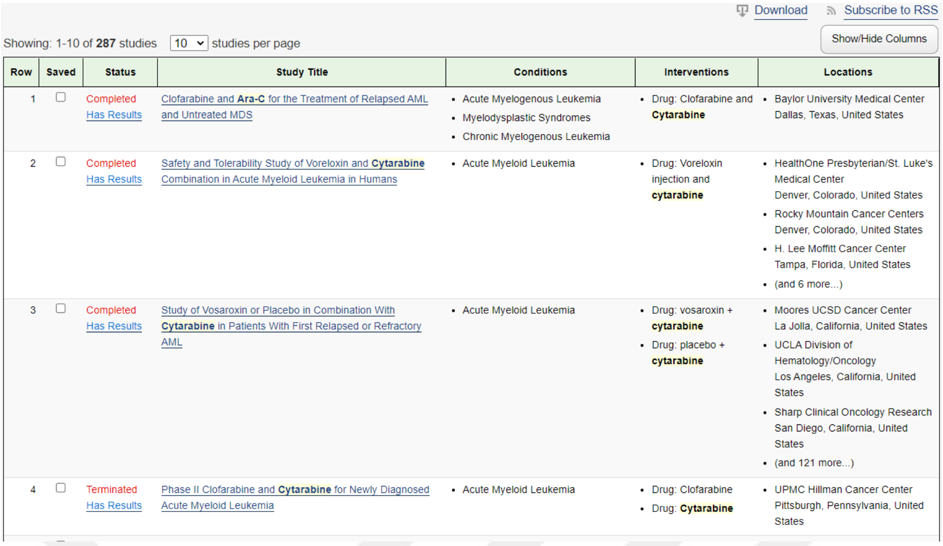
Outcome Measure: Adverse Events

Sponsor / Collaborator: □ Exact match

Sponsor (Lead): □ Exact match

Study IDs:

Figure 3.2. The advanced search form of the ClinicalTrials.gov [130].



Showing: 1-10 of **287** studies 10 studies per page

Row	Saved	Status	Study Title	Conditions	Interventions	Locations
1	□	Completed Has Results	Clofarabine and **Ara-C** for the Treatment of Relapsed AML and Untreated MDS	• Acute Myelogenous Leukemia • Myelodysplastic Syndromes • Chronic Myelogenous Leukemia	• Drug: Clofarabine and **Cytarabine**	• Baylor University Medical Center Dallas, Texas, United States
2	□	Completed Has Results	Safety and Tolerability Study of Voreloxin and **Cytarabine** Combination in Acute Myeloid Leukemia in Humans	• Acute Myeloid Leukemia	• Drug: Voreloxin injection and **cytarabine**	• HealthOne Presbyterian/St. Luke's Medical Center Denver, Colorado, United States • Rocky Mountain Cancer Centers Denver, Colorado, United States • H. Lee Moffitt Cancer Center Tampa, Florida, United States • (and 6 more...)
3	□	Completed Has Results	Study of Vosaroxin or Placebo in Combination With **Cytarabine** in Patients With First Relapsed or Refractory AML	• Acute Myeloid Leukemia	• Drug: vosaroxin + **cytarabine** • Drug: placebo + **cytarabine**	• Moores UCSD Cancer Center La Jolla, California, United States • UCLA Division of Hematology/Oncology Los Angeles, California, United States • Sharp Clinical Oncology Research San Diego, California, United States • (and 121 more...)
4	□	Terminated Has Results	Phase II Clofarabine and **Cytarabine** for Newly Diagnosed Acute Myeloid Leukemia	• Acute Myeloid Leukemia	• Drug: Clofarabine • Drug: **Cytarabine**	• UPMC Hillman Cancer Center Pittsburgh, Pennsylvania, United States

Figure 3.3. Result of the search for studies including Cytarabine as a treatment.

Each study is encoded as an individual XML file and all the study records were downloaded as a single zip file containing these XML files (347 files in total).

Sorafenib, Doxorubicin and Cytarabine were chosen because they have different characteristics and methods of actions. They are widely used in clinical trials which increases the diversity and the amount of the retrieved data respectively.

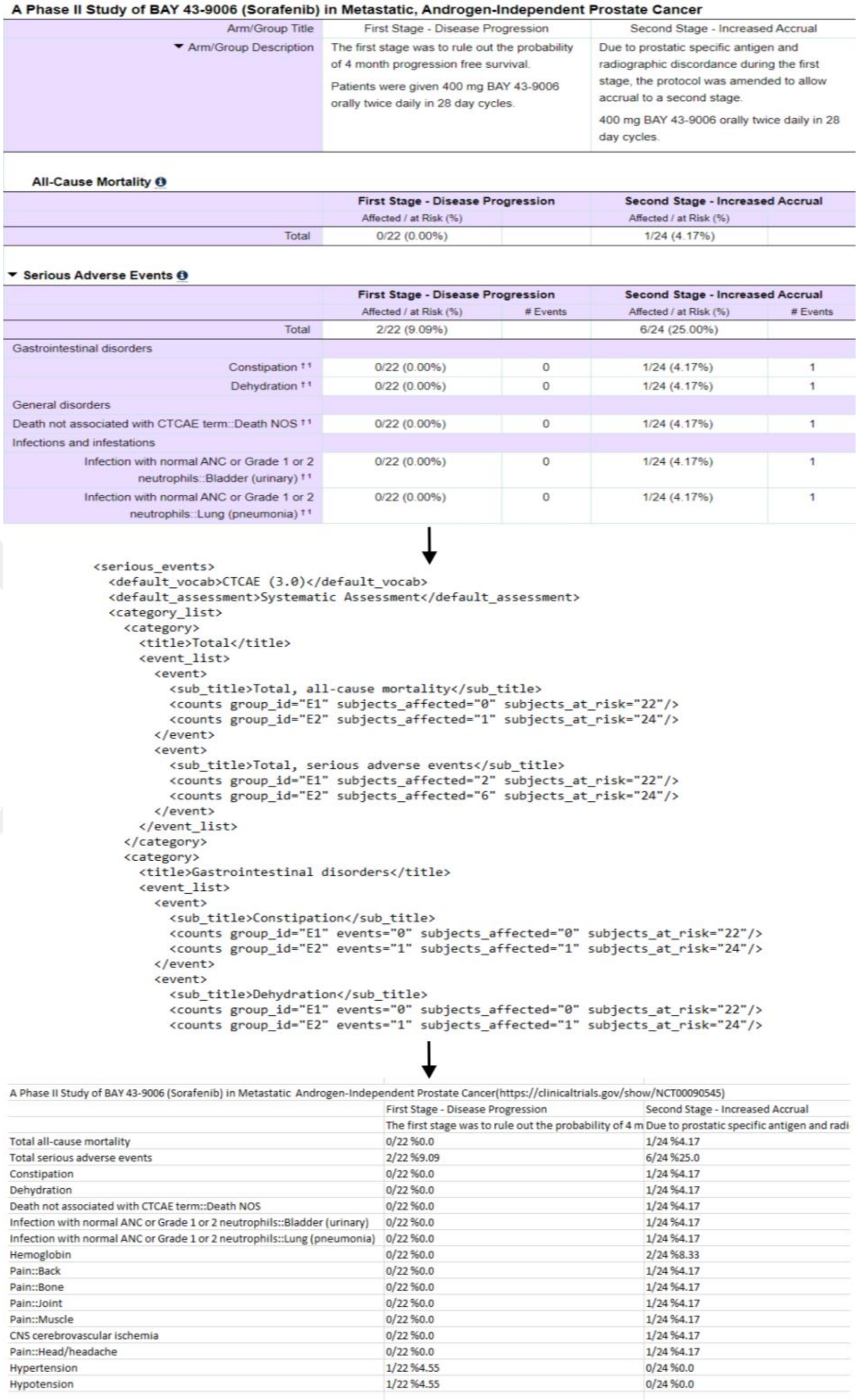
3.2.2 Data Extraction From XML Files

The desired information to be used in data analysis was extracted from downloaded XML files. The XML files in zip were extracted to a local folder and a Python script was written for the data extraction process. Study title, link of the study that also contains the Clinicaltrials.gov identifier, arm/group titles and arm/group descriptions of the study and for each serious adverse event from each study arm; serious adverse event term, number of affected participants, number of participants at risk and ratio of these two were extracted from the XML files.

ElementTree, which is part of the Python standard library was used in the script to read and parse the XML data. Since XML documents are in a tree-like structure, their elements were accessed by ElementTree. Using the methods of this library, a file was parsed into an element tree (ElementTree.parse()) and the root element of this tree was returned (tree.getroot()). That root was used to access the child and sub-child elements within a for loop.

The root element and all elements below it were iterated over until a specific element tag was found (using root.iter()). After finding the element tag, the element that had the tag was returned and the desired information was acquired either from its child or its sub-child with another root iteration (which would be an iteration over child elements this time).

Study URL was gathered from the “url” child of the “required_header” element and study name was gathered from the “brief_title” element. Title and description of groups were gathered from the “title” sub-child and “description” sub-child of the “group_list” child of the “reported_events” element respectively. Serious adverse event terms were gathered from the “sub_title” sub-child of the “event” child of the “serious_events” element. Likewise, numbers of affected participants and numbers of participants at risk were gathered from the “counts” sub-child of the same child and element; their ratios were also calculated. All the acquired information was separated by commas to form a table structure and written to a comma-separated values (CSV) file (Figure 3.4). Before that, the commas in the study title, group titles and descriptions, and serious adverse event terms were removed to avoid spurious columns in the final data. This process was repeated recursively for each XML file that was extracted from the three zip files containing clinical trial results of three different anticancer drugs. For this, the *glob* module was used to retrieve pathnames (the XML files in this case) recursively from the local folder where the XML files were extracted.



A Phase II Study of BAY 43-9006 (Sorafenib) in Metastatic, Androgen-Independent Prostate Cancer

Arm/Group Title	First Stage - Disease Progression	Second Stage - Increased Accrual
▼ Arm/Group Description	The first stage was to rule out the probability of 4 month progression free survival. Patients were given 400 mg BAY 43-9006 orally twice daily in 28 day cycles.	Due to prostatic specific antigen and radiographic discordance during the first stage, the protocol was amended to allow accrual to a second stage. 400 mg BAY 43-9006 orally twice daily in 28 day cycles.

All-Cause Mortality ℹ

	First Stage - Disease Progression		Second Stage - Increased Accrual	
	Affected / at Risk (%)		Affected / at Risk (%)	
Total	0/22 (0.00%)		1/24 (4.17%)	

▼ Serious Adverse Events ℹ

	First Stage - Disease Progression		Second Stage - Increased Accrual	
	Affected / at Risk (%)	# Events	Affected / at Risk (%)	# Events
Total	2/22 (9.09%)		6/24 (25.00%)	
Gastrointestinal disorders				
Constipation †1	0/22 (0.00%)	0	1/24 (4.17%)	1
Dehydration †1	0/22 (0.00%)	0	1/24 (4.17%)	1
General disorders				
Death not associated with CTCAE term::Death NOS †1	0/22 (0.00%)	0	1/24 (4.17%)	1
Infections and infestations				
Infection with normal ANC or Grade 1 or 2 neutrophils::Bladder (urinary) †1	0/22 (0.00%)	0	1/24 (4.17%)	1
Infection with normal ANC or Grade 1 or 2 neutrophils::Lung (pneumonia) †1	0/22 (0.00%)	0	1/24 (4.17%)	1

```
<serious_events>
  <default_vocab>CTCAE (3.0)</default_vocab>
  <default_assessment>Systematic Assessment</default_assessment>
  <category_list>
    <category>
      <title>Total</title>
      <event_list>
        <event>
          <sub_title>Total, all-cause mortality</sub_title>
          <counts group_id="E1" subjects_affected="0" subjects_at_risk="22"/>
          <counts group_id="E2" subjects_affected="1" subjects_at_risk="24"/>
        </event>
        <event>
          <sub_title>Total, serious adverse events</sub_title>
          <counts group_id="E1" subjects_affected="2" subjects_at_risk="22"/>
          <counts group_id="E2" subjects_affected="6" subjects_at_risk="24"/>
        </event>
      </event_list>
    </category>
    <category>
      <title>Gastrointestinal disorders</title>
      <event_list>
        <event>
          <sub_title>Constipation</sub_title>
          <counts group_id="E1" events="0" subjects_affected="0" subjects_at_risk="22"/>
          <counts group_id="E2" events="1" subjects_affected="1" subjects_at_risk="24"/>
        </event>
        <event>
          <sub_title>Dehydration</sub_title>
          <counts group_id="E1" events="0" subjects_affected="0" subjects_at_risk="22"/>
          <counts group_id="E2" events="1" subjects_affected="1" subjects_at_risk="24"/>
```

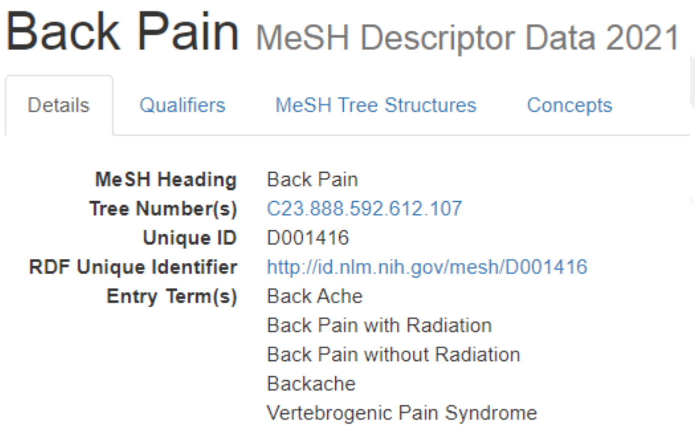
A Phase II Study of BAY 43-9006 (Sorafenib) in Metastatic Androgen-Independent Prostate Cancer(https://clinicaltrials.gov/show/NCT00090545)		
	First Stage - Disease Progression	Second Stage - Increased Accrual
	The first stage was to rule out the probability of 4 m	Due to prostatic specific antigen and radi
Total all-cause mortality	0/22 %0.0	1/24 %4.17
Total serious adverse events	2/22 %9.09	6/24 %25.0
Constipation	0/22 %0.0	1/24 %4.17
Dehydration	0/22 %0.0	1/24 %4.17
Death not associated with CTCAE term::Death NOS	0/22 %0.0	1/24 %4.17
Infection with normal ANC or Grade 1 or 2 neutrophils::Bladder (urinary)	0/22 %0.0	1/24 %4.17
Infection with normal ANC or Grade 1 or 2 neutrophils::Lung (pneumonia)	0/22 %0.0	1/24 %4.17
Hemoglobin	0/22 %0.0	2/24 %8.33
Pain::Back	0/22 %0.0	1/24 %4.17
Pain::Bone	0/22 %0.0	1/24 %4.17
Pain::Joint	0/22 %0.0	1/24 %4.17
Pain::Muscle	0/22 %0.0	1/24 %4.17
CNS cerebrovascular ischemia	0/22 %0.0	1/24 %4.17
Pain::Head/headache	0/22 %0.0	1/24 %4.17
Hypertension	1/22 %4.55	0/24 %0.0
Hypotension	1/22 %4.55	0/24 %0.0

Figure 3.4. Respectively; the website, the XML file and the created CSV file views of the serious adverse events data.

3.3 Creation of Term Dictionaries in Python

3.3.1 MeSH

The MeSH data (2020) [131] was downloaded from nlm.nih.gov in XML format. As in the extraction of clinical data, the necessary information in the MeSH XML file was extracted using ElementTree. Main heading terms (descriptor) and their entry terms (different terms that represent the same condition as main heading) and unique ID's of terms (Figure 3.5) were extracted from the XML file and stored in a Python dictionary with terms as keys and IDs as values (Figure 3.6). The intention in including entry terms in addition to main headings was to make it easier to match MeSH terms with terms used in clinical trials by increasing the variety of terms.

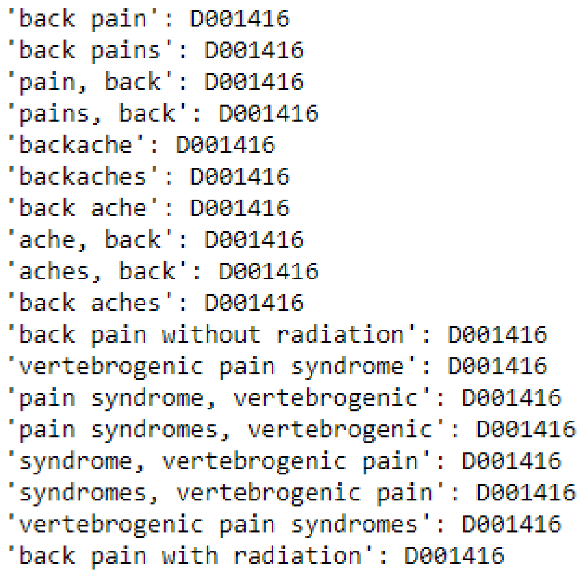


Back Pain MeSH Descriptor Data 2021

MeSH Heading: Back Pain
Tree Number(s): C23.888.592.612.107
Unique ID: D001416
RDF Unique Identifier: http://id.nlm.nih.gov/mesh/D001416
Entry Term(s): Back Ache
Back Pain with Radiation
Back Pain without Radiation
Backache
Vertebrogenic Pain Syndrome

Figure 3.5. Unique ID and entry terms of back pain MeSH heading [132].

The XML file was parsed into an element tree and the root element of this tree was returned. The root was iterated over and term ID was gathered from the "DescriptorUI" element, the term name was gathered from the "DescriptorName" element and entry terms of each term were gathered from the "TermList" child of the "ConceptList" element. Main heading and its entry terms are stored in a Python dictionary along with the unique ID. This process was repeated for each main heading in the XML file. After that, the created dictionary was saved to a file for later access by using the Python *pickle* module.

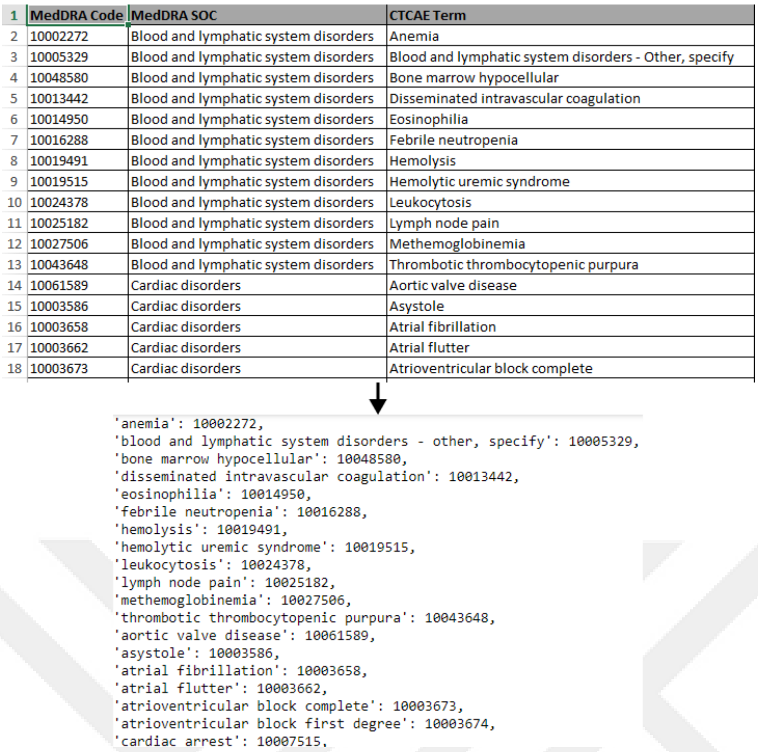


```
'back pain': D001416
'back pains': D001416
'pain, back': D001416
'pains, back': D001416
'backache': D001416
'backaches': D001416
'back ache': D001416
'ache, back': D001416
'aches, back': D001416
'back aches': D001416
'back pain without radiation': D001416
'vertebrogenic pain syndrome': D001416
'pain syndrome, vertebrogenic': D001416
'pain syndromes, vertebrogenic': D001416
'syndrome, vertebrogenic pain': D001416
'syndromes, vertebrogenic pain': D001416
'vertebrogenic pain syndromes': D001416
'back pain with radiation': D001416
```

Figure 3.6. A partial view of the dictionary. The unique ID and the entry terms of back pain main heading can be seen.

3.3.2 CTCAE

The CTCAE (v5.0) [133] terms and their unique MedDRA codes were downloaded from ctep.cancer.gov in XLSX format which is a XML-based format. The file was converted to a dataframe by using the read_excel method of *pandas* [134] which is a Python package. After that, the "CTCAE Term" and "MedDRA Code" columns of the dataframe were selected. The terms and codes in the selected columns were stored in a dictionary as keys and values respectively (Figure 3.7). The created dictionary was saved to a file by using pickle for later use.



1	MedDRA Code	MedDRA SOC	CTCAE Term
2	10002272	Blood and lymphatic system disorders	Anemia
3	10005329	Blood and lymphatic system disorders	Blood and lymphatic system disorders - Other, specify
4	10048580	Blood and lymphatic system disorders	Bone marrow hypocellular
5	10013442	Blood and lymphatic system disorders	Disseminated intravascular coagulation
6	10014950	Blood and lymphatic system disorders	Eosinophilia
7	10016288	Blood and lymphatic system disorders	Febrile neutropenia
8	10019491	Blood and lymphatic system disorders	Hemolysis
9	10019515	Blood and lymphatic system disorders	Hemolytic uremic syndrome
10	10024378	Blood and lymphatic system disorders	Leukocytosis
11	10025182	Blood and lymphatic system disorders	Lymph node pain
12	10027506	Blood and lymphatic system disorders	Methemoglobinemia
13	10043648	Blood and lymphatic system disorders	Thrombotic thrombocytopenic purpura
14	10061589	Cardiac disorders	Aortic valve disease
15	10003586	Cardiac disorders	Asystole
16	10003658	Cardiac disorders	Atrial fibrillation
17	10003662	Cardiac disorders	Atrial flutter
18	10003673	Cardiac disorders	Atrioventricular block complete

```
'anemia': 10002272,
'blood and lymphatic system disorders - other, specify': 10005329,
'bone marrow hypocellular': 10048580,
'disseminated intravascular coagulation': 10013442,
'eosinophilia': 10014950,
'febrile neutropenia': 10016288,
'hemolysis': 10019491,
'hemolytic uremic syndrome': 10019515,
'leukocytosis': 10024378,
'lymph node pain': 10025182,
'methemoglobinemia': 10027506,
'thrombotic thrombocytopenic purpura': 10043648,
'aortic valve disease': 10061589,
'asystole': 10003586,
'atrial fibrillation': 10003658,
'atrial flutter': 10003662,
'atrioventricular block complete': 10003673,
'atrioventricular block first degree': 10003674,
'cardiac arrest': 10007515,
```

Figure 3.7. Conversion of CTCAE data from XSLX file to a dictionary.

3.4 Assignment of IDs to Serious Adverse Event Terms

An algorithm (Figure 3.8) was designed to assign the stored unique IDs in the dictionaries to the clinical serious adverse event terms in the CSV file. First the non-alphabetic characters were removed from the clinical trial terms. The clinical trial terms were searched for in the dictionaries. The unique ID, which is the value of the key term in the dictionary, were assigned to clinical trial terms right away if they were exactly the same as in the dictionaries. If the algorithm could not find exactly the same term in the dictionaries then it would search for the most similar one in the dictionaries to match with the clinical trial term. The similarity calculation was based on scoring matched words and length of the two terms. For that, stop words[1]

[1] Stop words are words that occur very frequently in language, and that do not provide useful information (e.g. "the", "and", "of", etc.).

were removed from the clinical trial terms in order to reduce bias and chance of matching different conditions other than clinical trial terms indicated. Since the similarity calculation was based on the word matching it would fail if there were typos in the clinical trial term. So, for the terms that have typos in them, a similarity function that calculates the similarity based on letters was used. After that, all the found terms and their unique IDs were written to another CSV file to be used in data analysis tools.

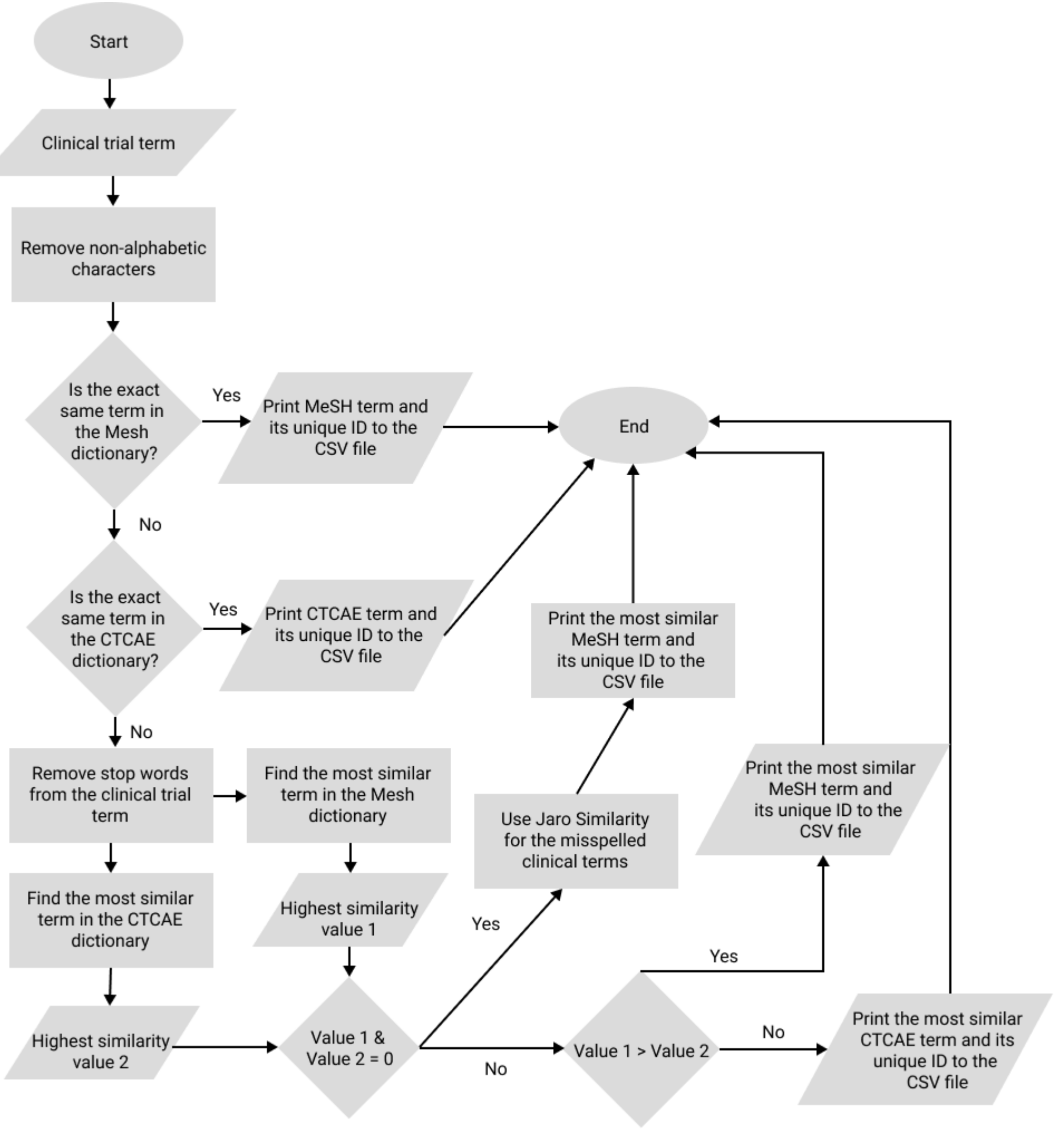

Start
Clinical trial term
Remove non-alphabetic characters
Is the exact same term in the Mesh dictionary?
Yes
Print MeSH term and its unique ID to the CSV file
End
No
Is the exact same term in the CTCAE dictionary?
Yes
Print CTCAE term and its unique ID to the CSV file
Print the most similar MeSH term and its unique ID to the CSV file
No
Remove stop words from the clinical trial term
Find the most similar term in the Mesh dictionary
Print the most similar MeSH term and its unique ID to the CSV file
Use Jaro Similarity for the misspelled clinical terms
Find the most similar term in the CTCAE dictionary
Highest similarity value 1
Yes
Yes
Highest similarity value 2
Value 1 & Value 2 = 0
No
Value 1 > Value 2
No
Print the most similar CTCAE term and its unique ID to the CSV file

Figure 3.8. The flowchart of the algorithm that finds terms in the dictionaries and assigns unique IDs to the reported serious adverse events.

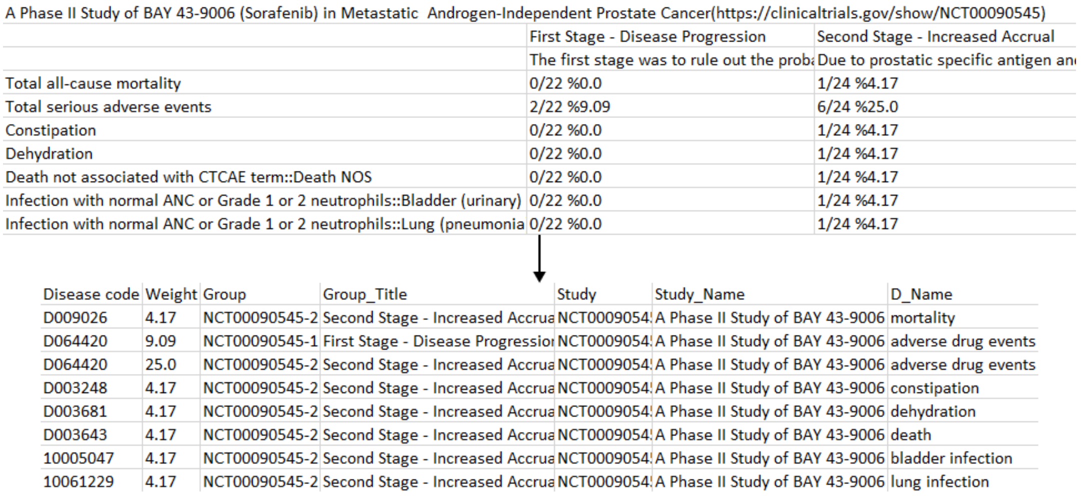
The assignment process was carried out with a defined function that takes one input which is a clinical trial term. The function was used in a while loop to take every clinical term of each clinical study from the CSV file. The CSV file was read line by line and the information of clinical trial study name, Clinicaltrail.gov identifier, group no (Clinicaltrail.gov identifier-1, 2, ..., number of groups) and group titles were acquired. When the clinical trial terms were being read each of them was fed into the defined function as inputs. The function utilizes the *re* (regular expression) module that was used to remove (re.sub()) non-alphabetic characters like brackets, numbers and dashes. As an exception, if the "GI" abbreviation existed in the clinical trial term it was replaced with "gastrointestinal". The exact same term was searched for by checking the clinical trial terms whether they were in the dictionaries or not. MeSH dictionary was used first for this process, since it is the backbone of this assignment process. If the exact match was found, the term and unique ID were written to a CSV file along with its weight (ratio of number of affected participants to number of participants at risk) and the other acquired information (study title, Clinicaltrial.gov identifier, group title etc.) in a single line. If it was not found in the MeSH dictionary then the CTCAE dictionary would be tried. If the exact same clinical trial term was not found in either, then the algorithm would search for the most similar one in the dictionaries. To find the most similar one, clinical trial terms were needed to be cleared from stop words. For that, a function that includes the stopwords corpus was defined inside the parent function. The stopwords in NLTK [135] (Natural Language Toolkit) were used to remove English stop words. For that, each word of each term was checked for stop words and removed if necessary. In addition to stop words, abbreviations such as NOS (not otherwise specified) and ANC (absolute neutrophil count) were also removed. After removing unnecessary words, the clinical trial term was compared with every term in the MeSH dictionary. The score that had been initiated as zero was increased by one for each identical word in two compared terms. For every clinical trial term and MeSH term pair, similarities were calculated as follows:

$$\frac{Score\ (number\ of\ common\ terms)}{\begin{array}{c} number\ of\ words\ in\ term \\ in\ the\ dictionary \end{array}}\ x\ \frac{Score\ (number\ of\ common\ terms)}{\begin{array}{c} number\ of\ words\ in \\ clinical\ trial\ term \end{array}}$$

After finding the MeSH term that had the maximum value of similarity, its title, unique ID and similarity value were stored. The same process was repeated for the CTCAE dictionary. After comparing clinical trial term with every term in the CTCAE dictionary, this time similarities were calculated for each clinical trial term and CTCAE term pair. To find the most similar term, the maximum similarity value of the MeSH term and the maximum similarity value of the CTCAE term were compared with each other. The term with the highest value was chosen and its term title, unique ID/code in the dictionary (Table 3.2) and weight (i.e the ratio of number of affected participants to number of participants at risk) were written to the CSV file. The algorithm could not find the exact same terms or a similarity value greater than zero for misspelled clinical trial terms. For that, the jaro_distance() function from the *Jellyfish* library was used. The function utilizes the Jaro Similarity which gives a value between zero and one for two strings, where zero represents complete dissimilarity and one represents identical strings. The misspelled clinical trial term and every term in the MeSH dictionary were compared based on the similarity (Table 3.1). The term with the value that was closest to one was selected and its term title, unique ID in the dictionary and weight were written to the CSV file (Figure 3.9). This assignment process was repeated for every clinical trial term of each study and terms with no affected participants were excluded from this process.

Table 3.1 Some of the misspelled clinical term corrections.

Misspelled Clinical Term	**The Term Found in the MeSH Dictionary by using Jaro Similarity**
Diarrhoea	Diarrhea
Hyperglycaemia	Hyperglycemia
Hypotnesion	Hypotension
Intussception	Intussusception



A Phase II Study of BAY 43-9006 (Sorafenib) in Metastatic Androgen-Independent Prostate Cancer(https://clinicaltrials.gov/show/NCT00090545)

	First Stage - Disease Progression	Second Stage - Increased Accrual
	The first stage was to rule out the prob	Due to prostatic specific antigen an
Total all-cause mortality	0/22 %0.0	1/24 %4.17
Total serious adverse events	2/22 %9.09	6/24 %25.0
Constipation	0/22 %0.0	1/24 %4.17
Dehydration	0/22 %0.0	1/24 %4.17
Death not associated with CTCAE term::Death NOS	0/22 %0.0	1/24 %4.17
Infection with normal ANC or Grade 1 or 2 neutrophils::Bladder (urinary)	0/22 %0.0	1/24 %4.17
Infection with normal ANC or Grade 1 or 2 neutrophils::Lung (pneumonia	0/22 %0.0	1/24 %4.17

↓

Disease code	Weight	Group	Group_Title	Study	Study_Name	D_Name
D009026	4.17	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	mortality
D064420	9.09	NCT00090545-1	First Stage - Disease Progressio	NCT0009054	A Phase II Study of BAY 43-9006	adverse drug events
D064420	25.0	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	adverse drug events
D003248	4.17	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	constipation
D003681	4.17	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	dehydration
D003643	4.17	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	death
10005047	4.17	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	bladder infection
10061229	4.17	NCT00090545-2	Second Stage - Increased Accrua	NCT0009054	A Phase II Study of BAY 43-9006	lung infection

Figure 3.9. Creation of the CSV file with the found terms in the dictionaries.

Table 3.2 Examples of unique ID assignments.

Clinical Trial Term (Input)	Term after Removing stop words and non-characters	Output	
		Term	ID
Infection with normal ANC or grade 1 or 2 neutrophils::bladder (urinary)	infection normal grade neutrophils bladder urinary	bladder infection	10005047 (CTCAE)
Pain::Head/headache	pain head headache	headache	D006261 (MeSH)
Headache	(Found directly)	headache	D006261 (MeSH)
Non-cardiac chest pain	(Found directly)	non-cardiac chest pain	10062501 (CTCAE)

3.5 Transforming Data into a Sparse Matrix

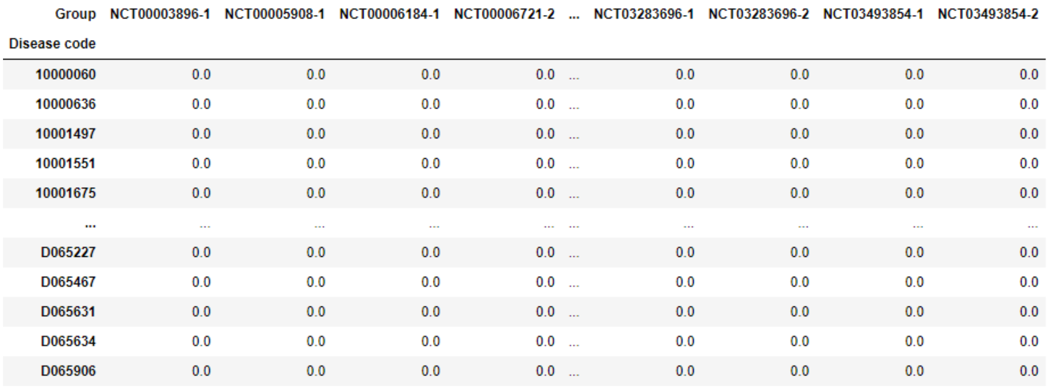
After the assignment process, the obtained dataset was transformed into a format that was appropriate for both the intended purpose and the methods to be used. For this, “pivot_table()” method of pandas was used in the data reshaping to transform the data structure into a suitable format for further analysis. Three parameters were given to derive a table out of the existing dataset. Since the aim of this study is to discover links between drugs, serious adverse events and clinical studies, the dataset was reshaped to suit this intention. The weight of the serious adverse events were specified as values, the serious adverse event ID's as indices and group numbers as columns. So, groups of the clinical trials were used as attributes to describe the serious adverse events. Because not all the serious adverse events in the dataset take place in every clinical trial, there are lots of cells without weights. Since they were filled with zeros, a sparse matrix with 1725 rows and 700 columns was created out of the existing dataset (Figure 3.10). The pivot operation is explained in Table 3.3 and Table 3.4.

Table 3.3 The structure of the data after the assignment process (the colors demonstrate the pivot operation in Table 3.4).

Disease Code	**Weight**	**Group No**
D009026	4.17	NCT00090545-2
D064420	9.09	NCT00090545-1
D064420	25.00	NCT00090545-2
D003248	4.17	NCT00090545-2
D003681	4.17	NCT00090545-2

Table 3.4 The structure of the data after the pivot operation.

	NCT00090545-1	NCT00090545-2
D009026	0.0	4.17
D064420	9.09	25.00
D003248	0.0	4.17
D003681	0.0	4.17



Group	NCT00003896-1	NCT00005908-1	NCT00006184-1	NCT00006721-2	...	NCT03283696-1	NCT03283696-2	NCT03493854-1	NCT03493854-2
Disease code									
10000060	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
10000636	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
10001497	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
10001551	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
10001675	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...
D065227	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
D065467	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
D065631	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
D065634	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
D065906	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0

Figure 3.10. The structure of the sparse matrix.

3.6 Data Visualization Using t-SNE and DBSCAN

The created sparse matrix data was visualised by using the t-SNE and DBSCAN algorithms from scikit-learn library [136]. Since t-SNE proved to be very effective in visualizing and clustering the data after many trials and DBSCAN was good at distinguishing these clusters and labeling them, the t-SNE/DBSCAN combination was used for this process. Before running t-SNE on the data, it was standardized by using the StandardScaler method from the sklearn.preprocessing module. With the standardization, formation of distant data points by t-SNE was prevented, so DBSCAN could label every data point on the two dimensional map. t-SNE was run on the standardized data and the two dimensional map was plotted. For that, the output of the t-SNE that contains the coordinates of the data points in the two dimensional map was plotted as a scatter plot by using the Seaborn's scatterplot function. Seaborn is a python based visualization library. Since the rows of the input sparse matrix are the serious adverse events, they were clustered based on the similarity between them according to the column attributes which are the groups in this case.

The output of t-SNE that contains the data point coordinates was used as the input of DBSCAN to label the data points based on the data point densities. A dataframe that consists of the names of serious adverse events, their unique IDs and DBSCAN labels was created to use after the data co-clustering and p-value calculations.

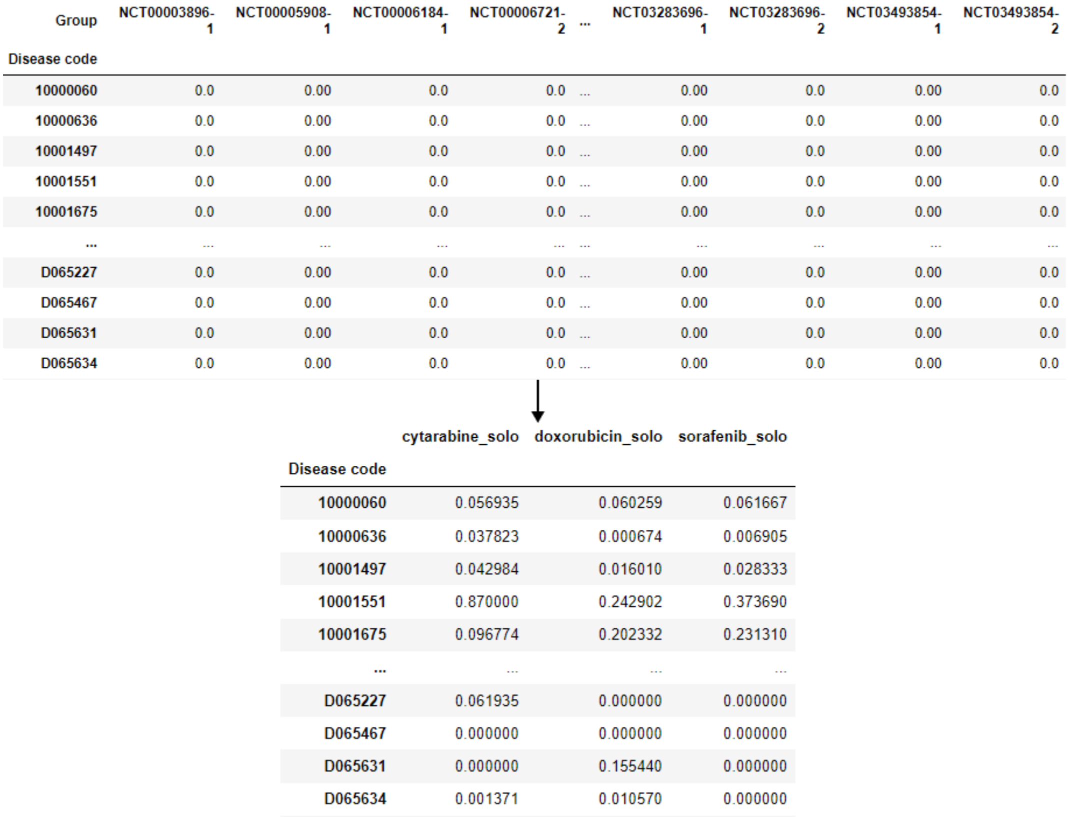
3.7 Filtering For Groups That Contain Any of the Three Drugs

Although the drugs Sorafenib, Cytarabine and Doxorubicin were indicated while doing the search for the clinical trials that involve the use of these drugs, not all the groups of clinical trials involve them necessarily. Since some clinical trials compare the effects of different drugs, the combination may vary with different drugs in groups. So, some of the groups of the clinical trials that were collected did not involve the indicated drugs and they were filtered out.

For filtering the groups that contain these 3 drugs, titles of studies and descriptions of groups were used. Because the group descriptions were not in the existing dataframe, another CSV file was created that contains the group numbers and group descriptions. After converting it into a dataframe, it was merged with the existing dataframe by using the pandas' merge method. After that, the names of the three drugs were searched both in the group descriptions and titles with the "*str.contains()*" method of pandas, and the ones that did not have any of them were discarded. The group numbers of the resulting data frame were used to select columns of the sparse matrix. By that, the groups that do not involve the indicated drugs were filtered out so that they would not affect subsequent results. Besides that, the groups that involve any two of the three drugs simultaneously were also discarded because it would not be clear which drugs affected the resulting serious adverse event weights. After the filtering process there were 401 groups left out of 700.

3.8 Column Aggregation of Weights by Three Drugs

After the sparse matrix transformation and filtering, the mean aggregation of weights was calculated based on the three drugs. For that, the group numbers from the dataframe created in the filtering process were used since they were already arranged in an order according to the drugs. These arranged group numbers were used in order to arrange the sparse matrix columns. For that, the *loc* method, which is a label based indexing method of pandas was used to select groups by their positions with the group names. After arranging the columns of the sparse matrix according to which drugs they belong to, their means were calculated separately for the three drugs. For that, the first and the last group numbers of three drugs were used to slice the sparse matrix with the loc method. Finally, the column-wise mean was calculated for each of the three drugs (Figure 3.11).



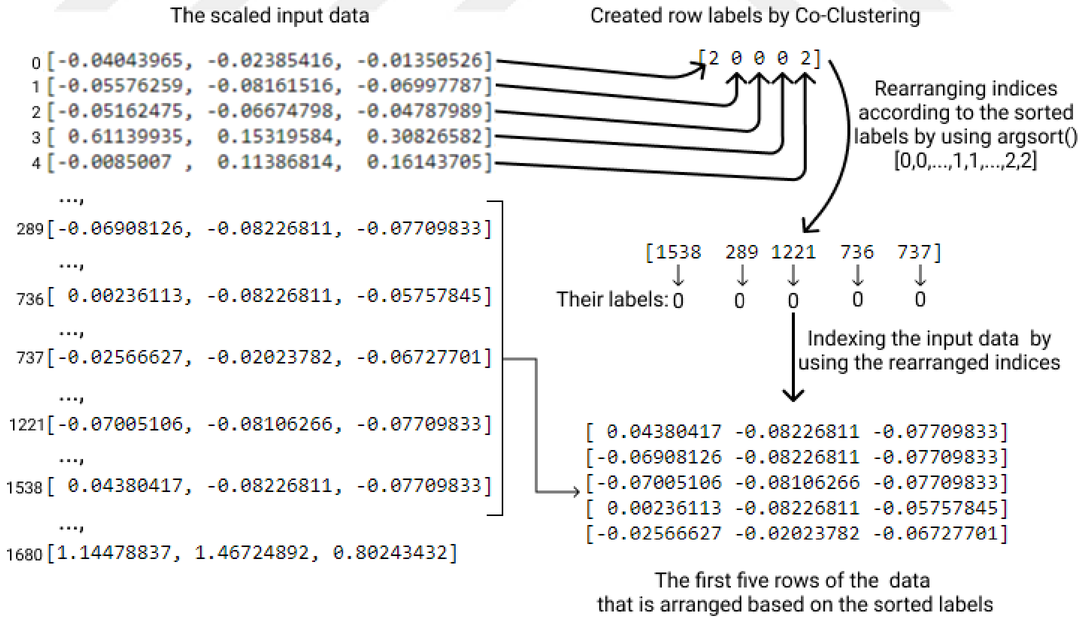
Group	NCT00003896-1	NCT00005908-1	NCT00006184-1	NCT00006721-2	...	NCT03283696-1	NCT03283696-2	NCT03493854-1	NCT03493854-2
Disease code									
10000060	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
10000636	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
10001497	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
10001551	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
10001675	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
...	...	...	...	...	...	...	...	...	...
D065227	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
D065467	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
D065631	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0
D065634	0.0	0.00	0.0	0.0	...	0.00	0.0	0.00	0.0

	cytarabine_solo	doxorubicin_solo	sorafenib_solo
Disease code			
10000060	0.056935	0.060259	0.061667
10000636	0.037823	0.000674	0.006905
10001497	0.042984	0.016010	0.028333
10001551	0.870000	0.242902	0.373690
10001675	0.096774	0.202332	0.231310
...	...	...	...
D065227	0.061935	0.000000	0.000000
D065467	0.000000	0.000000	0.000000
D065631	0.000000	0.155440	0.000000
D065634	0.001371	0.010570	0.000000

Figure 3.11. The transformation of the sparse matrix into the aggregated data.

3.9 Data Co-Clustering

The rows and columns of the aggregated data were clustered simultaneously. For this, the spectral co-clustering algorithm from the sklearn machine learning library was used. Before running the algorithm on the data, it was standardized by using the StandardScaler method to handle the numerical errors resulting from the data still having many zeros as mean weights. After that, the standardized data was clustered. The co-clustering algorithm creates cluster labels for columns and rows. Because there were three columns in the data, three different cluster labels were created for rows, which denote serious adverse events. The cluster memberships (labels) of the rows were found by using the row_labels_ attribute of the model. Since we needed to arranged the rows based on their cluster labels, the memberships that were returned by the row_labels_ attribute were rearranged by using the argsort() method of numpy. With argsort(), the arranged indices of rows are returned according to their labels. By indexing the data by using these rearranged indices the rows were put in order based on their labels (Figure 3.12).


The scaled input data
Created row labels by Co-Clustering
Rearranging indices according to the sorted labels by using argsort() [0,0,...,1,1,...,2,2]
Their labels: 0 0 0 0 0
Indexing the input data by using the rearranged indices
The first five rows of the data that is arranged based on the sorted labels

```
0   [-0.04043965, -0.02385416, -0.01350526]
1   [-0.05576259, -0.08161516, -0.06997787]
2   [-0.05162475, -0.06674798, -0.04787989]
3   [ 0.61139935,  0.15319584,  0.30826582]
4   [-0.0085007 ,  0.11386814,  0.16143705]
     ...,
289 [-0.06908126, -0.08226811, -0.07709833]
     ...,
736 [ 0.00236113, -0.08226811, -0.05757845]
     ...,
737 [-0.02566627, -0.02023782, -0.06727701]
     ...,
1221[-0.07005106, -0.08106266, -0.07709833]
     ...,
1538[ 0.04380417, -0.08226811, -0.07709833]
     ...,
1680[1.14478837, 1.46724892, 0.80243432]
```

```
[2 0 0 0 2]
```

```
[1538  289 1221  736  737]
```

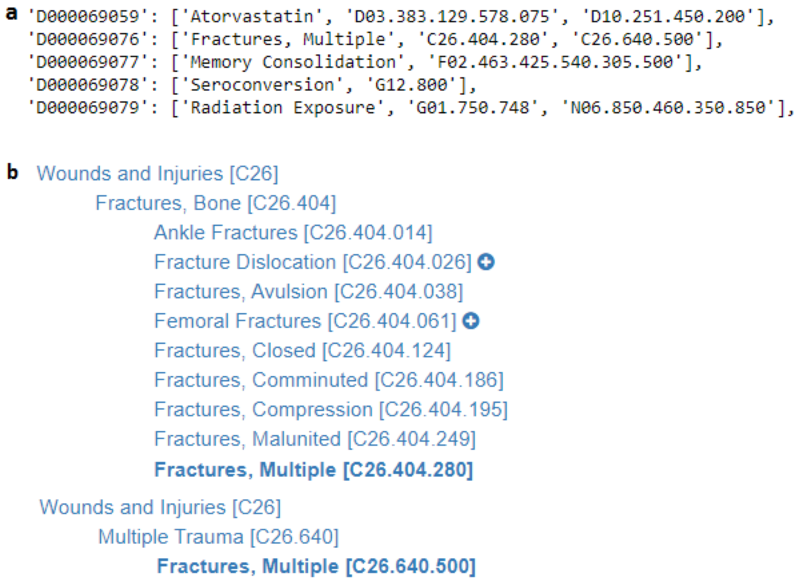
```
[ 0.04380417 -0.08226811 -0.07709833]
[-0.06908126 -0.08226811 -0.07709833]
[-0.07005106 -0.08106266 -0.07709833]
[ 0.00236113 -0.08226811 -0.05757845]
[-0.02566627 -0.02023782 -0.06727701]
```

Figure 3.12. Labeling and rearranging of the data matrix.

After arranging the data in an order which was indicated by the co-cluster labels, the matrix created was displayed. For this, the matshow function of matplotlib which is a data visualization library was used. Since the values were between zero and one, they were raised to the power of 0.1 to make them visible on the plot. By that, serious adverse events that were clustered in three different anticancer drugs were displayed. These serious adverse events were stored based on their drug clusters.

3.10 Applying Hierarchical Structure of Dictionaries to Serious Adverse Events

After adverse serious events were separated according to the drugs, and they were clustered by co-clustering, two dictionaries (one for the MeSH, one for the CTCAE terms) were created to make use of the hierarchical structure of the MeSH (Figure 3.13). For that, the downloaded MeSH data in XML format was used again. This time the unique ID’s of terms, main heading terms and their tree numbers in the hierarchical structure were extracted from the XML file and stored in a Python dictionary with IDs as keys and lists that consist terms and tree numbers as values.



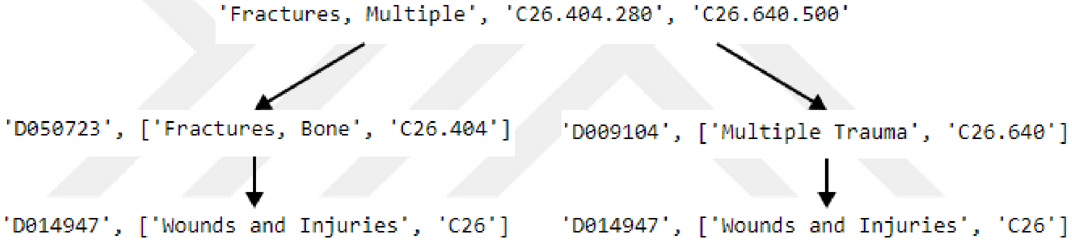
```
'D000069059': ['Atorvastatin', 'D03.383.129.578.075', 'D10.251.450.200'],
'D000069076': ['Fractures, Multiple', 'C26.404.280', 'C26.640.500'],
'D000069077': ['Memory Consolidation', 'F02.463.425.540.305.500'],
'D000069078': ['Seroconversion', 'G12.800'],
'D000069079': ['Radiation Exposure', 'G01.750.748', 'N06.850.460.350.850'],
```

```
Wounds and Injuries [C26]
    Fractures, Bone [C26.404]
        Ankle Fractures [C26.404.014]
        Fracture Dislocation [C26.404.026] ⊕
        Fractures, Avulsion [C26.404.038]
        Femoral Fractures [C26.404.061] ⊕
        Fractures, Closed [C26.404.124]
        Fractures, Comminuted [C26.404.186]
        Fractures, Compression [C26.404.195]
        Fractures, Malunited [C26.404.249]
        Fractures, Multiple [C26.404.280]
Wounds and Injuries [C26]
    Multiple Trauma [C26.640]
        Fractures, Multiple [C26.640.500]
```

Figure 3.13. A partial view of the dictionary (a), and the hierarchical structure of the “Fractures, Multiple” serious adverse event (b).

As Figure 3.13 shows, the tree numbers are separated by dots and there can be more than one hierarchical groups for one serious adverse event. Figure 3.14 shows the created root and parent events of the “Fractures, Multiple” serious adverse event.

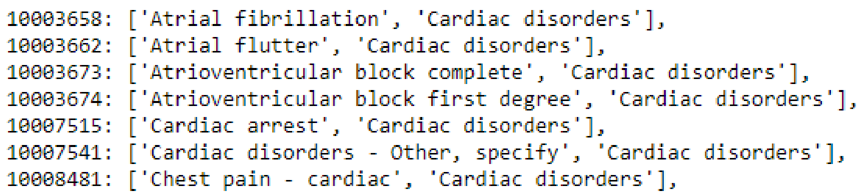
By including the tree numbers of the serious adverse events in the dictionary as values, their parents and roots in the hierarchical structure were made accessible. After the co-clustering, serious adverse event terms were used to get their roots (highest level in the hierarchy) and parents (one level lower from the root). The IDs of the serious adverse events were searched for in the dictionary and their element tree numbers were found, after that, the first three numbers were used to find the root event term of the serious adverse events and the first six numbers were used to find the parent terms. If the serious adverse event terms belong to more than one hierarchical tree, all its roots and parents were found regardless.



```
'Fractures, Multiple', 'C26.404.280', 'C26.640.500'
'D050723', ['Fractures, Bone', 'C26.404']
'D014947', ['Wounds and Injuries', 'C26']
'D009104', ['Multiple Trauma', 'C26.640']
'D014947', ['Wounds and Injuries', 'C26']
```

Figure 3.14. Different hierarchical levels of the “Fractures, Multiple” serious adverse event in the created Python dictionary.

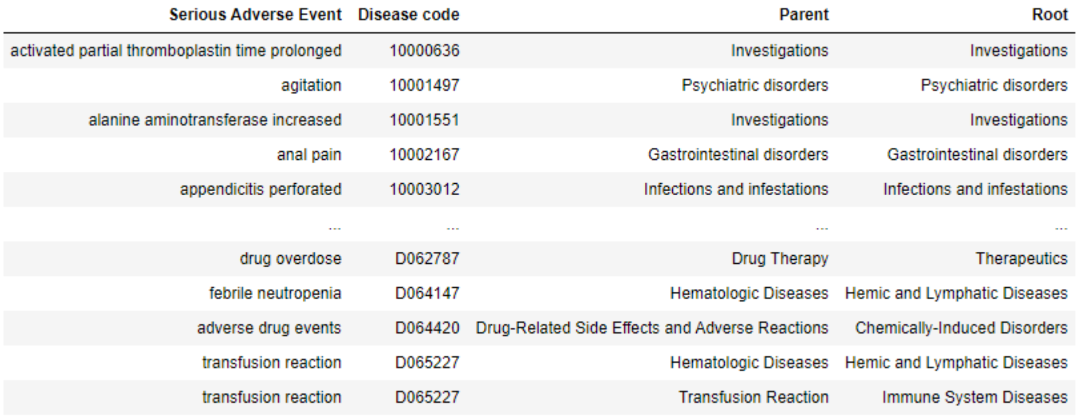
The other dictionary created was for the CTCAE to find the parents and roots of serious adverse events that were found in the CTCAE. For that, the downloaded CTCAE data in XLSX format was used. Since the CTCAE does not have a hierarchical structure like MeSH has, just the MedDRA SOC (system organ classes) level was used to drive parents and roots (Figure 3.15). SOCs are the groupings that could be based on the purpose, site of the condition or causality. The unique ID’s of terms, CTCAE terms and MedDRA SOC were extracted from the XLSX file and stored in a Python dictionary with IDs as keys and lists that consist of terms and MedDRA SOCs as values.



```
10003658: ['Atrial fibrillation', 'Cardiac disorders'],
10003662: ['Atrial flutter', 'Cardiac disorders'],
10003673: ['Atrioventricular block complete', 'Cardiac disorders'],
10003674: ['Atrioventricular block first degree', 'Cardiac disorders'],
10007515: ['Cardiac arrest', 'Cardiac disorders'],
10007541: ['Cardiac disorders - Other, specify', 'Cardiac disorders'],
10008481: ['Chest pain - cardiac', 'Cardiac disorders'],
```

Figure 3.15. Partial view of the dictionary. IDs, adverse evet terms and MedDRA SOCs can be seen respectively.

The created dictionaries were used in the finding parent and root event terms of the clustered serious adverse events (Figure 3.16) in the three distinct dataframes.

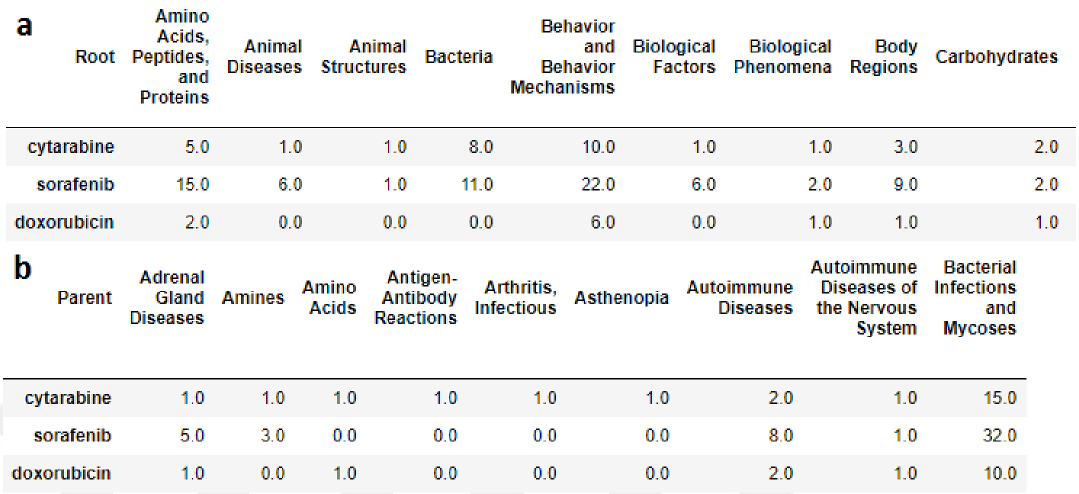


Serious Adverse Event	Disease code	Parent	Root
activated partial thromboplastin time prolonged	10000636	Investigations	Investigations
agitation	10001497	Psychiatric disorders	Psychiatric disorders
alanine aminotransferase increased	10001551	Investigations	Investigations
anal pain	10002167	Gastrointestinal disorders	Gastrointestinal disorders
appendicitis perforated	10003012	Infections and infestations	Infections and infestations
...	...	...	...
drug overdose	D062787	Drug Therapy	Therapeutics
febrile neutropenia	D064147	Hematologic Diseases	Hemic and Lymphatic Diseases
adverse drug events	D064420	Drug-Related Side Effects and Adverse Reactions	Chemically-Induced Disorders
transfusion reaction	D065227	Hematologic Diseases	Hemic and Lymphatic Diseases
transfusion reaction	D065227	Transfusion Reaction	Immune System Diseases

Figure 3.16. Partial view of dataframe that contains serious adverse events with their root and parent events.

3.11 Calculating P-Values

The p-values of each root and parent event were calculated to find the drug specific events. The number of root event in the dataframes that were created after applying the hierarchical structure of the serious adverse events were found for each dataframe of the three drugs. For this, the dataframes were grouped by the root event terms and they were counted by using the groupby() and the size() methods of pandas. After that, the three dataframes were merged together and one dataframe that contains the numbers of root events for each drug was created (Figure 3.17-a). The same things were done for the parent events and another dataframe that contains the numbers of parent events for each drug was created (Figure 3.17-b).



a

Root	Amino Acids, Peptides, and Proteins	Animal Diseases	Animal Structures	Bacteria	Behavior and Behavior Mechanisms	Biological Factors	Biological Phenomena	Body Regions	Carbohydrates
cytarabine	5.0	1.0	1.0	8.0	10.0	1.0	1.0	3.0	2.0
sorafenib	15.0	6.0	1.0	11.0	22.0	6.0	2.0	9.0	2.0
doxorubicin	2.0	0.0	0.0	0.0	6.0	0.0	1.0	1.0	1.0

b

Parent	Adrenal Gland Diseases	Amines	Amino Acids	Antigen-Antibody Reactions	Arthritis, Infectious	Asthenopia	Autoimmune Diseases	Autoimmune Diseases of the Nervous System	Bacterial Infections and Mycoses
cytarabine	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	15.0
sorafenib	5.0	3.0	0.0	0.0	0.0	0.0	8.0	1.0	32.0
doxorubicin	1.0	0.0	1.0	0.0	0.0	0.0	2.0	1.0	10.0

Figure 3.17. Partial view of the merged dataframes that contain the number of root (a) and parent (b) events in the drug clusters.

Since we found the root and parent events from two different source (the CTCAE and the MeSH) there are different event terms that actually represent the same root or parent event. This inconsistency would affect the p-value calculations and potentially lead to inaccurate results. For this reason, the CTCAE root events (MedDRA SOC) were merged with the appropriate MeSH root and parent terms (Table 3.5) and the numbers of serious adverse events they contain were summed up.

Table 3.5 Merged CTCAE and MeSH event terms.

CTCAE Root Events	**MeSH Root Events**	**MeSH Parent Events**
Blood and lymphatic system disorders	Hemic and lymphatic diseases	Lymphatic diseases
Cardiac disorders	Cardiovascular diseases	Heart diseases
Ear and labyrinth disorders	Otorhinolaryngologic Diseases	Ear Diseases

Table 3.5 Merged CTCAE and MeSH event terms (continued).

CTCAE Root Events	**MeSH Root Events**	**MeSH Parent Events**
Endocrine disorders	Endocrine System Diseases	Endocrine Gland Neoplasms
Eye disorders	Eye Diseases	Eye Abnormalities
Gastrointestinal disorders	Digestive System Diseases	Gastrointestinal Diseases
General disorders and administration site conditions	Pathological Conditions, Signs and Symptoms	Signs and Symptoms
Hepatobiliary disorders	Digestive System Diseases	Biliary Tract Diseases
Immune system disorders	Immune System Diseases	Immune System
Infections and infestations	Infections	-
Injury, poisoning and procedural complications	Wounds and Injuries	-
Investigations	Investigative Techniques	Clinical Laboratory Techniques
Metabolism and nutrition disorders	Nutritional and Metabolic Diseases	Metabolic Diseases
Musculoskeletal and connective tissue disorders	Musculoskeletal Diseases	-
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	Neoplasms	Neoplasms by Histologic Type
Nervous system disorders	Nervous System Diseases	Neurologic Manifestations
Psychiatric disorders	Behavior and Behavior Mechanisms	Behavior
Renal and urinary disorders	Female Urogenital Diseases and Pregnancy Complications - Male Urogenital Diseases	Female Urogenital Diseases - Urologic Diseases
Respiratory, thoracic and mediastinal disorders	Respiratory Tract Diseases	Respiration Disorders
Skin and subcutaneous tissue disorders	Skin and Connective Tissue Diseases	Skin Diseases
Vascular disorders	Cardiovascular Diseases	Vascular Diseases

While doing the CTCAE root and MeSH parent merging, the found serious adverse events under the CTCAE roots were considered. For example, since the most of the serious adverse events under the “Blood and lymphatic system disorders” CTCAE root event were related to lymphatic diseases so the “Lymphatic diseases” MeSH parent event was chosen instead of “Hematologic diseases” which is another sub-element of “Hemic and lymphatic diseases” root events. Another thing that was considered was that some of the CTCAE root events could not be merged with the MeSH parent events since they are too specific for a generalized root term. For example, the “Infections and infestations” CTCAE root event was not merged with any of the “Infections” MeSH root sub-elements (catheter-related infections, pelvic infections, soft tissue infections).

After root events are merged the p-values of each root event were calculated. For this, the χ^2 test (chi-square) from the scipy.stats module was used. To calculate the p-value, the incidence rates of the root events were found first. For each root event, the total number of a specific root event were found by the summation of the numbers for each drug and divided by the total serious adverse events which was 1725.

$$\frac{\mathit{Total\ number\ of\ a\ specific\ root\ term}}{\mathit{Total\ number\ of\ serious\ adverse\ events}} = \alpha\ (\mathit{the\ incidence\ rate})$$

After finding the incidence rate of that root term, its expected numbers for each drug were calculated. For this the numbers of the serious adverse events in the drug clusters that were created by the co-clustering method were used. The calculation is as follows:

$$\begin{array}{c}\mathit{Total\ number\ of\ serious\ adverse}\\ \mathit{event\ terms\ in}\\ \mathit{the\ Cytarabine\ cluster}\end{array}\ x\ \alpha = \begin{array}{c}\mathit{Expected\ number\ of\ root}\\ \mathit{term\ in\ the\ Cytarabine\ cluster}\end{array}$$

$$\begin{array}{c}\mathit{Total\ number\ of\ serious\ adverse}\\ \mathit{event\ terms\ in}\\ \mathit{the\ Doxorubicin\ cluster}\end{array}\ x\ \alpha = \begin{array}{c}\mathit{Expected\ number\ of\ root}\\ \mathit{term\ in\ the\ Doxorubicin\ cluster}\end{array}$$

$$\begin{matrix}\textit{Total number of serious adverse}\\ \textit{even terms in}\\ \textit{the Sorafenib cluster}\end{matrix}\; x\ \alpha = \begin{matrix}\textit{Expected number of root}\\ \textit{term in the Sorafenib cluster}\end{matrix}$$

The observed and expected numbers of each root term were used in the χ^2 function to calculate the p-values. The process explained above was also repeated for p-values of parent events which are lower by one level from the root events in the hierarchical structure.

3.12 Finding DBSCAN Labels of Co-Clustered Serious Adverse Events that Belong to the Significant Root or Parent Events

Since we found the DBSCAN labels of each serious adverse events, which drugs they clustered in and the significant root and parent events of serious adverse events; these findings could be used simultaneously. The serious adverse events of the significant root and parent events were searched for in the three different dataframes that were created as a result of the co-clustering method. Since the created dataframes contains the root and parent event terms of each serious adverse event, the serious adverse events of the desired root or parent events can be searched for. So, the serious adverse events that belong to the significant root and parent events were found in each dataframe. After that, the dataframes that contains the serious adverse events of significant root and parent events were merged with the found DBSCAN cluster labels. To see the serious adverse events that has the same DBSCAN labels, labels that belonged to just one serious adverse events were filtered out. For this, the labels were counted in the dataframes and those that appeared only once were discarded from the dataframes. By that, all the results that were acquired by using t-SNE/DBSCAN, spectral co-clustering and p-value calculations were used all together.

CHAPTER 4

RESULTS

4.1 Summary of the Initial Data

The collected data are summarized in Table 4.1 As a result of the search on the ClinicalTrials.gov we found 71 studies for Sorafenib, 177 for Doxorubicin and 99 for Cytarabine (347 in total). Since some of them were the same trials or did not report any serious adverse events which is the main focus of this study, they were discarded from the data (31 in total). This data contains study titles, ClinicalTrial.gov identifiers of studies, group numbers of studies and their descriptions, event terms, their assigned IDs and weights of the serious adverse events.

Table 4.1 The data summary.

Data Elements	Example	Data Summary
Trial studies	NCT00003896, NCT00005908	316
Trial groups	NCT00003896-1, NCT00005908-1	700
Distinct serious adverse events	Nausea, vomiting, fever	1725

The most seen 15 serious adverse events in the data can be seen in Figure 4.1.

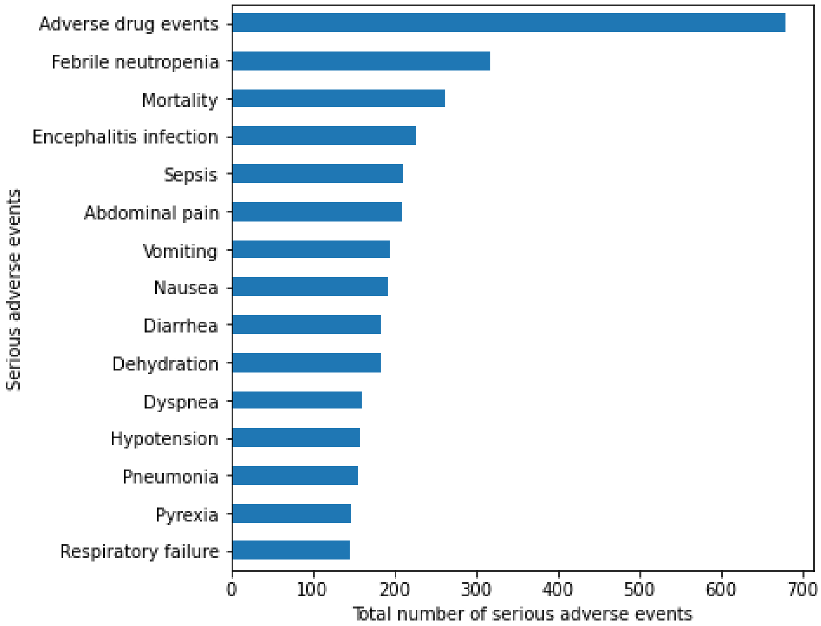


Figure 4.1. Total numbers of serious adverse events in the data.

The prevalence and the incidence of the serious adverse events were calculated (Table 4.2 and Table 4.3). In addition to the serious adverse events listed below, mortality has a prevalence at 38.92% and incidence at 20.59%.

Prevalence is the pecentage of the clinical trials that reported the serious adverse event. Incidence is the probability of occurrence of the serious adverse events. To calculate the incidence, the number of affected participants and number of participants at risk were added to the data. In the incidance calculation, the events that were reported in lower than five trial groups were excluded to discard uncommon events with high weights, since they would provide misleading statistics. For example, "lumbar puncture" found in one of the groups has a weight of 0.20, because the group had five participants. Since, the incidence is based on the average weight, it would be affected by these events greatly.

Table 4.2 Top twelve serious adverse events based on the prevalence.

Serious Adverse Events	Prevalence (%)
Drug-related side effects and adverse reactions	99.05
Febrile neutropenia	50.95
Fever	42.41
Pneumonia	39.87
Dyspnea	39.56
Nausea	38.92
Abdominal pain	38.29
Diarrhea	38.29
Dehydration	37.66
Vomiting	37.03
Sepsis	36.08
Hypotension	32.27

Table 4.3 Top nine serious adverse events based on the incidence.

Serious Adverse Events	Incidence (%)
Drug-related side effects and adverse reactions	25.57
Mucositis	8.05
Hand-foot syndrome	7.84
Febrile neutropenia	6.79
Disease progression	6.56
Catheter-related infections	6.38
Hemoglobin	5.63
Vomiting and Nausea	5.43
Hyponatremia (low blood sodium)	5.09

Table 4.4 Top ten serious adverse events based on the ranking of weight percentages.

Serious Adverse Events	Weight (%)
Drug-related side effects and adverse reactions	25.23
Febrile neutropenia	4.76
Pneumonia	1.64
Fever	1.56
Infectious Encephalitis	1.33
Neutropenia	1.13
Diarrhea	1.09
Nausea	1.00
Abdominal Pain	0.98
Vomiting	0.97

In addition to the events listed above, mortality constitutes 11% of the total weight.

4.2 Application of t-SNE and DBSCAN

4.2.1 Visualization of the Data

We used t-SNE to cluster the similar serious adverse events together. The unfiltered raw data was used to visualize the serious adverse events as data points, since the more attributes there are in the data, the better the prediction gets. The created clusters of the serious adverse events on the two dimensional map can be seen in Figure 4.2-a. This two dimensional map was plotted by using the output of the t-SNE that contains the coordinates of the data points which are the serious adverse events in this case. As mentioned in Section 3.6, the same output was also used in the DBSCAN application to label the data points based on the density that was created by the data points. With t-SNE and DBSCAN 344 clusters were found. Figure 4.2-b shows the two dimensional map which is colored by the DBSCAN labels.

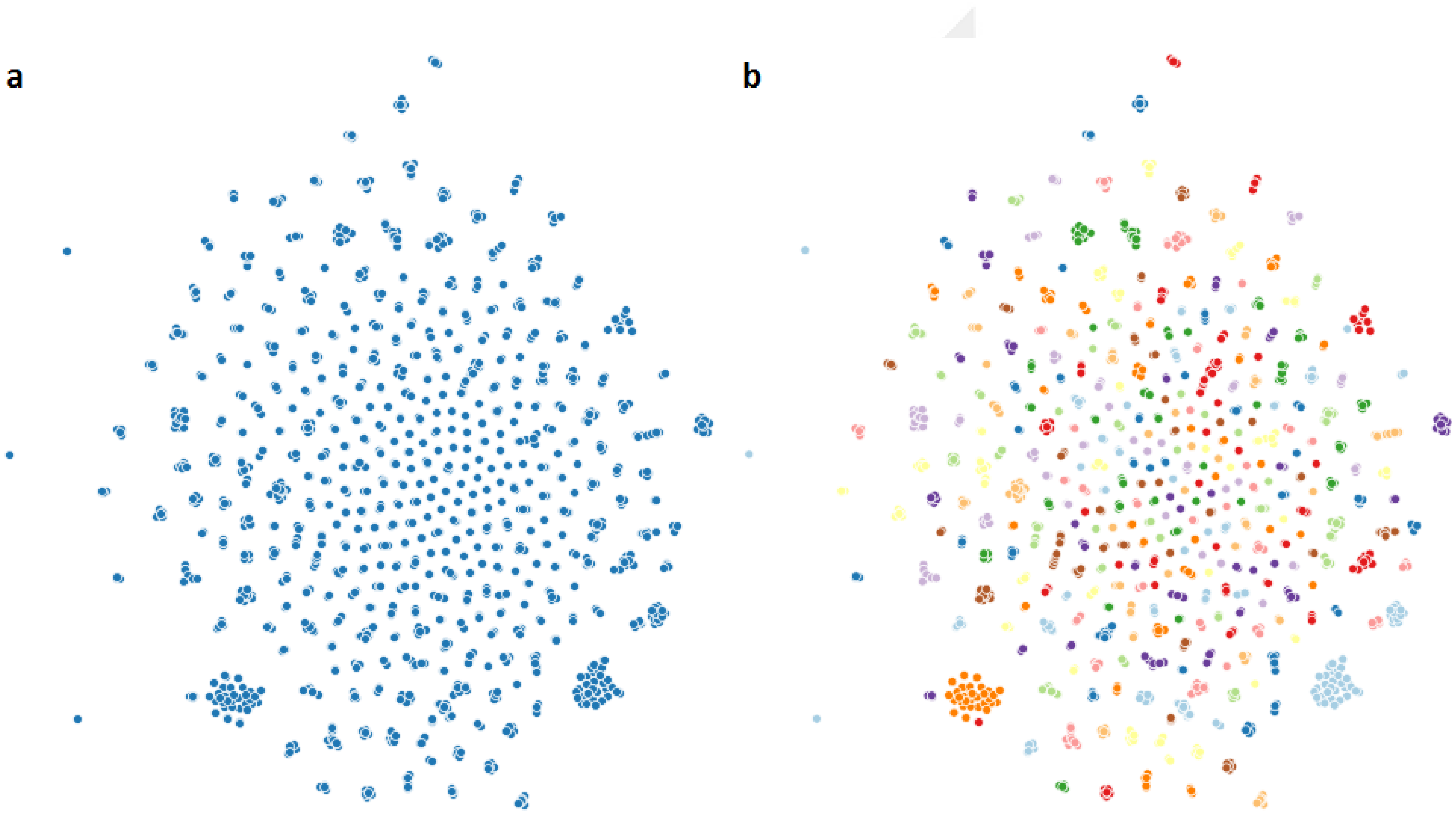


Figure 4.2. The two dimensional map created with t-SNE (a) and the colored points based on the DBSCAN labels (b).

4.2.2 Clusters of Similarity

The found DBSCAN labels and serious adverse event terms were merged together. Terms, IDs and corresponding DBSCAN labels of 1725 serious adverse events were stored in a dataframe to use later. Some of the serious adverse events and their DBSCAN labels can be seen in Table 4.5.

Table 4.5 Some of the serious adverse events and their DBSCAN labels.

DBSCAN Labels	Serious Adverse Events
1	Blood glucose, deoxyglucose, INR increased, serum albumin, activated partial thromboplastin time prolonged, fibrinogen decreased, calcium, magnesium, potassium
3	Alanine aminotransferase increased, aspartate aminotransferase increased, blood bilirubin increased, death NOS.
17	Influenza, dizziness, bone pain, joint pain, gingipains
92	Hyperglycemia, hyponatremia, hypoalbuminemia, fatigue, anorexia, postnasal drip
116	Tooth infection, gingivitis
195	Nausea, vomiting, constipation, diarrhea, abdominal pain, dehydration, headache, pleural effusion, cancer pain
275	Fever, febrile neutropenia, dyspnea, sepsis, pneumonia, infections upper respiratory, extrasystoles, renal cell carcinoma, adverse drug events, mortality
332	Sinus infections, respiratory failure, neuralgia

4.3 Application of Co-Clustering and p-Value Calculations

4.3.1 Assignment of Serious Adverse Events to the Drugs

As mentioned in Section 3.7, the groups that did not involve the indicated drugs (Sorafenib, Cytarabine and Doxorubicin) and the groups that involve the indicated drugs simultaneously were discarded from the dataset so they would not affect subsequent results. The number of studies was dropped to 229 and for groups to 401 after the filtering process.

The sparse matrix that has 1725 rows (serious adverse events) and 700 columns (groups) was created by the application of pivot operation on the filtered data. The weight of the serious adverse events were specified as values, the adverse event IDs as indices and group numbers as columns. Since we know which group involves which drug, the weights of the serious adverse events were aggregated and their means were taken based on drugs involved. Table 4.6 shows the mean weights of some serious adverse events for each drug type.

Table 4.6 Some of the serious adverse events and their means in the drug clusters.

Serious Adverse Events	Cytarabine	Doxorubicin	Sorafenib
Drug-related side effects and adverse reactions	45.24	39.28	36.53
Febrile neutropenia	12.33	6.76	1.55
Pneumonia	4.64	1.37	0.40
Fever	3.15	2.54	1.39
Platelet count decreased	0.73	0.91	1.96
Respiratory tract infections	0.11	0.23	0.09

We used spectral co-clustering to simultaneously cluster the rows and columns of the aggregated data which contains the mean weight of each serious adverse event for each drug type. The rows of the data which are the serious adverse events in this case, were labeled. Before the visualization of the clustered data the rows were put in order based on their labels by using their rearranged indices. After displaying the labeled rows based on which drug they were clustered in, the dense regions were selected for each column as indicated in Figure 4.3. Since the column index 1 belongs to Cytarabine, index 2 to Doxorubicin and index 3 to Sorafenib, the dense regions of the rows were assigned to three different dataframes along with their serious adverse event terms and corresponding IDs. For that, the indices of the sorted rows are selected from 0 to 330 for Cytarabine, from 330 to 640 for Doxorubicin and from 630 to 1725 for Sorafenib (Figure 4.4). Selected indices were used to index serious adverse events and store them in dataframes for each drug type along with their unique IDs. Overall, 330 serious adverse events were clustered in Cytarabine, 310 in Doxorubicin and 1085 in Sorafenib (Figure 4.5).

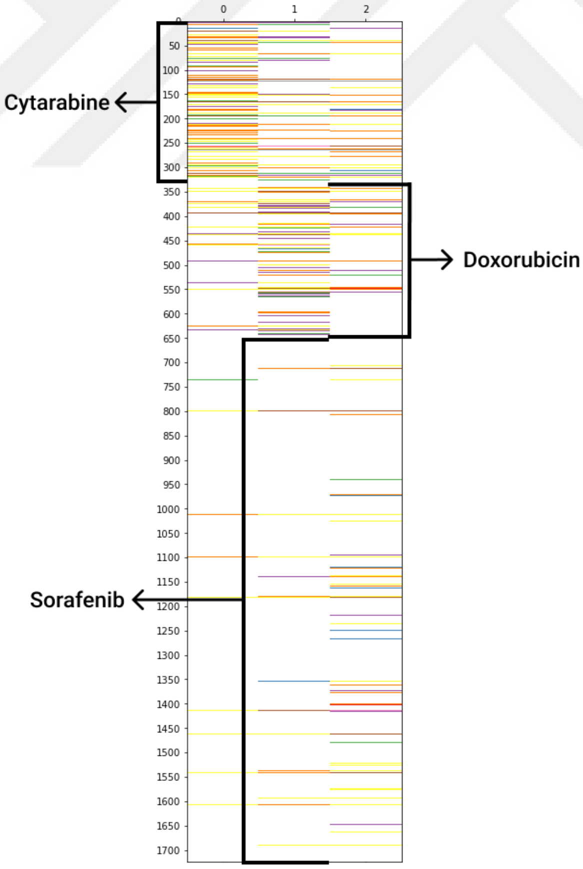
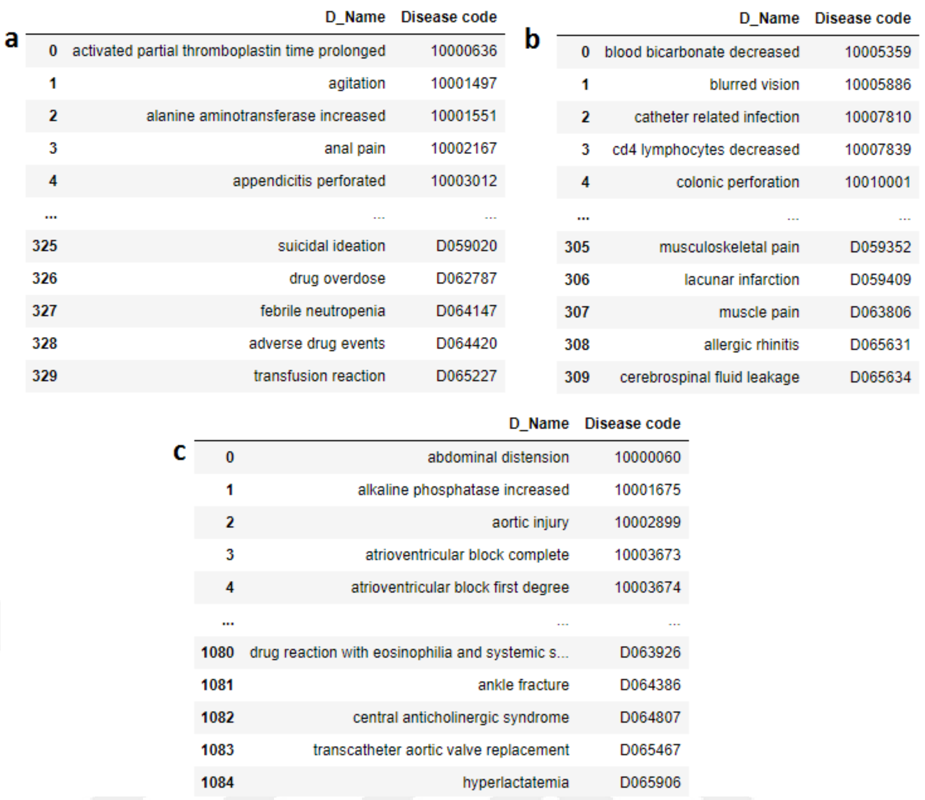


Figure 4.3. The output of spectral co-cluster and the drug selection.



a

	D_Name	Disease code
0	activated partial thromboplastin time prolonged	10000636
1	agitation	10001497
2	alanine aminotransferase increased	10001551
3	anal pain	10002167
4	appendicitis perforated	10003012
...	...	...
325	suicidal ideation	D059020
326	drug overdose	D062787
327	febrile neutropenia	D064147
328	adverse drug events	D064420
329	transfusion reaction	D065227

b

	D_Name	Disease code
0	blood bicarbonate decreased	10005359
1	blurred vision	10005886
2	catheter related infection	10007810
3	cd4 lymphocytes decreased	10007839
4	colonic perforation	10010001
...	...	...
305	musculoskeletal pain	D059352
306	lacunar infarction	D059409
307	muscle pain	D063806
308	allergic rhinitis	D065631
309	cerebrospinal fluid leakage	D065634

c

	D_Name	Disease code
0	abdominal distension	10000060
1	alkaline phosphatase increased	10001675
2	aortic injury	10002899
3	atrioventricular block complete	10003673
4	atrioventricular block first degree	10003674
...	...	...
1080	drug reaction with eosinophilia and systemic s...	D063926
1081	ankle fracture	D064386
1082	central anticholinergic syndrome	D064807
1083	transcatheter aortic valve replacement	D065467
1084	hyperlactatemia	D065906

Figure 4.4. The partial view of the dataframes that contains the serious adverse events that were clustered in Cytarabine (a), Doxorubicin (b) and Sorafenib (c).

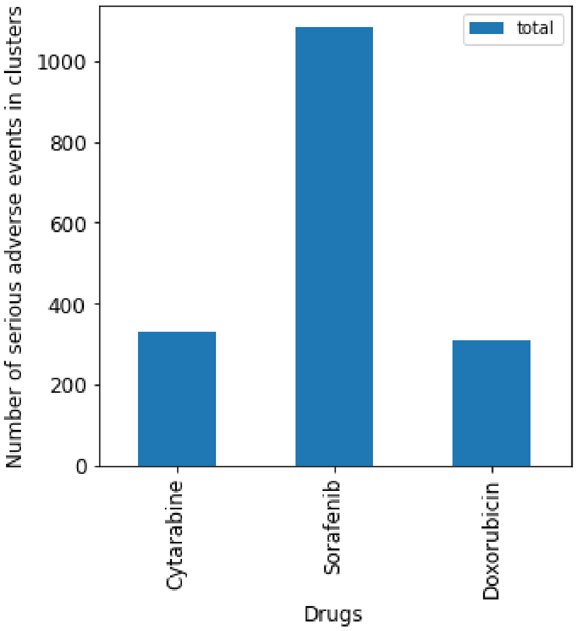


Figure 4.5. The number of serious adverse events in drug clusters.

4.3.2 P-Value Calculations

The p-values were calculated at a significance level of 0.05 based on the distributions of the root and the parent events of the serious adverse events among the three drugs. We found the root and parent events of the serious adverse events in the dictionaries. The found root and parent terms were added to the dataframes. Hence, the hierarchical structures of the dictionaries were applied to the dataframes.

Table 4.7 Some of the elements of the Cytarabine dataframe.

Index	Serious Adverse Event	Disease Code	Parent Event	Root Event
0	Activated partial thromboplastin time prolonged	10000636	Investigations	Investigations
1	Agitation	10001497	Psychiatric disorders	Psychiatric disorders
2	Alanine aminotransferase increased	10001551	Investigations	Investigations
3	Anal pain	10002167	Gastrointestinal disorders	Gastrointestinal disorders
4	Appendicitis perforated	10003012	Infections and infestations	Infections and infestations
5	Aspartate aminotransferase increased	10003481	Investigations	Investigations
6	Aspiration	10003504	Respiratory, thoracic and mediastinal disorders	Respiratory, thoracic and mediastinal disorders
7	Bladder infection	10005047	Infections and infestations	Infections and infestations
...	...	...	...	...
596	Transfusion reaction	D065227	Hematologic Diseases	Hemic and Lymphatic Diseases
597	Transfusion reaction	D065227	Transfusion Reaction	Immune System Diseases

The parent and root terms of some of the serious adverse events in the Cytarabine cluster can be seen in Table 4.7. The rows are increased considerably since some of the serious adverse events can belong to more than one tree in the hierarchical structure.

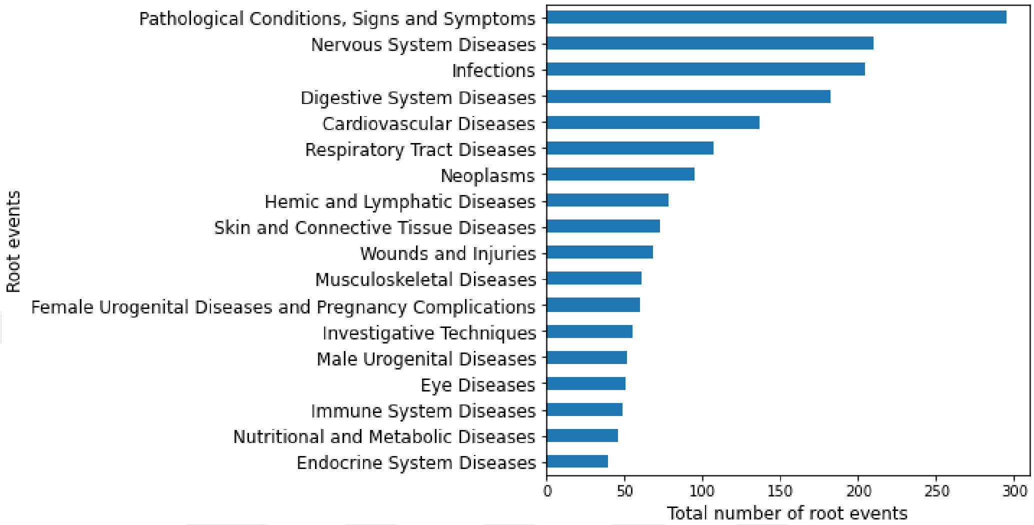


Figure 4.6. Total numbers of root events

The number of the root events (Figure 4.6) were counted for each dataframe and the results merged in a single dataframe (Table 4.8). Their p-values were calculated based on the distribution of root events among the three drugs by using the observed and expected numbers of root terms (Table 4.9).

We also repeated the same process for the parent events. The number of the parent events (Figure 4.8) were counted for each dataframe and the results merged in a single dataframe (Table 4.10). With the p-value calculations the significant drug specific events were found (Table 4.11).

Table 4.8 Numbers of root events in the drug clusters. The full version of the table is available in Appendix A.

Roots Drugs	Amino Acids, Peptides, and Proteins	Animal Diseases	Animal Structures	...	Urogenital System	Integumentary System Physiological Phenomena	Lipids
Cytarabine	5.0	1.0	1.0	...	0.0	0.0	0.0
Sorafenib	15.0	6.0	1.0	...	4.0	0.0	0.0
Doxorubicin	2.0	0.0	0.0	...	2.0	1.0	1.0

Table 4.9 Significant root events.

Root Events	p-Values
Bacteria	0.012698
Health Care Facilities, Manpower, and Services	0.013382
Heterocyclic Compounds	0.013573
Eye Diseases	0.015775
Hemic and Lymphatic Diseases	0.021762
Nutritional and Metabolic Diseases	0.025969
Infections	0.036822
Nervous System Diseases	0.053600

The number of expected and observed serious adverse events of the “Bacteria” root event in the drug clusters can be seen in Figure 4.7. The rates of all the root events are available in Appendix B.

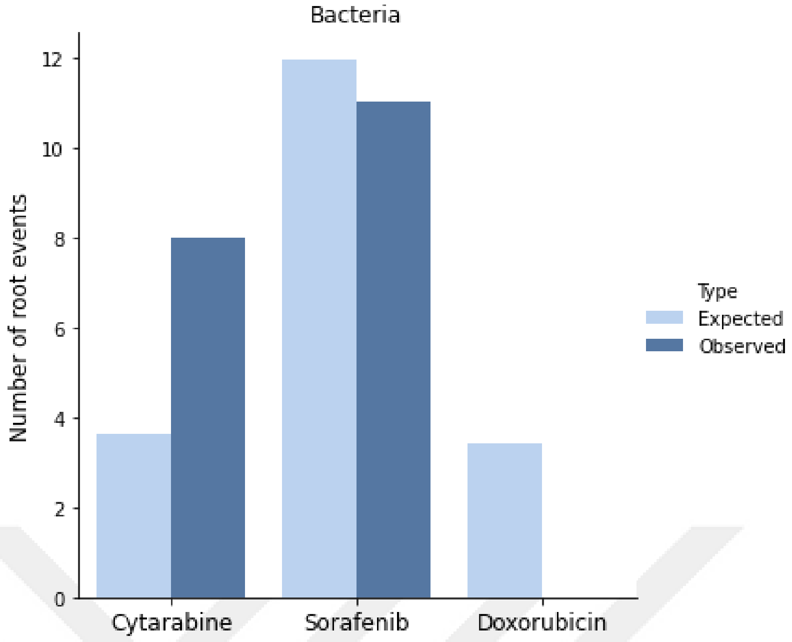


Figure 4.7. The expected and observed rates of the “Bacteria” root event.

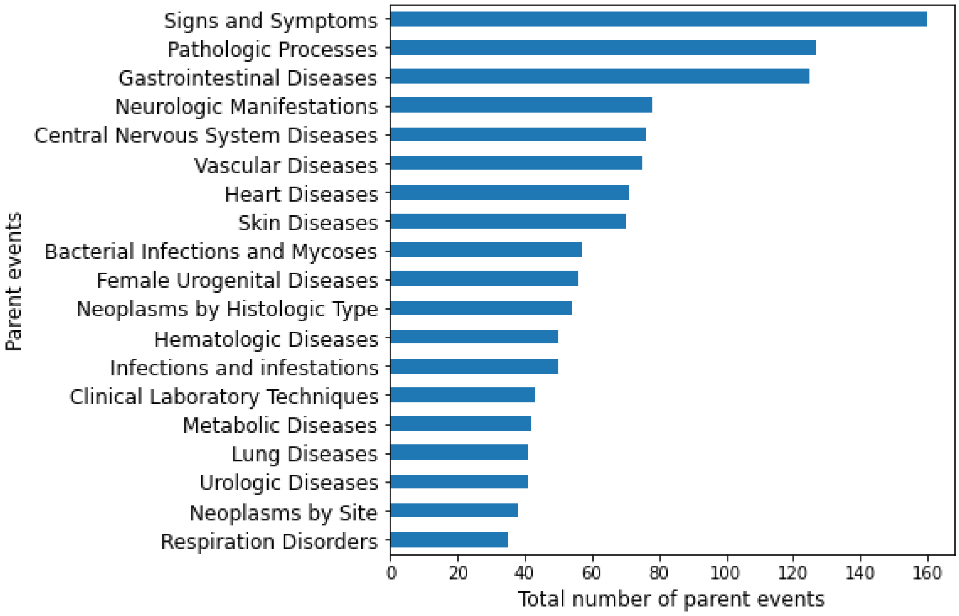


Figure 4.8. Total numbers of parent events

Table 4.10 Numbers of parent events in the drug clusters. The full version of the table is available in Appendix C.

Parents Drugs	Adrenal Gland Diseases	Amines	Amino Acids	...	Skin Physiological Phenomena	Urinary Tract Infections	Urogenital Abnormalities
Cytarabine	5.0	1.0	1.0	...	0.0	0.0	0.0
Sorafenib	15.0	6.0	1.0	...	4.0	0.0	0.0
Doxorubicin	2.0	0.0	0.0	...	2.0	1.0	1.0

Table 4.11 Significant parent events.

Parent Events	p-Values
Sulfur Compounds	0.000120
Fractures, Bone	0.000942
Hematologic Diseases	0.001168
Metabolic Diseases	0.006675
Gram-Negative Bacteria	0.007227
Health Facilities	0.010415
Proteobacteria	0.011919
Gingivitis	0.014592
Neurotoxicity Syndromes	0.014592
Neoplasms by Site	0.014787
Pathological Conditions, Anatomical	0.025986
Heterocyclic Compounds, 1-Ring	0.027130
Pathologic Processes	0.036845
Central Nervous System Infections	0.046825
Sepsis	0.053935

The number of expected and observed serious adverse events of the “Fractures, Bone” parent event in the drug clusters can be seen in Figure 4.9. The rates of all the parent events are available in Appendix D.

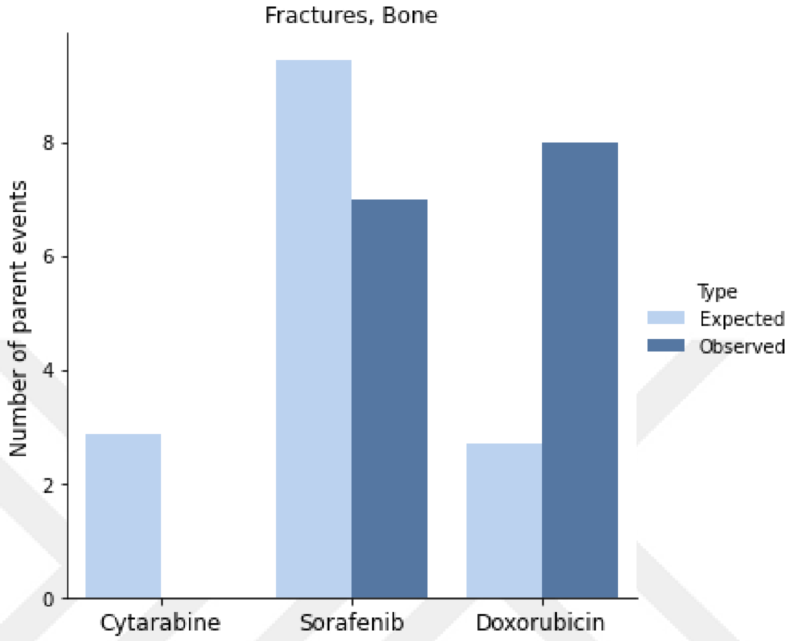


Figure 4.9. The expected and observed rates of the “Fractures, Bone” parent event.

4.4 Finding Similar or Simultaneous Co-Clustered Serious Adverse Events of Significant Root or Parent Events

The serious adverse events that belong to the significant root and parent events were selected from the dataframes that were created based on the drug clusters from the co-clustering method. After selection, they were merged with the dataframe that contains the DBSCAN labels of the serious adverse events. Since t-SNE uses attributes of data points to place them on a two dimensional map, the serious adverse events with the same labels should have similar attributes. So serious adverse events with the same label are either similar events or they occur together. By finding the DBSCAN labels of serious adverse events that belong to the significant root and parent events, results of all the methods that were used in this study were combined.

The tables that contains the serious adverse events of significant roots and parents that occur together can be seen below (Table 4.12 – Table 4.21) based on which drug they were clustered under and each with their similarity labels. Each background shade in tables denotes serious adverse events that belong to the same DBSCAN cluster. It should be noted some significant events and drug clusters have no serious adverse events associated with their DBSCAN labels.

Table 4.12 Serious adverse events of the significant "Eye Diseases" root events.

Drugs	Serious Adverse Events
Cytarabine	Eyestrain
	Blindness transient
Sorafenib	Retinopathy
	Watering eyes
	Papilledema
	Optic nerve disorder
	Endophthalmitis
	Strabismus
	Myopia
	Angle closure glaucoma
	Choroidal effusions
	Cataract

Table 4.13 Serious adverse events of the "Significant Hemic and Lymphatic Diseases" root event.

Drugs	Serious Adverse Events
Cytarabine	Mantle cell lymphoma
	Autoimmune haemolytic anaemia
	Lymph node pain
	Leukemia chronic myeloid
	Blast crisis
	Neutrophils
	Lymphocytes
Sorafenib	Anemia iron-deficiency
	Lymphocytic leukemia chronic

Table 4.13 Serious adverse events of the “Significant Hemic and Lymphatic Diseases” root event (continued).

Sorafenib	Von willebrand diseases
	Monoclonal gammopathies benign
	Granuloma
	Lymphoma b-cell marginal zone
	B cell lymphoma
Doxorubicin	Anemia
	Thrombocytopenia
	Leukopenia
	Neutropenia
	Platelets
	Granulocytes
	Leukocytes

Table 4.14 Serious adverse events of the significant “Nutritional and Metabolic Diseases” root event.

Drugs	**Serious Adverse Events**
Cytarabine	Hypermagnesemia
	Hypokalemia
	Hyperglycemia
	Hyponatremia
Sorafenib	Nutrition disorders
	Hyperlactatemia
	Obesity
	Latent autoimmune diabetes in adults
	Anemia iron-deficiency
	Water-electrolyte imbalance
	Pickwickian syndrome
Doxorubicin	Hypomagnesemia
	Hypocalcemia

Table 4.15 Serious adverse events of the significant "Infections" root event. Since there are 64 distinct serious adverse events in the Sorafenib cluster, some of them are listed.

Drugs	Serious Adverse Events
Cytarabine	Infection
	Typhlitis
	Abscess
	Staphylococcal infection
	Clostridium enterocolitis
	Arthritis septic
	Fungal infections
	Bronchopneumonia
	Pneumonia
	Sepsis
	Paratuberculosis
	Herpes simplex oral
Sorafenib	Amoebiasis
	Mycoplasma infection
	Cranial nerve infection
	Vaginal infection
	Malaria
	Herpetic meningoencephalitis
	Prostate infection
	Opisthorchis felineus infection
	Hidradenitis suppurativa
	...
	Flavivirus infection
	Retropharyngeal abscess
	Pseudomonas infections
	Meningitis pneumococcal
	Herpetic keratitis
	Staphylococcal scalded skin syndrome
Doxorubicin	Stoma site infection
	Paronychia
	Catheter-related infections
	Dipetalonema infections

Table 4.16 Serious adverse events of the significant "Nervous System Diseases" root event. Since there are 76 distinct events in the Sorafenib cluster, some of them are listed below.

Drugs	Serious Adverse Events
Cytarabine	Anoxic encephalopathy
	Multiple sclerosis
	Hemiparesis
	Aseptic meningitis
	Polyneuropathy
	Neurotoxicity syndrome manganese
	Peripheral sensory neuropathy
	Peripheral motor neuropathy
	Toxic encephalopathy
	Tinnitus
	Dyskinesia
Sorafenib	Facial muscle weakness
	Accessory nerve disorder
	Muscle weakness left-sided
	Optic nerve
	Trigeminal nerve disorder
	Hemianopia
	Deafness bilateral
	Temporal lobe epilepsy
	Alzheimer type senile dementia
	Complicated migraine
	Carpal tunnel syndrome
	...
	Listeria cerebritis
	Neurologic symptoms
	Paraplegia
	Paraparesis
Doxorubicin	Ataxia
	Dysarthria
	Agnosia for pain
	Muscle pain
	Insomnia

Table 4.16 Serious adverse events of the significant "Nervous System Diseases" root event. Since there are 76 distinct events in the Sorafenib cluster, some of them are listed below (continued).

Doxorubicin	Progeria-like syndrome
	Voice hoarseness
	Visual impairment
	Acute inflammatory demyelinating polyneuropathy
	Orthostatic hypotension
	Diabetic autonomic neuropathy
	Cerebrovascular accident
	Primary exertional headache
	Lacunar infarction
	Meninges

Table 4.17 Serious adverse events of the significant "Hematologic Diseases" parent event.

Drugs	**Serious Adverse Events**
Cytarabine	Blast crisis
	Leukemia chronic myeloid
	Lymphocytes
	Neutrophils
Sorafenib	Monoclonal gammopathies benign
	Anemia iron-deficiency
	Von willebrand diseases
Doxorubicin	Anemia
	Thrombocytopenia
	Leukopenia
	Neutropenia
	Platelets
	Granulocytes
	Leukocytes

Table 4.18 Serious adverse events of the significant “Metabolic Diseases” parent event in the Cytarabine cluster.

Drugs	Serious Adverse Events
Cytarabine	Hypermagnesemia
	Hypokalemia
	Hyperglycemia
	Hyponatremia
Sorafenib	Water-electrolyte imbalance
	Anemia iron-deficiency
Doxorubicin	Hypomagnesemia
	Hypocalcemia

Table 4.19 Serious adverse events of the significant “Neoplasm by Site” parent event.

Drugs	Serious Adverse Events
Sorafenib	Bone neoplasm
	Breast carcinoma in situ
	Laryngeal cancer
	Brain neoplasm
	Rectal cancer
	Bronchioloalveolar carcinoma
	Pancreatic neoplasm
	Acoustic neuroma
	Nasopharyngeal carcinoma
	Neoplasms cardiac
	Adrenal cancer

Table 4.20 Serious adverse events of the “Significant Pathological Conditions, Anatomical” parent event.

Drugs	Serious Adverse Events
Sorafenib	Cardiomegaly
	Hepatomegaly
	Muscle atrophy

Table 4.21 Serious adverse events of the significant “Central Nervous System Infections” parent event.

Drugs	Serious Adverse Events
Sorafenib	Cryptococcal meningitis
	Tuberculosis central nervous system

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Incidence Rate of Serious Adverse Events

For the interpretation of the serious adverse event incidence rates found in the data summary section, the difference between *serious adverse events* and *adverse events* should be well understood. It should be noted that serious adverse events are adverse events that are life threatening, and that require hospitalization [137]. So, vomiting that could lead to death is reported as a serious adverse event instead of an adverse event. Naturally, the incidence rate of serious adverse events are lower than their adverse event counterparts.

5.2 Similarity of Labels

To be sure about the result of the t-SNE/DBSCAN combination, they were run several times with different random states to check if the created clusters were random or not. Although the locations of the clusters and, accordingly, the two dimensional map changed in each run, the content of the clusters did not change and this indicates a strong relationship between clustered serious adverse events.

Some of the clusters were examined to check if they make sense or not. For instance, a cluster that contains nausea in it also contains; vomiting, diarrhea, constipation, abdominal pain, dehydration, headache, pleural effusion and cancer pain. Nausea and vomiting are very common symptoms for a cancer patient [138] and they often occur together [139]. Dehydration is one of the consequences of vomiting [140] and diarrhea [141]. It is also indicated that nausea, abdominal pain and vomiting are some of the constipation-related symptoms [142]. Headache could be related to any of these serious adverse events. While cancer pain is a very general term for any cluster, pleural effusion seems a bit unrelated.

As it can be seen in Table 4.5, serious adverse events in cluster number 1 are all about blood coagulation, since glucose metabolism affects coagulation [143], INR (International Normalized Ratio) and thromboplastin time are used in monitoring coagulation status [144], albumin is an anticoagulant which is often low in cancer patients [145], fibrinogen is involved in clotting [146] and calcium, magnesium and potassium are involved in blood coagulation [147], [148]. Cluster number 3 is related to the liver, since alanine aminotransferase, aspartate aminotransferase and serum bilirubin are used in liver function tests [149]. Hyperglycemia, hyponatremia, hypoalbuminemia, fatigue, anorexia and postnasal drip are clustered together in cluster number 92. Hyperglycemia is associated with a low level of sodium in the blood [150] which is called hyponatremia. Hyponatremia can be seen in patients with anorexia [151]. And fatigue is one of the symptoms of hyponatremia [152]. Cluster number 116 involves both tooth infection and gingivitis, which are related to each other [153]. Cluster number 275 is highly interesting since febrile neutropenia is the fever that occurs in patients with neutropenia [154]. Sepsis and pneumonia are the leading cause of death in cancer patients that have neutropenia [155]. Also, pneumonia is one of the causes of dyspnea [156]. Sinus infection and neuralgia are clustered together in the cluster 321 and their association was found in a study [157].

It should be noted, we can not expect t-SNE to be totally accurate since it is an unsupervised learning with no prior knowledge. Also, we used real-world data which is another challenge for t-SNE since real-word data can be complex and it is hard to predict a model which works best with such data [158]. Moreover, the data was a type of electronic health record (EHR). EHR’s are known to be difficult to represent and model due to sparness, heterogeneity, noise and clinical phenotypes being often expressed using different terminologies [159]. Considering all these, it can be said that the clusters created as a result of the t-SNE/DBSCAN combination are promising despite there being some dissimilar serious adverse events clustered together. Some of these, however, could indeed indicate potential hidden links.

5.3 Co-Clustered Data

The spectral co-clustering method was run several times with different random states on the aggregated data that contains the mean weights of the serious adverse events to check if the created clusters were random or not just like we did with the t-SNE/DBSCAN combination. The serious adverse event content in the dense regions does not change even though the whole body of the dense regions move up and down along the columns in each run. This indicates a strong relationship between clustered serious adverse events and drug types.

330 serious adverse events were clustered in Cytarabine, 310 in Doxorubicin and 1085 in Sorafenib and they were selected according to the dense regions on the heatmap that was created by the co-clustering method. It is clear that more serious adverse events were clustered in Sorafenib than others. This makes sense because the multi-tyrosine kinase inhibitors, like Sorafenib [160], tend to cause more serious adverse events [161], just like C. Federer et al. proposed [127] as mentioned in Section 2.2.

5.4 p-Values of Root and Parent Events

The drug and significant root or parent event relationships are evaluated below within the scope of available sources in the literature.

- The "Bacteria" root event is significant (p=0.0127), because the observed rate in the Doxorubicin cluster is much lower than the expected. Doxorubicin has antibacterial activity [162] since it is an antibiotic [81] and can inhibit the division of escherichia coli [163] which is one of the bacteria types in the "Bacteria" root event.

- Eye diseases are significant (p=0.0158) due to the observed rate in Sorafenib cluster. Sorafenib is associated with retinal tears, ocular irritation or blurred vision like some other anti-VEGF drugs [164].
- Hemic and lymphatic diseases are significant (p=0.0218) as a root event due to the observed rate in Cytarabine. But the reason for this significance could be the same reason why Cytarabine is used. Since Cytarabine is mostly used in the treatment and management of lymphoma and leukemia [72], the hemic and lymphatic diseases could already be observed in the patients such as leukocytosis [165] and lymphopenia [166] that were also found in the Cytarabine cluster that created with co-clustering. Also according to the Cytarabine prescription; hematologic diseases such as anemia, leukopenia and thrombocytopenia can be expected as a result of Cytarabine administration [167]. Thus, hematologic diseases as parent events should be significant as we found due to the observed rate of Cytarabine and this could be another reason why the hemic and lymphatic diseases are significant.
- Due to the increase in the observed number of "Infections" root event in the Cytarabine drugs, infections are significant (p=0.0368). High dose Cytarabine is found to increase the risk of infection and it is mainly used for adult patients with AML [168]. This could be the reason why the "Infections" root event is significant.
- The "Nervous System Diseases" root event is found to be almost significant (p=0.0536) because of the increase in the observed rate in Sorafenib cluster. But only one study was found to include this relationship by inducing neuropathy due administration of Sorafenib in the rat [169].
- Although nutritional and metabolic diseases are significant because of the observed rate in Cytarabine, no findings were found about the relationships of these diseases and related drugs.
- The other root events such as "Health Care Facilities, Manpower, and Services" and "Heterocyclic Compounds" were found to be significant but they do not mean anything. Most of the terms of these root terms were found

from clinical trials inaccurately due to reporting formats that are not compatible with a certain rule, superficial clinical report terms that have no the exact match in the two dictionaries or defects in the algorithm could lead to finding inappropriate terms other than indicated. For example, creatinine which is a chemical compound that is used to assess kidney function [170] was reported under the section of “Renal and urinary disorders”. Although creatinine is an element in the “Heterocyclic Compounds” MeSH tree. Fortunately, these mistakes are not that major to affect the significance of other events. Also, because we applied the hierarchical structure of the dictionaries, the found terms that are different from the indicated terms in clinical reports do not matter when their root or parent are considered only. For example, “infection without neutropenia” which is a superficial term to be used and it is not compatible with the reporting format of the two dictionaries was reported in a study. The found term was encephalitis infection for this clinical term and they are both in the “Infection” root event in the hierarchical structure.

- For the significant parent events, the “Fractures, Bone” parent event is significant (p=0.0009) due to the observed rate in the Doxorubicin cluster. It is known that Doxorubicin increases bone loss [171] and fractures due to its destructive effects on bones [172].
- Hematologic diseases are significant (p=0. 0012) due to the observed rate of Cytarabine and Doxorubicin. Like in the case of hemic and lymphatic diseases root events, hematologic diseases such as leukopenia and neutropenia are common in patients with leukemia [173]. Besides that, Cytarabine can lead to anemia, leukopenia and thrombocytopenia as mentioned before. Also, hematologic diseases such as febrile neutropenia, that was clustered in Cytarabine drug, can be expected with high dose Cytarabine administration [174]. Also, risk of platelet aggregation is increased after Doxorubicin treatment [175]. Due to the platelet cytotoxicity of Doxorubicin; thrombocytopenia, which is one of the serious adverse

events in the Doxorubicin cluster, can occur [176] and this could lead to platelet aggregation [177].

- Metabolic diseases are significant (p=0.007) due to the observed rate of Cytarabine, the “Neoplasm by site” parent event is significant (p=0.015) due to the observed rate in Sorafenib and central nervous system infections (p=0.047) due to Sorafenib. Although, no findings were found about the relationships of these parent events and related drugs.
- Both of the “Gram-Negative Bacteria” (p=0.007) and “Proteobacteria” (p=0.012) parent events are significant for the same reason the “Bacteria” parent event is. Their observed rates in the Doxorubicin cluster are very low due to the antibacterial activity, as explained above.
- The “Gingivitis” parent event is significant (0.015) due to the observed rate of Cytarabine, although nothing was found that can show the relation of gingivitis either with Cytarabine or leukemia.
- Neurotoxicity syndromes are significant (p=0.015) due to the observed rate of Cytarabine, and it is a well known effect of Cytarabine [178].
- The “Pathological conditions, Anatomical” parent event is significant (p=0.026) due to the observed rate of Sorafenib. This parent event mostly contains different kinds of hernias and fistulas in the Sorafenib cluster. Sorafenib can promote hernias due to its VEGFR and PDGFR inhibitor properties that have angiogenesis properties [179], and different kinds of fistulas can form during sorafenib treatment [180]–[182].
- The “Sepsis” parent event is almost significant (p=0.0539) due to the observed rate of Cytarabine. Sepsis is one of the consequences of treatments with high-dose Cytarabine [183], although it can be seen even with low-dose Cytarabine [184].
- The “Pathologic Processes” parent event is also significant (p=0.037) but serious adverse events can not be interpreted since they are very general.
- The “Heart Diseases” parent event (p=0.082) is not significant but it was expected to be since Doxorubicin is known for its cardiotoxicity [185].

Although Doxorubicin is cardiotoxic, all the serious adverse events whose parent events are heart diseases were clustered in Cytarabine.

- The other root events such as "Sulfur Compounds", "Health Facilities", "Heterocyclic Compounds, 1-Ring" are all significant but they do not mean anything. Again, their serious adverse events were found inaccurately.

5.5 Conclusion

We propose a method to discover the links between serious adverse events and drugs. We used ClinicalTrials.gov to download the clinical trial results that reported serious adverse events and in particular studied the anticancer drugs Cytarabine, Sorafenib and Doxorubicin. We retrieved 1725 distinct serious adverse events from 417 clinical trials and 700 trial groups. We used the MeSH and the CTCAE thesaurus to assign unique IDs to the serious adverse events to handle the inconsistency in the reports of serious adverse events. The serious adverse events were clustered on a two dimensional map based on their similarity by using t-SNE. For the labeling of these clusters, DBSCAN was used. To cluster the serious adverse events based on the drugs, we used spectral co-clustering. The hierarchical structure of the dictionaries were also applied to the problem of finding high level events of the serious adverse events. We calculated the p-values of the root and parent events to find significant ones that are encountered more or less, relatively, in a specific drug. Overall, we used data analysis methods to analyze drug-serious adverse events relationships.

Discovering links between drugs and serious adverse events and finding significant serious adverse events by using these drug-serious adverse event relationships are the motivations of this study. To achieve them we transformed various clinical trial data into a table structure and important insights were derived from the data based on the obtained clusters. We showed that, by analyzing the clustered serious adverse events, events that are significantly associated with drugs can be detected. We believe that the methods we proposed for this study would provide opportunities to detect serious adverse events that are seen more or less in the treatment with a specific drug. Since most of the serious adverse events that occur in anticancer drug therapy occur in all drugs, it is important to detect drug-specific serious adverse events. By comparing observed rates of serious adverse events of a new drug with other drugs, serious adverse events that are specific to a new treatment option can be discovered. Also, with the similarity labeling, the serious adverse events that trigger each other, occur together or that are similar to each other can be found.

Different reporting formats of the serious adverse events made data extraction and analysis difficult. Some of the results were affected by this inconsistency. Clinical trial reports that are compatible with a certain standard may provide more accurate results for studies like this. With the increasing number of clinical trials, the reliability of the results may increase in the future. And also, increased drug diversity could derive more insights about the drug-serious adverse event relationships. The in focus disease could vary with the chosen drug types. For instance; blood pressure, antibiotic, diabetic and allergy drugs could be selected and analyzed with the proposed methods. Clinical studies that include drug combinations could be used to analyze the effect of multi-drug applications on serious adverse events. Also, the gene targets of the chosen drugs can be included in the study to find the pathways triggered and this would lead to pathway enrichment analysis and projection of symptoms from pathways. Choosing different clustering algorithms that are compatible with the data, roll up or drill down operations that involve different detail levels of the data, and refinement of the ID assignment algorithm that would find terms even more accurately could have an important influence on the results. Although some similarity clusters contain unrelated serious adverse events and some of the drug-serious adverse event relationships are not available in the literature, these could indicate links that would be discovered in the future. This study can find certain patterns, turn data into actionable knowledge and provide insights to decision makers such as doctors or clinicians about the potential serious adverse events. So it has the potential to influence treatment decisions, medical search and health policies.

We hope that our study will contribute to the scientific literature with both the proposed methodology and the results obtained, in the matter of using data science to find the relationships between anticancer drugs and serious adverse events.

REFERENCES

[1] G. Gopal, C. Suter-Crazzolara, L. Toldo, and W. Eberhardt, “Digital transformation in healthcare - Architectures of present and future information technologies,” in *Clinical Chemistry and Laboratory Medicine*, Mar. 2019, vol. 57, no. 3, pp. 328–335, doi: 10.1515/cclm-2018-0658.

[2] E. Hutchings, M. Loomes, P. Butow, and F. M. Boyle, “A systematic literature review of researchers’ and healthcare professionals’ attitudes towards the secondary use and sharing of health administrative and clinical trial data,” *Syst. Rev.*, vol. 9, no. 1, Oct. 2020, doi: 10.1186/s13643-020-01485-5.

[3] B. Goldacre, “Are clinical trial data shared sufficiently today? No,” *BMJ (Online)*, vol. 347, no. 7916. BMJ, Jul. 13, 2013, doi: 10.1136/bmj.f1880.

[4] R. Pastorino *et al.*, “Benefits and challenges of Big Data in healthcare: An overview of the European initiatives,” *Eur. J. Public Health*, vol. 29, no. Suppl 3, pp. 23–27, Oct. 2019, doi: 10.1093/eurpub/ckz168.

[5] I. of M. (US), L. Olsen, R. S. Saunders, and J. M. McGinnis, “Clinical Data as a Public Good for Discovery,” 2011, Accessed: Apr. 08, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK92065/.

[6] “Clinical Data - Data Resources in the Health Sciences - Library Guides at University of Washington Libraries.” https://guides.lib.uw.edu/hsl/data/findclin#s-lg-box-1908462 (accessed Apr. 08, 2021).

[7] J. M. Novitzke, “The significance of clinical trials,” Zeenat Qureshi Stroke Research Center, 2008. Accessed: Apr. 08, 2021. [Online]. Available: /pmc/articles/PMC3317309/.

[8] I. of M. (US) R. on V. & S.-D. H. Care, “Summary,” 2010, Accessed: Apr. 08, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK54290/.

[9] A. Schultz, B. R. Saville, J. A. Marsh, and T. L. Snelling, “An introduction to clinical trial design,” *Paediatric Respiratory Reviews*, vol. 32. W.B. Saunders Ltd, pp. 30–35, Nov. 01, 2019, doi: 10.1016/j.prrv.2019.06.002.

[10] S. J. Nass, L. A. Levit, and L. O. Gostin, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press, 2009.

[11] G. Moody, K. Addison, R. Cannings-John, J. Sanders, C. Wallace, and M. Robling, “Monitoring adverse social and medical events in public health trials: Assessing predictors and interpretation against a proposed model of adverse event reporting,” *Trials*, vol. 20, no. 1, Dec. 2019, doi: 10.1186/s13063-019-3961-8.

[12] R. Chou *et al.*, “AHRQ Series Paper 4: Assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program,” *Journal of Clinical Epidemiology*, vol. 63, no. 5. J Clin Epidemiol, pp. 502–512, May 2010, doi: 10.1016/j.jclinepi.2008.06.007.

[13] “Glossary of Common Site Terms - ClinicalTrials.gov.” https://clinicaltrials.gov/ct2/about-studies/glossary (accessed Apr. 09, 2021).

[14] M. Watanabe *et al.*, “Serious adverse events and compensation in registration trials: A review of data from a Japanese university hospital,” *BMC Res. Notes*, vol. 7, no. 1, p. 245, Apr. 2014, doi: 10.1186/1756-0500-7-245.

[15] L. A. Ladewski *et al.*, “Dissemination of information on potentially fatal adverse drug reactions for cancer drugs from 2000 to 2002: First results from the research on adverse drug events and reports project,” *J. Clin. Oncol.*, vol. 21, no. 20, pp. 3859–3866, Oct. 2003, doi:

10.1200/JCO.2003.04.537.

[16] D. A. Dorr *et al.*, “Quality of reporting of serious adverse drug events to an institutional review board: A case study with the novel cancer agent, imatinib mesylate,” *Clin. Cancer Res.*, vol. 15, no. 11, pp. 3850–3855, Jun. 2009, doi: 10.1158/1078-0432.CCR-08-1811.

[17] I. Ingrand *et al.*, “Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey,” *Br. J. Clin. Pharmacol.*, vol. 86, no. 4, pp. 711–722, Apr. 2020, doi: 10.1111/bcp.14159.

[18] K. Kawasumi, A. Kujirai, R. Matsui, Y. Kawano, M. Yamaguchi, and T. Aoyama, “Survey of serious adverse events and safety evaluation of oral anticancer drug treatment in Japan: A retrospective study,” *Mol. Clin. Oncol.*, vol. 14, no. 1, pp. 1–9, 2020, doi: 10.3892/mco.2020.2174.

[19] A. Sharma, J. Thomas, K. Bairy, Km. Kumari, and H. Manohar, “Pattern of adverse drug reactions due to cancer chemotherapy in a tertiary care hospital in South India,” *Perspect. Clin. Res.*, vol. 6, no. 2, p. 109, 2015, doi: 10.4103/2229-3485.154014.

[20] L. Crielaard and P. Papapetrou, “Explainable predictions of adverse drug events from electronic health records via oracle coaching,” in *IEEE International Conference on Data Mining Workshops, ICDMW*, Feb. 2019, vol. 2018-November, pp. 707–714, doi: 10.1109/ICDMW.2018.00108.

[21] J. S. Ross, G. K. Mulvey, E. M. Hines, S. E. Nissen, and H. M. Krumholz, “Trial Publication after Registration in ClinicalTrials.Gov: A Cross-Sectional Analysis,” *PLoS Med.*, vol. 6, no. 9, p. e1000144, Sep. 2009, doi: 10.1371/journal.pmed.1000144.

[22] “ClinicalTrials.gov Background - ClinicalTrials.gov.” https://clinicaltrials.gov/ct2/about-site/background (accessed Apr. 12, 2021).

[23] “Advanced Search - ClinicalTrials.gov.”

https://clinicaltrials.gov/ct2/search/advanced?cond=&term=&cntry=&state=&city=&dist= (accessed Apr. 12, 2021).

[24] B. R. Hirsch *et al.*, “Characteristics of oncology clinical trials: Insights from a systematic analysis of clinicaltrials.gov,” *JAMA Intern. Med.*, vol. 173, no. 11, pp. 972–979, Jun. 2013, doi: 10.1001/jamainternmed.2013.627.

[25] T. Tse, K. M. Fain, and D. A. Zarin, “How to avoid common problems when using clinicaltrials.gov in research: 10 issues to consider,” *BMJ*, vol. 361, 2018, doi: 10.1136/bmj.k1452.

[26] “About the Results Database - ClinicalTrials.gov.” https://clinicaltrials.gov/ct2/about-site/results (accessed Apr. 12, 2021).

[27] D. A. Zarin and T. Tse, “Sharing Individual Participant Data (IPD) within the Context of the Trial Reporting System (TRS),” *PLoS Med.*, vol. 13, no. 1, 2016, doi: 10.1371/journal.pmed.1001946.

[28] S. Pranić and A. Marušić, “Changes to registration elements and results in a cohort of Clinicaltrials.gov trials were not reflected in published articles,” *J. Clin. Epidemiol.*, vol. 70, pp. 26–37, Feb. 2016, doi: 10.1016/j.jclinepi.2015.07.007.

[29] D. A. Zarin, N. C. Ide, T. Tse, W. R. Harlan, J. C. West, and D. A. B. Lindberg, “Issues in the registration of clinical trials,” *Journal of the American Medical Association*, vol. 297, no. 19. American Medical Association, pp. 2112–2120, May 16, 2007, doi: 10.1001/jama.297.19.2112.

[30] A. Tasneem *et al.*, “The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty,” *PLoS One*, vol. 7, no. 3, p. e33677, Mar. 2012, doi: 10.1371/journal.pone.0033677.

[31] H. Singh, R. Singh, A. Malhotra, and M. Kaur, “Developing a biomedical expert finding system using medical subject headings,” *Healthc. Inform.*

Res., vol. 19, no. 4, pp. 243–249, Dec. 2013, doi: 10.4258/hir.2013.19.4.243.

[32] X. Ma *et al.*, “Development of a controlled vocabulary for semantic interoperability of mineral exploration geodata for mining projects,” *Comput. Geosci.*, vol. 36, no. 12, pp. 1512–1522, Dec. 2010, doi: 10.1016/j.cageo.2010.05.014.

[33] “Frequently Asked Questions about Indexing.” https://www.nlm.nih.gov/bsd/indexfaq.html (accessed May 03, 2021).

[34] G. Tsatsaronis, I. Varlamis, N. Kanhabua, and K. Nørvåg, “LNCS 7816 - Temporal Classifiers for Predicting the Expansion of Medical Subject Headings.” Accessed: Apr. 25, 2021. [Online]. Available: http://www.geneontology.org/.

[35] S. J. Nelson, W. D. Johnston, and B. L. Humphreys, “Relationships in Medical Subject Headings (MeSH),” 2001, pp. 171–184.

[36] A. Gelbukh, Ed., *Computational Linguistics and Intelligent Text Processing*, vol. 7816. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[37] H. J. Lowe, “Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches,” *JAMA J. Am. Med. Assoc.*, vol. 271, no. 14, p. 1103, Apr. 1994, doi: 10.1001/jama.1994.03510380059038.

[38] A. Kabirzadeh, H. Siamian, E. B. F. Abadi, and B. M. Saravi, “Survey of keyword adjustment of published articles medical subject headings in journal of mazandaran university of medical sciences (2009-2010),” *Acta Inform. Medica*, vol. 21, no. 2, pp. 98–102, 2013, doi: 10.5455/aim.2013.21.98-102.

[39] A. A. Chang, K. M. Heskett, and T. M. Davidson, “Searching the literature using medical subject headings versus text word with PubMed,” *Laryngoscope*, vol. 116, no. 2, pp. 336–340, Feb. 2006, doi:

10.1097/01.mlg.0000195371.72887.a2.

[40] "MeSH Browser." https://meshb.nlm.nih.gov/record/ui?ui=D013274 (accessed Apr. 27, 2021).

[41] A. Kartika Sari, "Mapping of Change Operations from Gene Ontology into Medical Subject Headings," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 4, 2020, doi: 10.22266/ijies2020.0831.05.

[42] C. E. Lipscomb, "Medical Subject Headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3. Medical Library Association, pp. 265–266, 2000, Accessed: Apr. 25, 2021. [Online]. Available: /pmc/articles/PMC35238/.

[43] J. Voss, "Collaborative thesaurus tagging the Wikipedia way," *ArXiv*, 2006.

[44] I. Dhammi and S. Kumar, "Medical subject headings (MeSH) terms," *Indian Journal of Orthopaedics*, vol. 48, no. 5. Medknow Publications, pp. 443–444, Sep. 01, 2014, doi: 10.4103/0019-5413.139827.

[45] N. Baumann, "How to use the medical subject headings (MeSH)," *Int. J. Clin. Pract.*, vol. 70, no. 2, pp. 171–174, Feb. 2016, doi: 10.1111/ijcp.12767.

[46] C. Moreau-Bachelard, E. Coquan, and C. Le Tourneau, "Imputability of Adverse Events to Anticancer Drugs," *N. Engl. J. Med.*, vol. 380, no. 19, pp. 1873–1874, May 2019, doi: 10.1056/NEJMc1900053.

[47] Z. Wang *et al.*, "Risk of serious adverse event and fatal adverse event with molecular target anticancer drugs in cancer patients: A meta-analysis," *Journal of Cancer Research and Therapeutics*, vol. 15, no. 7. Wolters Kluwer Medknow Publications, pp. 1435–1449, Oct. 01, 2019, doi: 10.4103/jcrt.JCRT_577_18.

[48] A. P. Chen *et al.*, "Grading dermatologic adverse events of cancer treatments: The common terminology criteria for adverse events version 4.0," *Journal of the American Academy of Dermatology*, vol. 67, no. 5.

Mosby Inc., pp. 1025–1039, Nov. 01, 2012, doi: 10.1016/j.jaad.2012.02.010.

[49] A. Trotti *et al.*, “CTCAE v3.0: Development of a comprehensive grading system for the adverse effects of cancer treatment,” in *Seminars in Radiation Oncology*, 2003, vol. 13, no. 3, pp. 176–181, doi: 10.1016/S1053-4296(03)00031-6.

[50] S. Zhang, Q. Chen, and Q. Wang, “The use of and adherence to CTCAE v3.0 in cancer clinical trial publications,” *Oncotarget*, vol. 7, no. 40, pp. 65577–65588, 2016, doi: 10.18632/oncotarget.11576.

[51] T. P. Miller *et al.*, “Unintended consequences of evolution of the Common Terminology Criteria for Adverse Events,” *Pediatr. Blood Cancer*, vol. 66, no. 7, p. e27747, Jul. 2019, doi: 10.1002/pbc.27747.

[52] L. M. Wintner, J. M. Giesinger, M. Sztankay, A. Bottomley, and B. Holzner, “Evaluating the use of the EORTC patient-reported outcome measures for improving inter-rater reliability of CTCAE ratings in a mixed population of cancer patients: study protocol for a randomized controlled trial,” *Trials*, vol. 21, no. 1, Dec. 2020, doi: 10.1186/s13063-020-04745-w.

[53] H. Iihara *et al.*, “Evaluation of clinical pharmacist interventions for adverse events in hospitalized patients with thoracic cancer receiving cancer chemotherapy,” *Mol. Clin. Oncol.*, vol. 14, no. 6, Apr. 2021, doi: 10.3892/mco.2021.2278.

[54] A. C. Dueck *et al.*, “Validity and reliability of the us national cancer institute’s patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE),” *JAMA Oncol.*, vol. 1, no. 8, pp. 1051–1059, Nov. 2015, doi: 10.1001/jamaoncol.2015.2639.

[55] S.-C. Lee *et al.*, “Common Terminology Criteria for Adverse Events (CTCAE): Redesign and Life Cycle Management,” Apr. 2010. Accessed: May 03, 2021. [Online]. Available: https://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03/Documentation/CTCAE_

Governance_2010-03-11.pdf.

[56] C. A. Brandt, C. C. Lu, and P. M. Nadkarni, “Automating identification of adverse events related to abnormal lab results using standard vocabularies.,” *AMIA Annu. Symp. Proc.*, vol. 2005, p. 903, 2005, Accessed: May 03, 2021. [Online]. Available: http://ctep.cancer.gov/forms/CTCAEv3.pdf.

[57] Y. Yu *et al.*, “Coverage Evaluation of CTCAE for Capturing the Immune-related Adverse Events Leveraging Text Mining Technologies.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2019, pp. 771–778, 2019, Accessed: May 03, 2021. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/31259034.

[58] National Cancer Institute and U.S. Department of Health and Human Services, “Common Terminology Criteria for Adverse Events (CTCAE) v5.0,” 2017. Accessed: May 03, 2021. [Online]. Available: ctep.cancer.gov.

[59] B. I. Rini, “Sorafenib,” *Expert Opin. Pharmacother.*, vol. 7, no. 4, pp. 453–461, Mar. 2006, doi: 10.1517/14656566.7.4.453.

[60] V. Gupta-Abramson *et al.*, “Phase II trial of sorafenib in advanced thyroid cancer,” *J. Clin. Oncol.*, vol. 26, no. 29, pp. 4714–4719, Oct. 2008, doi: 10.1200/JCO.2008.16.3279.

[61] O. Hahn and W. Stadler, “Sorafenib,” *Current Opinion in Oncology*, vol. 18, no. 6. pp. 615–621, Nov. 2006, doi: 10.1097/01.cco.0000245316.82391.52.

[62] G. M. Keating and A. Santoro, “Sorafenib: A review of its use in advanced hepatocellular carcinoma,” *Drugs*, vol. 69, no. 2. Springer, pp. 223–240, Oct. 15, 2009, doi: 10.2165/00003495-200969020-00006.

[63] H. B. El-Serag and K. L. Rudolph, “Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis,” *Gastroenterology*, vol. 132, no. 7. Gastroenterology, pp. 2557–2576, Jun. 2007, doi: 10.1053/j.gastro.2007.04.061.

[64] N. Ferrara, H. P. Gerber, and J. LeCouter, “The biology of VEGF and its receptors,” *Nature Medicine*, vol. 9, no. 6. Nat Med, pp. 669–676, Jun. 01, 2003, doi: 10.1038/nm0603-669.

[65] L. F. Allen, J. Sebolt-Leopold, and M. B. Meyer, “CI-1040 (PD184352), a Targeted Signal Transduction Inhibitor of MEK (MAPKK),” in *Seminars in Oncology*, 2003, vol. 30, no. 5 SUPPL. 16, pp. 105–116, doi: 10.1053/j.seminoncol.2003.08.012.

[66] Y. Liu, Y. Ding, Y. Nie, and M. Yang, “EMP1 promotes the proliferation and invasion of ovarian cancer cells through activating the MAPK pathway,” *Onco. Targets. Ther.*, vol. 13, pp. 2047–2055, 2020, doi: 10.2147/OTT.S240028.

[67] J. Coventon, “A review of the mechanism of action and clinical applications of sorafenib in advanced osteosarcoma,” *J. Bone Oncol.*, vol. 8, pp. 4–7, Sep. 2017, doi: 10.1016/j.jbo.2017.07.001.

[68] S. M. Wilhelm *et al.*, “BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis,” *Cancer Res.*, vol. 64, no. 19, pp. 7099–7109, Oct. 2004, doi: 10.1158/0008-5472.CAN-04-1443.

[69] S. L. Chan and W. Yeo, “Targeted therapy of hepatocellular carcinoma: Present and future,” *Journal of Gastroenterology and Hepatology (Australia)*, vol. 27, no. 5. Blackwell Publishing, pp. 862–872, May 01, 2012, doi: 10.1111/j.1440-1746.2012.07096.x.

[70] J. C. M. Uitdehaag *et al.*, “Comparison of the cancer gene targeting and biochemical selectivities of all targeted kinase inhibitors approved for clinical use,” *PLoS One*, vol. 9, no. 3, p. 92146, Mar. 2014, doi: 10.1371/journal.pone.0092146.

[71] B. J. Rider, “Cytarabine,” in *xPharm: The Comprehensive Pharmacology*

Reference, Elsevier Inc., 2007, pp. 1–5.

[72] H. I. El-Subbagh and A. A. Al-Badr, “Chapter 2 Cytarabine,” in *Profiles of Drug Substances, Excipients and Related Methodology*, vol. 34, Academic Press Inc., 2009, pp. 37–113.

[73] D. L. Betcher and N. Burnham, “Cytarabine,” *J. Pediatr. Oncol. Nurs.*, vol. 7, no. 4, pp. 154–157, Jan. 1990, doi: 10.1177/104345429000700406.

[74] M. C. Chamberlain, S. Khatibi, J. C. Kim, S. B. Howell, E. Chatelut, and S. Kim, “Treatment of Leptomeningeal Metastasis With Intraventricular Administration of Depot Cytarabine (DTC 101): A Phase I Study,” *Arch. Neurol.*, vol. 50, no. 3, pp. 261–264, Mar. 1993, doi: 10.1001/ARCHNEUR.1993.00540030027009.

[75] F. L. Graham and G. F. Whitmore, “The Effect of 1-β-d-Arabinofuranosylcytosine on Growth, Viability, and DNA Synthesis of Mouse L-cells,” *Cancer Res.*, vol. 30, no. 11, 1970.

[76] J. Stentoft, “The Toxicity of Cytarabine,” *Drug Safety*, vol. 5, no. 1. Springer, pp. 7–27, Oct. 17, 1990, doi: 10.2165/00002018-199005010-00003.

[77] N. Shahabadi, M. Falsafi, and M. Maghsudi, “DNA-binding study of anticancer drug cytarabine by spectroscopic and molecular docking techniques,” *http://dx.doi.org/10.1080/15257770.2016.1218021*, 2016, doi: 10.1080/15257770.2016.1218021.

[78] R. Di Francia *et al.*, “Response and toxicity to cytarabine therapy in leukemia and lymphoma: From dose puzzle to pharmacogenomic biomarkers,” *Cancers*, vol. 13, no. 5. MDPI AG, pp. 1–39, Mar. 01, 2021, doi: 10.3390/cancers13050966.

[79] J. H. Lass, H. M. Lazarus, M. D. Reed, and R. H. Herzig, “Topical corticosteroid therapy for corneal toxicity from systemically administered

cytarabine,” *Am. J. Ophthalmol.*, vol. 94, no. 5, pp. 617–621, Nov. 1982, doi: 10.1016/0002-9394(82)90006-X.

[80] M. C. Chamberlain, S. Khatibi, J. C. Kim, S. B. Howell, E. Chatelut, and S. Kim, “Treatment of Leptomeningeal Metastasis with Intraventricular Administration of Depot Cytarabine (DTC 101): A Phase I Study,” *Arch. Neurol.*, vol. 50, no. 3, pp. 261–264, Mar. 1993, doi: 10.1001/archneur.1993.00540030027009.

[81] P. Krzesiński, R. Wierzbowski, G. Gielerak, J. Hałka, O. Matysiak, and P. Smurzyński, “Impedance cardiography in the diagnosis of capillary leak syndrome caused by doxorubicin therapy in a patient with myeloma multiplex,” *CASE Rep. Cardiol. J.*, vol. 17, no. 1, pp. 88–91, 2010, Accessed: Jun. 08, 2021. [Online]. Available: www.cardiologyjournal.org.

[82] K. Johnson-Arbor and R. Dubey, “Doxorubicin,” *StatPearls. StatPearls Publ.*, Jan. 2021.

[83] D. M. Rayner and S. M. Cutts, “Anthracyclines,” in *Side Effects of Drugs Annual*, vol. 36, Elsevier B.V., 2014, pp. 683–694.

[84] K. Agrawal, “Doxorubicin,” in *xPharm: The Comprehensive Pharmacology Reference*, Elsevier Inc., 2007, pp. 1–5.

[85] C. F. Thorn *et al.*, “Doxorubicin pathways: Pharmacodynamics and adverse effects,” *Pharmacogenet. Genomics*, vol. 21, no. 7, pp. 440–446, 2011, doi: 10.1097/FPC.0b013e32833ffb56.

[86] S. Yang *et al.*, “Cancer-activated doxorubicin prodrug nanoparticles induce preferential immune response with minimal doxorubicin-related toxicity,” *Biomaterials*, vol. 272, p. 120791, May 2021, doi: 10.1016/j.biomaterials.2021.120791.

[87] M. Wattenberg, F. Viégas, and I. Johnson, “How to Use t-SNE Effectively,” *Distill*, vol. 1, no. 10, p. e2, Jan. 2017, doi: 10.23915/distill.00002.

[88] H. Perez and J. H. M. Tah, “Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE,” *Mathematics*, vol. 8, no. 5, p. 662, May 2020, doi: 10.3390/MATH8050662.

[89] D. Agis and F. Pozo, “A frequency-based approach for the detection and classification of structural changes using t-SNE†,” *Sensors (Switzerland)*, vol. 19, no. 23, p. 5097, Dec. 2019, doi: 10.3390/s19235097.

[90] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” 2008. Accessed: Jun. 28, 2021. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html.

[91] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen, “Short Review of Dimensionality Reduction Methods Based on Stochastic Neighbour Embedding,” in *Advances in Intelligent Systems and Computing*, 2014, vol. 295, pp. 65–74, doi: 10.1007/978-3-319-07695-9_6.

[92] M. Verleysen and J. A. Lee, “Nonlinear dimensionality reduction for visualization,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8226 LNCS, no. PART 1, pp. 617–622, doi: 10.1007/978-3-642-42054-2_77.

[93] B. Melit Devassy and S. George, “Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE,” *Forensic Sci. Int.*, vol. 311, p. 110194, Jun. 2020, doi: 10.1016/j.forsciint.2020.110194.

[94] M. ai Li, X. yong Luo, and J. fu Yang, “Extracting the nonlinear features of motor imagery EEG using parametric t-SNE,” *Neurocomputing*, vol. 218, pp. 371–381, Dec. 2016, doi: 10.1016/j.neucom.2016.08.083.

[95] K. K. Varadarajan, P. R. Suhasini, K. Manikantan, and S. Ramachandran, “Face recognition using block based feature extraction with CZT and Goertzel-algorithm as a preprocessing technique,” in *Procedia Computer*

Science, Jan. 2015, vol. 46, pp. 1458–1467, doi: 10.1016/j.procs.2015.02.065.

[96] G. Traven *et al.*, “The Galah Survey: Classification and Diagnostics with t-SNE Reduction of Spectral Information,” *Astrophys. J. Suppl. Ser.*, vol. 228, no. 2, p. 24, Feb. 2017, doi: 10.3847/1538-4365/228/2/24.

[97] D. Kobak, G. Linderman, S. Steinerberger, Y. Kluger, and P. Berens, “Heavy-Tailed Kernels Reveal a Finer Cluster Structure in t-SNE Visualisations,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 11906 LNAI, pp. 124–139, doi: 10.1007/978-3-030-46150-8_8.

[98] M. Husnain, M. M. S. Missen, S. Mumtaz, M. M. Luqman, M. Coustaty, and J. M. Ogier, “Visualization of high-dimensional data by pairwise fusion matrices using t-SNE,” *Symmetry (Basel).*, vol. 11, no. 1, p. 107, Jan. 2019, doi: 10.3390/sym11010107.

[99] D. M. Chan, R. Rao, F. Huang, and J. F. Canny, “T-SNE-CUDA: GPU-Accelerated T-SNE and its Applications to Modern Data,” in *Proceedings - 2018 30th International Symposium on Computer Architecture and High Performance Computing, SBAC-PAD 2018*, Feb. 2019, pp. 330–338, doi: 10.1109/CAHPC.2018.8645912.

[100] W. Lai, M. Zhou, F. Hu, K. Bian, and Q. Song, “A New DBSCAN Parameters Determination Method Based on Improved MVO,” *IEEE Access*, vol. 7, pp. 104085–104095, 2019, doi: 10.1109/ACCESS.2019.2931334.

[101] A. Musdholifah, S. Zaiton, and M. Hashim, “Cluster Analysis on High-Dimensional Data: A Comparison of Density-based Clustering Algorithms,” *Aust. J. Basic Appl. Sci.*, vol. 7, no. 2, pp. 380–389, 2013.

[102] M. Gaonkar and K. Sawant, “AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Dataset,” *Int. J. Adv. Comput. Theory Eng.*, 2013.

[103] K. Khan, S. U. Rehman, K. Aziz, S. Fong, S. Sarasvady, and A. Vishwa, “DBSCAN: Past, present and future,” *5th Int. Conf. Appl. Digit. Inf. Web Technol. ICADIWT 2014*, pp. 232–238, 2014, doi: 10.1109/ICADIWT.2014.6814687.

[104] M. H. Jeong, Y. Cai, C. J. Sullivan, and S. Wang, “Data depth based clustering analysis,” *GIS Proc. ACM Int. Symp. Adv. Geogr. Inf. Syst.*, Oct. 2016, doi: 10.1145/2996913.2996984.

[105] T. Boonchoo, X. Ao, Y. Liu, W. Zhao, F. Zhuang, and Q. He, “Grid-based DBSCAN: Indexing and inference,” *Pattern Recognit.*, vol. 90, pp. 271–284, Jun. 2019, doi: 10.1016/J.PATCOG.2019.01.034.

[106] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 1996, Accessed: Jul. 11, 2021. [Online]. Available: www.aaai.org.

[107] I. K. Savvas and D. Tselios, “Parallelizing DBSCaN algorithm using MPI,” *Proc. - 25th IEEE Int. Conf. Enabling Technol. Infrastruct. Collab. Enterp. WETICE 2016*, pp. 77–82, Aug. 2016, doi: 10.1109/WETICE.2016.26.

[108] G. H. Shah, “An improved DBSCAN, a density based clustering algorithm with parameter selection for high dimensional data sets,” *3rd Nirma Univ. Int. Conf. Eng. NUiCONE 2012*, 2012, doi: 10.1109/NUICONE.2012.6493211.

[109] B. Borah and D. K. Bhattacharyya, “An Improved Sampling-Based DBSCAN for Large Spatial Databases,” *Proc. Int. Conf. Intell. Sens. Inf. Process. ICISIP 2004*, pp. 92–96, 2004, doi: 10.1109/ICISIP.2004.1287631.

[110] M. Kryszkiewicz and P. Lasek, “TI-DBSCAN: Clustering with DBSCAN by means of the triangle inequality,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6086 LNAI, pp. 60–69, 2010, doi: 10.1007/978-3-642-13529-3_8.

[111] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN,” *ACM Trans. Database Syst.*, vol. 42, no. 3, Jul. 2017, doi: 10.1145/3068335.

[112] I. Mohamad, U. Johor Bahru, J. Darul Ta, I. Bin Mohamad, D. Usman, and J. Bahru, “Standardization and Its Effects on K-Means Clustering Algorithm,” *Artic. Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 17, pp. 3299–3303, 2013, doi: 10.19026/rjaset.6.3638.

[113] T. Wu, A. R. Benson, and D. F. Gleich, “General Tensor Spectral Co-clustering for Higher-Order Data,” *Adv. Neural Inf. Process. Syst.*, pp. 2567–2575, Mar. 2016, Accessed: Jun. 16, 2021. [Online]. Available: http://arxiv.org/abs/1603.00395.

[114] P. M. Bhagat, P. S. Halgaonkar, and V. M. Wadhai, “Review of Clustering Algorithm for Categorical Data,” *Int. J. Eng. Adv. Technol.*, vol. 3, no. 2, 2013.

[115] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 89–98, doi: 10.1145/956750.956764.

[116] J. Zhao and S. Conrad, “Subspace clustering with distance-density function and entropy in high-dimensional data,” in *DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications*, 2013, pp. 14–22, doi: 10.5220/0004486600140022.

[117] R. Xu and D. C. Wunsch II, “Cluster Analysis,” in *Clustering*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008.

[118] L. Rokach and O. Maimon, “Clustering Methods,” in *Data Mining and Knowledge Discovery Handbook*, Springer-Verlag, 2006, pp. 321–352.

[119] S. Papadimitriou, “DisCo: Distributed Co-clustering with Map-Reduce: A

case study towards petabyte-scale end-to-end mining," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2008, pp. 512–521, doi: 10.1109/ICDM.2008.142.

[120] S. Huang, H. Wang, D. Li, Y. Yang, and T. Li, "Spectral co-clustering ensemble," *Knowledge-Based Syst.*, vol. 84, pp. 46–55, Aug. 2015, doi: 10.1016/j.knosys.2015.03.027.

[121] A. Banerjee, J. Ghosh, S. Merugu, and D. S. Modha, "A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation," 2007. Accessed: Jun. 17, 2021. [Online]. Available: http://jmlr.org/papers/v8/banerjee07a.html.

[122] G. Bisson and F. Hussain, "χ-Sim: A new similarity measure for the co-clustering task," in *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008*, 2008, pp. 211–217, doi: 10.1109/ICMLA.2008.103.

[123] F. Wang, S. Lin, and P. S. Yu, "Collaborative co-clustering across multiple social media," in *Proceedings - IEEE International Conference on Mobile Data Management*, Jul. 2016, vol. 2016-July, pp. 142–151, doi: 10.1109/MDM.2016.31.

[124] E. Hoseini, S. Hashemi, and A. Hamzeh, "Link prediction in social network using co-clustering based approach," in *Proceedings - 26th IEEE International Conference on Advanced Information Networking and Applications Workshops, WAINA 2012*, 2012, pp. 795–800, doi: 10.1109/WAINA.2012.189.

[125] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognit.*, vol. 41, no. 1, pp. 176–190, Jan. 2008, doi: 10.1016/j.patcog.2007.05.018.

[126] M. Nasiri, B. Minaei, and Z. Sharifi, "Adjusting data sparsity problem using linear algebra and machine learning algorithm," *Appl. Soft Comput. J.*, vol.

61, pp. 1153–1159, Dec. 2017, doi: 10.1016/j.asoc.2017.05.042.

[127] C. Federer, M. Yoo, and A. C. Tan, "Big Data Mining and Adverse Event Pattern Analysis in Clinical Drug Trials," *Assay Drug Dev. Technol.*, vol. 14, no. 10, pp. 557–566, Dec. 2016, doi: 10.1089/ADT.2016.742/ASSET/IMAGES/MEDIUM/FIGURE1.GIF.

[128] J. Luo and R. A. Cisler, "Discovering Outliers of Potential Drug Toxicities Using a Large-scale Data-driven Approach," *Cancer Inform.*, vol. 15, p. 211, Oct. 2016, doi: 10.4137/CIN.S39549.

[129] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, Jul. 2018, doi: 10.1093/BIOINFORMATICS/BTY294.

[130] "Advanced Search - ClinicalTrials.gov." https://clinicaltrials.gov/ct2/search/advanced (accessed Aug. 25, 2021).

[131] "Download MeSH Data." https://www.nlm.nih.gov/databases/download/mesh.html (accessed Aug. 25, 2021).

[132] "MeSH Browser." https://meshb.nlm.nih.gov/record/ui?ui=D001416 (accessed Aug. 25, 2021).

[133] "Common Terminology Criteria for Adverse Events (CTCAE) | Protocol Development | CTEP." https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm (accessed Aug. 25, 2021).

[134] W. McKinney, "{D}ata {S}tructures for {S}tatistical {C}omputing in {P}ython," in *{P}roceedings of the 9th {P}ython in {S}cience {C}onference*, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[135] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media Inc., 2009.

[136] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Aug. 25, 2021. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[137] B. Zhao, H. Zhao, and J. Zhao, “Serious adverse events and fatal adverse events associated with nivolumab treatment in cancer patients,” *J. Immunother. Cancer*, vol. 6, no. 1, p. 101, Dec. 2018, doi: 10.1186/S40425-018-0421-Z.

[138] L. Grassi *et al.*, “Role of Psychosocial Variables on Chemotherapy-Induced Nausea and Vomiting and Health-Related Quality of Life among Cancer Patients: A European Study,” *Psychother. Psychosom.*, vol. 84, no. 6, pp. 339–347, Oct. 2015, doi: 10.1159/000431256.

[139] D. G. Harris, “Nausea and vomiting in advanced cancer,” *Br. Med. Bull.*, vol. 96, no. 1, pp. 175–185, Dec. 2010, doi: 10.1093/BMB/LDQ031.

[140] G. J. Sanger and P. L. R. Andrews, “Treatment of nausea and vomiting: Gaps in our knowledge,” *Auton. Neurosci.*, vol. 129, no. 1–2, pp. 3–16, Oct. 2006, doi: 10.1016/J.AUTNEU.2006.07.009.

[141] S. C and T. L, “Treatment-related diarrhea in patients with cancer,” *Clin. J. Oncol. Nurs.*, vol. 16, no. 4, pp. 413–417, 2012, doi: 10.1188/12.CJON.413-417.

[142] G. Santucci and J. W. Mack, “Common Gastrointestinal Symptoms in Pediatric Palliative Care: Nausea, Vomiting, Constipation, Anorexia, Cachexia,” *Pediatr Clin N Am*, vol. 54, pp. 673–689, 2007, doi: 10.1016/j.pcl.2007.06.001.

[143] F. A. van der Toorn, R. de Mutsert, W. M. Lijfering, F. R. Rosendaal, and A. van H. Vlieg, “Glucose metabolism affects coagulation factors: The NEO study,” *J. Thromb. Haemost.*, vol. 17, no. 11, pp. 1886–1897, Nov. 2019, doi: 10.1111/JTH.14573.

[144] R. Yang and L. Moosavi, “Prothrombin Time,” *Transfus. Med. Hemost. Clin. Lab. Asp. Second Ed.*, pp. 799–803, May 2021, Accessed: Aug. 16, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK544269/.

[145] M. Paar *et al.*, “Anticoagulant action of low, physiologic, and high albumin levels in whole blood,” *PLoS One*, vol. 12, no. 8, Aug. 2017, doi: 10.1371/JOURNAL.PONE.0182997.

[146] M. Hirashima *et al.*, “High-level expression and preparation of recombinant human fibrinogen as biopharmaceuticals,” *J. Biochem.*, vol. 159, no. 2, p. 261, Jun. 2016, doi: 10.1093/JB/MVV099.

[147] S. C. Larsson *et al.*, “Serum magnesium and calcium levels in relation to ischemic stroke: Mendelian randomization study,” *Neurology*, vol. 92, no. 9, p. e944, Feb. 2019, doi: 10.1212/WNL.0000000000007001.

[148] N. Sevastos, G. Theodossiades, and A. J. Archimandritis, “Pseudohyperkalemia in Serum: A New Insight into an Old Phenomenon,” *Clin. Med. Res.*, vol. 6, no. 1, p. 30, May 2008, doi: 10.3121/CMR.2008.739.

[149] S. A. Maleknia and N. Ebrahimi, “Evaluation of Liver Function Tests and Serum Bilirubin Levels After Laparoscopic Cholecystectomy,” *Med. Arch.*, vol. 74, no. 1, p. 24, Feb. 2020, doi: 10.5455/MEDARH.2020.74.24-27.

[150] J. M. Roscoe, M. L. Halperin, F. S. Rolleston, and M. B. Goldstein, “Hyperglycemia-induced hyponatremia: metabolic considerations in calculation of serum sodium depression.,” *Can. Med. Assoc. J.*, vol. 112, no. 4, p. 452, 1975, Accessed: Aug. 24, 2021. [Online]. Available: /pmc/articles/PMC1956157/?report=abstract.

[151] L.-S. Y, D. D, V. I, K. B, S. D, and M.-M. D, “Hyponatremia and decreased bone density in adolescent inpatients diagnosed with anorexia nervosa,” *Nutrition*, vol. 32, no. 10, pp. 1097–1102, Oct. 2016, doi: 10.1016/J.NUT.2016.03.015.

[152] M. Kulkarni and A. Bhat, “Asymptomatic hyponatremia: is it time to abandon this entity?,” *J. Nephropharmacology*, vol. 4, no. 2, p. 78, 2015, Accessed: Aug. 24, 2021. [Online]. Available: /pmc/articles/PMC5297491/.

[153] D. Assimiti, “The Use of Beetroot as Natural Solutions for Reducing Inflammation - Case Studies from Thailand (P12-046-19),” *Curr. Dev. Nutr.*, vol. 3, no. Suppl 1, Jun. 2019, doi: 10.1093/CDN/NZZ035.P12-046-19.

[154] K. Patel and H. (Jack) West, “Febrile Neutropenia,” *JAMA Oncol.*, vol. 3, no. 12, pp. 1751–1751, Dec. 2017, doi: 10.1001/JAMAONCOL.2017.1114.

[155] S. E. Evans and D. E. Ost, “Pneumonia in the neutropenic cancer patient,” *Curr. Opin. Pulm. Med.*, vol. 21, no. 3, p. 260, May 2015, doi: 10.1097/MCP.0000000000000156.

[156] D. Berliner, N. Schneider, T. Welte, and J. Bauersachs, “The Differential Diagnosis of Dyspnea,” *Dtsch. Arztebl. Int.*, vol. 113, no. 49, p. 834, Dec. 2016, doi: 10.3238/ARZTEBL.2016.0834.

[157] S. RA, “Trigeminal neuralgia associated with sinusitis,” *ORL. J. Otorhinolaryngol. Relat. Spec.*, vol. 62, no. 3, pp. 160–163, 2000, doi: 10.1159/000027738.

[158] G. Cho, J. Yim, Y. Choi, J. Ko, and S.-H. Lee, “Review of Machine Learning Algorithms for Diagnosing Mental Illness,” *Psychiatry Investig.*, vol. 16, no. 4, pp. 262–269, Apr. 2019, doi: 10.30773/PI.2018.12.21.2.

[159] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records,” *Sci. Reports 2016 61*, vol. 6, no. 1, pp. 1–10, May 2016, doi: 10.1038/srep26094.

[160] D. J. Pinato *et al.*, “Integrated analysis of multiple receptor tyrosine kinases identifies Axl as a therapeutic target and mediator of resistance to sorafenib

in hepatocellular carcinoma," *Br. J. Cancer*, vol. 120, no. 5, pp. 512–521, Mar. 2019, doi: 10.1038/S41416-018-0373-6.

[161] W. AL-Busairi and M. Khajah, "The Principles behind Targeted Therapy for Cancer Treatment," *Tumor Progress. Metastasis*, Jun. 2019, doi: 10.5772/INTECHOPEN.86729.

[162] M. Y, Y. O, T. I, K. S, and B. B, "Antibacterial activity of adriamycin against bacillus Calmette-Guerin," *Oncol. Rep.*, vol. 4, no. 5, pp. 909–911, 1997, doi: 10.3892/OR.4.5.909.

[163] P. P, T. AC, S. S, K. MM, D. A, and B. TK, "Doxorubicin inhibits E. coli division by interacting at a novel site in FtsZ," *Biochem. J.*, vol. 471, no. 3, pp. 335–346, Nov. 2015, doi: 10.1042/BJ20150467.

[164] F. T. Fraunfelder and F. W. Fraunfelder, "Oral Anti-Vascular Endothelial Growth Factor Drugs and Ocular Side Effects," doi: 10.1089/jop.2018.0019.

[165] M. NS and L. AE, "Acute leukemia with a very high leukocyte count: confronting a medical emergency," *Cleve. Clin. J. Med.*, vol. 71, no. 8, pp. 633–637, 2004, doi: 10.3949/CCJM.71.8.633.

[166] Y. R. Kim *et al.*, "Lymphopenia is an important prognostic factor in peripheral T-cell lymphoma (NOS) treated with anthracycline-containing chemotherapy," *J. Hematol. Oncol.*, vol. 4, p. 34, 2011, doi: 10.1186/1756-8722-4-34.

[167] nlm.nih.gov, "DRUG LABEL INFORMATION, LABEL: CYTARABINE injection." https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=34803a0e-f54d-4147-8f9f-7d0e3a007756 (accessed Aug. 20, 2021).

[168] L. Wang *et al.*, "Does High-Dose Cytarabine Cause More Fungal Infection in Patients With Acute Myeloid Leukemia Undergoing Consolidation Therapy: A Multicenter, Prospective, Observational Study in China,"

Medicine (Baltimore)., vol. 95, no. 4, 2016, doi: 10.1097/MD.0000000000002560.

[169] L. Di Cesare Mannelli, M. Maresca, C. Farina, M. W. Scherz, and C. Ghelardini, “A model of neuropathic pain induced by sorafenib in the rat: Effect of dimiracetam,” *Neurotoxicology*, vol. 50, pp. 101–107, Sep. 2015, doi: 10.1016/J.NEURO.2015.08.002.

[170] K. K, R. MH, and O. M, “Creatinine: From physiology to clinical application,” *Eur. J. Intern. Med.*, vol. 72, pp. 9–14, Feb. 2020, doi: 10.1016/J.EJIM.2019.10.025.

[171] T. Rana, A. Chakrabarti, M. Freeman, and S. Biswas, “Doxorubicin-Mediated Bone Loss in Breast Cancer Bone Metastases Is Driven by an Interplay between Oxidative Stress and Induction of TGFβ,” *PLoS One*, vol. 8, no. 10, p. e78043, 2013, doi: 10.1371/JOURNAL.PONE.0078043.

[172] L. Zhou, F. Kuai, Q. Shi, and H. Yang, “Doxorubicin restrains osteogenesis and promotes osteoclastogenesis in vitro,” *Am. J. Transl. Res.*, vol. 12, no. 9, p. 5640, 2020, Accessed: Aug. 24, 2021. [Online]. Available: /pmc/articles/PMC7540161/.

[173] D. D, B.-B. I, F. P, S. N, and R. B, “T-cell acute lymphoblastic leukemia with severe leukopenia: evidence for suppression of myeloid progenitor cells by leukemic blasts,” *Acta Haematol.*, vol. 80, no. 4, pp. 185–189, 1988, doi: 10.1159/000205634.

[174] C. Siebenaller *et al.*, “Hospital readmission rate for febrile neutropenia (FN) following high dose cytarabine (HiDAC) consolidation chemotherapy for acute myeloid leukemia (AML).,” *https://doi.org/10.1200/JCO.2017.35.15_suppl.e18513*, vol. 35, no. 15_suppl, pp. e18513–e18513, May 2017, doi: 10.1200/JCO.2017.35.15_SUPPL.E18513.

[175] L. H *et al.*, “Doxorubicin contributes to thrombus formation and vascular

injury by interfering with platelet function," *Am. J. Physiol. Heart Circ. Physiol.*, vol. 319, no. 1, pp. H1333–H143, Jul. 2020, doi: 10.1152/AJPHEART.00456.2019.

[176] K. EJ *et al.*, "Doxorubicin-induced platelet cytotoxicity: a new contributory factor for doxorubicin-mediated thrombocytopenia," *J. Thromb. Haemost.*, vol. 7, no. 7, pp. 1172–1183, 2009, doi: 10.1111/J.1538-7836.2009.03477.X.

[177] C. Robier, "Platelet morphology," *J. Lab. Med.*, vol. 44, no. 5, pp. 231–239, Oct. 2020, doi: 10.1515/LABMED-2020-0007.

[178] B. WJ, R. GL, and W. RB, "Cytarabine and neurologic toxicity," *J. Clin. Oncol.*, vol. 9, no. 4, pp. 679–693, 1991, doi: 10.1200/JCO.1991.9.4.679.

[179] L. Barbier, F. Muscari, S. Le Guellec, A. Pariente, P. Otal, and B. Suc, "Liver Resection after Downstaging Hepatocellular Carcinoma with Sorafenib," *Int. J. Hepatol.*, vol. 2011, pp. 1–5, 2011, doi: 10.4061/2011/791013.

[180] K. Imoto *et al.*, "Successful endoscopic treatment of hepatoduodenal fistula formed during sorafenib treatment for hepatocellular carcinoma with duodenal invasion," *Acta Hepatol. Jpn.*, vol. 60, no. 3, pp. 91–98, 2019, doi: 10.2957/KANZO.60.91.

[181] E. Song, K. M. Song, W. G. Kim, and C. M. Choi, "Development of Tracheoesophageal Fistula after the Use of Sorafenib in Locally Advanced Papillary Thyroid Carcinoma: a Case Report," *Int. J. Thyroidol.*, vol. 9, no. 2, pp. 210–214, Nov. 2016, doi: 10.11106/IJT.2016.9.2.210.

[182] M. E. L. Alaminos *et al.*, "Skin fistula after sorafenib use in differentiated thyroid cancer," *Endocr. Abstr.*, vol. 63, May 2019, doi: 10.1530/ENDOABS.63.P1224.

[183] F. Rossetti, S. Cesaro, M. C. Putti, and L. Zanesco, "High-Dose Cytosine Arabinoside and Viridans Streptococcus Sepsis in Children with Leukemia,"

http://dx.doi.org/10.3109/08880019509029589, vol. 12, no. 4, pp. 387–392, 2009, doi: 10.3109/08880019509029589.

[184] J. E. Cortes *et al.*, "Randomized comparison of low dose cytarabine with or without glasdegib in patients with newly diagnosed acute myeloid leukemia or high-risk myelodysplastic syndrome," *Leuk. 2018 332*, vol. 33, no. 2, pp. 379–389, Dec. 2018, doi: 10.1038/s41375-018-0312-9.

[185] L. Zhao and B. Zhang, "Doxorubicin induces cardiotoxicity through upregulation of death receptors mediated apoptosis in cardiomyocytes," *Sci. Reports 2017 71*, vol. 7, no. 1, pp. 1–11, Mar. 2017, doi: 10.1038/srep44735.

APPENDICES

A. Numbers of Root Events in the Drug Clusters

Table A.1 Numbers of all root events in the drug clusters

Root Events	Cytarabine	Sorafenib	Doxorubicin
Amino Acids, Peptides, and Proteins	5.0	15.0	2.0
Animal Diseases	1.0	6.0	0.0
Animal Structures	1.0	1.0	0.0
Bacteria	8.0	11.0	0.0
Behavior and Behavior Mechanisms	10.0	22.0	6.0
Biological Factors	1.0	6.0	0.0
Biological Phenomena	1.0	2.0	1.0
Body Regions	3.0	9.0	1.0
Carbohydrates	2.0	2.0	1.0
Cardiovascular Diseases	29.0	88.0	20.0
Cardiovascular System	1.0	14.0	2.0
Cells	2.0	3.0	4.0
Chemical Actions and Uses	1.0	9.0	1.0
Chemically-Induced Disorders	5.0	10.0	2.0
Complex Mixtures	2.0	1.0	0.0
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	6.0	24.0	6.0
Dentistry	1.0	2.0	0.0
Diagnosis	3.0	18.0	7.0
Digestive System Diseases	27.0	114.0	41.0
Endocrine System Diseases	4.0	30.0	6.0
Environment and Public Health	3.0	10.0	5.0
Enzymes and Coenzymes	4.0	8.0	2.0

Equipment and Supplies	1.0	2.0	2.0
Eukaryota	1.0	13.0	5.0
Eye Diseases	5.0	42.0	4.0
Female Urogenital Diseases and Pregnancy Complications	10.0	38.0	12.0
Fluids and Secretions	1.0	2.0	0.0
Genetic Phenomena	1.0	2.0	0.0
Health Care Facilities, Manpower, and Services	1.0	3.0	5.0
Health Care Quality, Access, and Evaluation	1.0	3.0	3.0
Health Occupations	1.0	3.0	0.0
Health Services Administration	1.0	3.0	0.0
Hemic and Lymphatic Diseases	24.0	39.0	15.0
Heterocyclic Compounds	2.0	8.0	8.0
Immune System Diseases	11.0	29.0	9.0
Immune System Phenomena	1.0	2.0	0.0
Infections	53.0	114.0	37.0
Inorganic Chemicals	3.0	3.0	0.0
Investigative Techniques	14.0	27.0	14.0
Male Urogenital Diseases	11.0	29.0	12.0
Mental Disorders	2.0	15.0	2.0
Metabolism	1.0	0.0	0.0
Musculoskeletal Diseases	11.0	44.0	6.0
Musculoskeletal and Neural Physiological Phenomena	1.0	8.0	5.0
Neoplasms	13.0	66.0	16.0
Nervous System Diseases	32.0	149.0	29.0
Nucleic Acids, Nucleotides, and Nucleosides	1.0	1.0	0.0
Nutritional and Metabolic Diseases	16.0	23.0	7.0
Organic Chemicals	3.0	9.0	7.0
Otorhinolaryngologic Diseases	7.0	23.0	7.0

Pathological Conditions, Signs and Symptoms	57.0	187.0	51.0
Physical Phenomena	2.0	3.0	1.0
Population Characteristics	1.0	2.0	2.0
Psychological Phenomena	3.0	6.0	4.0
Reproductive and Urinary Physiological Phenomena	3.0	4.0	0.0
Reproductive system and breast disorders	1.0	5.0	3.0
Respiratory System	1.0	2.0	0.0
Respiratory Tract Diseases	22.0	64.0	21.0
Skin and Connective Tissue Diseases	9.0	54.0	10.0
Social Sciences	1.0	6.0	2.0
Stomatognathic Diseases	6.0	22.0	5.0
Stomatognathic System	2.0	1.0	1.0
Surgical Procedures, Operative	3.0	21.0	3.0
Technology, Industry, and Agriculture	1.0	1.0	1.0
Therapeutics	6.0	18.0	3.0
Tissues	1.0	8.0	1.0
Viruses	1.0	5.0	3.0
Wounds and Injuries	8.0	43.0	17.0
Anesthesia and Analgesia	0.0	1.0	1.0
Archaea	0.0	1.0	0.0
Biomedical and Dental Materials	0.0	1.0	1.0
Cell Physiological Phenomena	0.0	2.0	1.0
Education	0.0	1.0	0.0
Embryonic Structures	0.0	1.0	0.0
Endocrine System	0.0	1.0	0.0
Geographic Locations	0.0	2.0	0.0
Health Care Economics and Organizations	0.0	3.0	0.0
Hormones, Hormone Substitutes, and Hormone Antagonists	0.0	3.0	0.0

Human Activities	0.0	1.0	1.0
Humanities	0.0	1.0	0.0
Information Science	0.0	4.0	0.0
Integumentary System	0.0	2.0	0.0
Macromolecular Substances	0.0	1.0	1.0
Microbiological Phenomena	0.0	2.0	0.0
Musculoskeletal System	0.0	12.0	4.0
Non-Medical Public and Private Facilities	0.0	2.0	1.0
Persons	0.0	2.0	2.0
Pharmaceutical Preparations	0.0	1.0	0.0
Polycyclic Compounds	0.0	3.0	2.0
Pregnancy, puerperium and perinatal conditions	0.0	1.0	0.0
Sense Organs	0.0	1.0	0.0
Surgical and medical procedures	0.0	1.0	0.0
Urogenital System	0.0	4.0	2.0
Integumentary System Physiological Phenomena	0.0	0.0	1.0
Lipids	0.0	0.0	1.0

B. Rates of all the Root Events

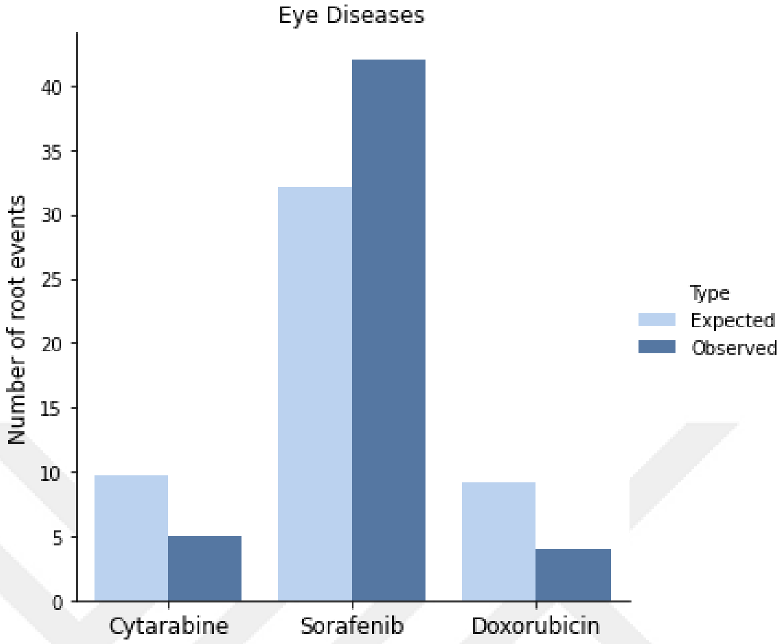


Figure B.1. The expected and observed rates of the “Eye Diseases” root event.

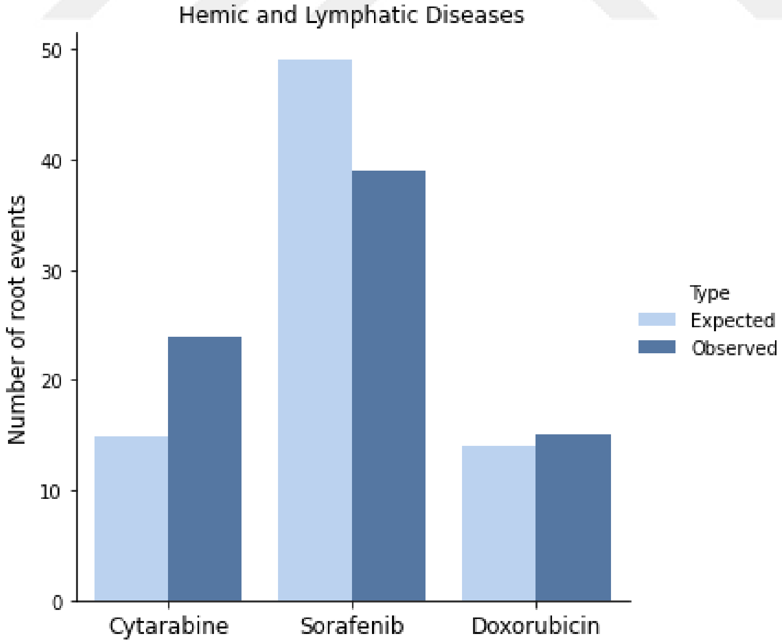


Figure B.2. The expected and observed rates of the “Hemic and Lymphatic Diseases” root event.

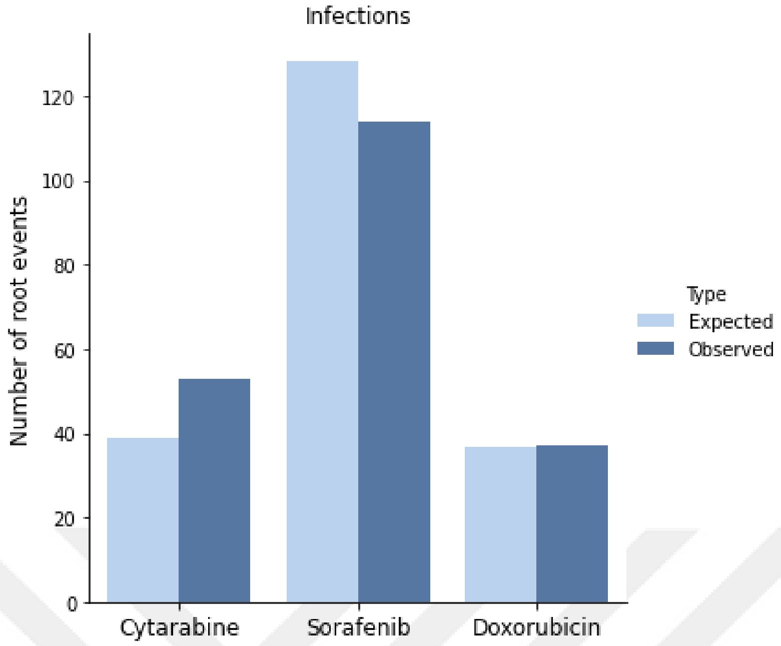


Figure B.3. The expected and observed rates of the “Infections” root event.

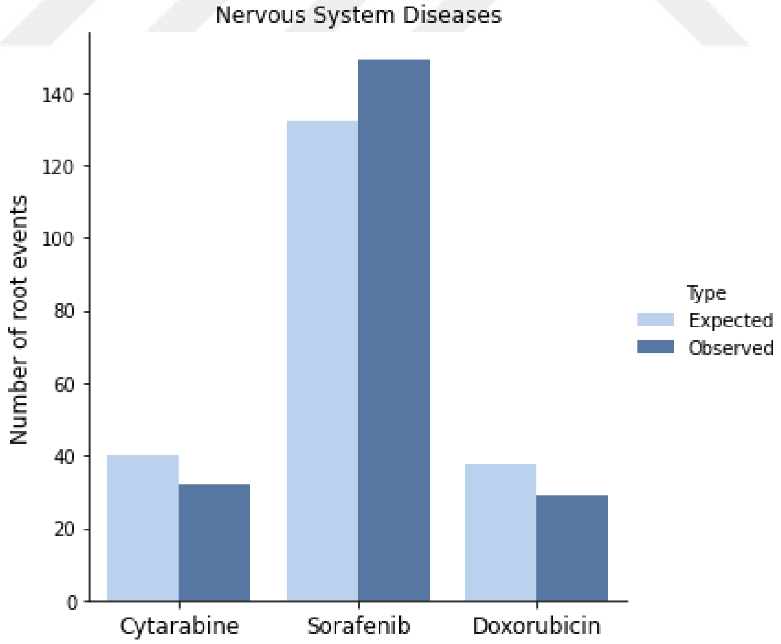


Figure B.4. The expected and observed rates of the “Nervous System Diseases” root event.

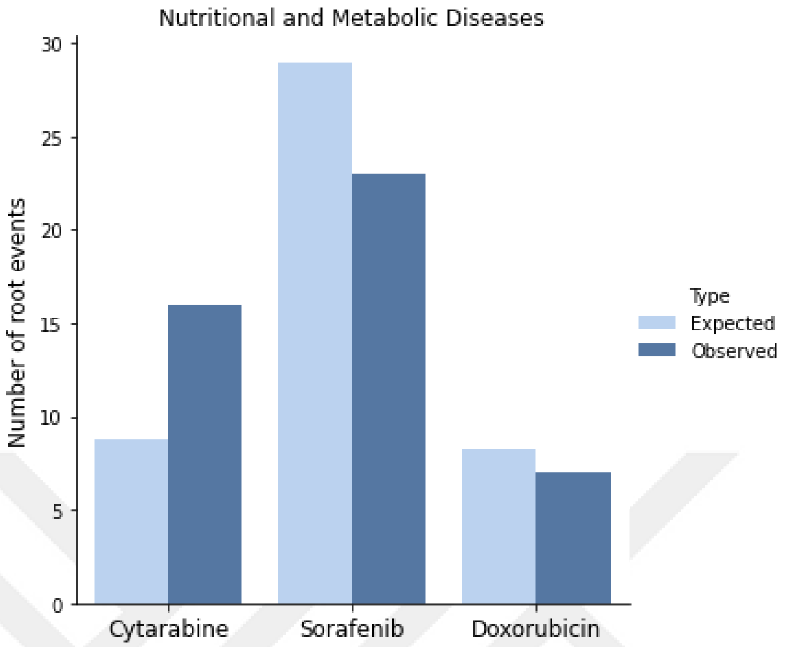


Figure B.5. The expected and observed rates of the “Nutritional and Metabolic Diseases” root event.

C. Numbers of Root Events in the Drug Clusters

Table C.1 Numbers of all parent events in the drug clusters

Parent Events	Cytarabine	Sorafenib	Doxorubicin
Adrenal Gland Diseases	1.0	5.0	1.0
Amines	1.0	3.0	0.0
Amino Acids	1.0	0.0	1.0
Antigen-Antibody Reactions	1.0	0.0	0.0
Arthritis, Infectious	1.0	0.0	0.0
Asthenopia	1.0	0.0	0.0
Autoimmune Diseases	2.0	8.0	2.0
Autoimmune Diseases of the Nervous System	1.0	1.0	1.0

Bacterial Infections and Mycoses	15.0	32.0	10.0
Behavior	6.0	8.0	2.0
Biliary Tract Diseases	1.0	17.0	2.0
Biological Products	1.0	1.0	0.0
Biotransformation	1.0	0.0	0.0
Bipolar and Related Disorders	1.0	0.0	0.0
Blood Physiological Phenomena	1.0	1.0	2.0
Blood Vessels	1.0	10.0	2.0
Body Constitution	1.0	1.0	2.0
Body Fluids	1.0	2.0	0.0
Bone Diseases	3.0	15.0	1.0
Bone Diseases, Infectious	1.0	4.0	0.0
Bronchial Diseases	2.0	5.0	2.0
Cardiovascular Abnormalities	1.0	2.0	0.0
Cardiovascular Physiological Phenomena	2.0	3.0	2.0
Central Nervous System Diseases	11.0	57.0	8.0
Central Nervous System Infections	1.0	14.0	0.0
Congenital Abnormalities	2.0	12.0	4.0
Connective Tissue Diseases	2.0	7.0	1.0
Corneal Diseases	1.0	3.0	0.0
Cranial Nerve Diseases	1.0	10.0	2.0
Culture Media	1.0	0.0	0.0
Delivery of Health Care	1.0	0.0	1.0
Demography	1.0	0.0	1.0
Demyelinating Diseases	1.0	3.0	2.0
Deoxy Sugars	1.0	0.0	0.0
Diagnosis, Oral	1.0	1.0	0.0
Diagnostic Techniques and Procedures	3.0	18.0	7.0
Diet, Food, and Nutrition	1.0	1.0	1.0

Digestive System Abnormalities	2.0	4.0	2.0
Digestive System Surgical Procedures	1.0	2.0	0.0
Drug Therapy	1.0	3.0	2.0
Drug-Related Side Effects and Adverse Reactions	2.0	7.0	2.0
Dwarfism	1.0	1.0	0.0
Ear Diseases	3.0	12.0	2.0
Ecological and Environmental Phenomena	1.0	1.0	0.0
Electrolytes	1.0	0.0	0.0
Elements	2.0	3.0	0.0
Endocrine Gland Neoplasms	2.0	6.0	0.0
Environment	1.0	2.0	0.0
Enzymes	4.0	7.0	2.0
Epidemiologic Methods	1.0	2.0	2.0
Exocrine Glands	1.0	0.0	0.0
Extremities	1.0	4.0	0.0
Fasciitis	1.0	0.0	0.0
Female Urogenital Diseases	10.0	35.0	11.0
Free Radicals	1.0	0.0	0.0
Fungal Viruses	1.0	0.0	0.0
Fungi	1.0	2.0	0.0
Gastrointestinal Diseases	23.0	70.0	32.0
Genetic Diseases, Inborn	3.0	8.0	3.0
Genetic Variation	1.0	0.0	0.0
Genital Diseases, Male	2.0	5.0	2.0
Genotype	1.0	0.0	0.0
Geological Phenomena	1.0	0.0	0.0
Gingivitis	2.0	0.0	0.0
Graft vs Host Disease	1.0	0.0	0.0
Gram-Negative Bacteria	7.0	7.0	0.0

Gram-Positive Bacteria	1.0	1.0	0.0
Growth and Development	1.0	2.0	0.0
Head	1.0	1.0	1.0
Health Services	1.0	3.0	2.0
Heart Diseases	21.0	39.0	11.0
Hematologic Diseases	17.0	19.0	14.0
Hemic and Lymphatic Diseases	1.0	0.0	0.0
Heterocyclic Compounds, 1-Ring	1.0	6.0	6.0
Heterocyclic Compounds, Fused-Ring	1.0	4.0	2.0
Hydrocarbons	1.0	5.0	3.0
Hypersensitivity	2.0	9.0	4.0
Immune System	4.0	4.0	2.0
Immunologic Deficiency Syndromes	1.0	1.0	0.0
Immunoproliferative Disorders	3.0	10.0	3.0
Infant, Newborn, Diseases	1.0	5.0	0.0
Infections	1.0	0.0	0.0
Infections and infestations	14.0	28.0	8.0
Injury, poisoning and procedural complications	5.0	14.0	2.0
Intraabdominal Infections	2.0	2.0	0.0
Intubation	1.0	0.0	0.0
Joint Diseases	4.0	10.0	3.0
Lacrimal Apparatus Diseases	1.0	0.0	1.0
Liver Diseases	1.0	15.0	4.0
Lung	1.0	0.0	0.0
Lung Diseases	11.0	21.0	9.0
Lymphatic Diseases	7.0	16.0	4.0
Medicine	1.0	3.0	0.0
Mental Processes	1.0	1.0	0.0
Metabolic Diseases	16.0	19.0	7.0
Metals	2.0	2.0	0.0

Mouth	2.0	1.0	1.0
Mouth Diseases	4.0	11.0	5.0
Musculoskeletal and connective tissue disorders	3.0	9.0	0.0
Myeloid Cells	1.0	0.0	1.0
Neoplasms	1.0	0.0	0.0
Neoplasms by Histologic Type	7.0	35.0	12.0
Neoplasms by Site	5.0	32.0	1.0
Neoplastic Processes	2.0	2.0	1.0
Nervous System Neoplasms	1.0	1.0	0.0
Nervous System Physiological Phenomena	1.0	4.0	4.0
Neurobehavioral Manifestations	3.0	8.0	1.0
Neurocognitive Disorders	1.0	5.0	0.0
Neurologic Manifestations	15.0	53.0	10.0
Neuromuscular Diseases	2.0	15.0	6.0
Neurotoxicity Syndromes	2.0	0.0	0.0
Nose Diseases	2.0	3.0	1.0
Nucleotides	1.0	0.0	0.0
Nutrition Therapy	1.0	0.0	0.0
Opportunistic Infections	1.0	0.0	0.0
Orbital Diseases	1.0	1.0	0.0
Oxygen Compounds	1.0	0.0	0.0
Parasitic Diseases	2.0	6.0	2.0
Paratuberculosis	1.0	0.0	0.0
Pathologic Processes	34.0	78.0	15.0
Patient Care	1.0	4.0	0.0
Patient Care Management	1.0	1.0	0.0
Peptides	1.0	2.0	0.0
Perianal Glands	1.0	0.0	0.0
Personality	1.0	1.0	0.0

Phagocytes	1.0	0.0	0.0
Pharmacokinetics	1.0	0.0	0.0
Pharmacological and Toxicological Phenomena	1.0	1.0	0.0
Pharyngeal Diseases	2.0	4.0	1.0
Pleural Diseases	1.0	5.0	1.0
Poisoning	2.0	2.0	0.0
Proteins	4.0	14.0	1.0
Proteobacteria	7.0	8.0	0.0
Psychological Theory	1.0	0.0	0.0
Psychophysiology	1.0	4.0	4.0
Public Health	2.0	8.0	5.0
Radiation	1.0	2.0	1.0
Radiation Injuries	1.0	2.0	1.0
Rehabilitation	1.0	2.0	0.0
Reproductive Physiological Phenomena	2.0	3.0	0.0
Reproductive system and breast disorders	1.0	5.0	3.0
Respiration Disorders	7.0	22.0	6.0
Respiratory Tract Infections	8.0	15.0	8.0
Rheumatic Diseases	2.0	4.0	0.0
Sepsis	3.0	1.0	1.0
Signs and Symptoms	25.0	100.0	35.0
Skin Diseases	11.0	49.0	10.0
Skin Diseases, Infectious	1.0	4.0	2.0
Sociology	1.0	5.0	1.0
Soft Tissue Infections	1.0	0.0	0.0
Soil	1.0	0.0	0.0
Specialty Uses of Chemicals	1.0	2.0	0.0
Substance-Related Disorders	1.0	2.0	0.0
Sugars	1.0	1.0	1.0

Sulfur Compounds	1.0	0.0	5.0
Suppuration	2.0	7.0	2.0
Thyroid Diseases	2.0	6.0	1.0
Torso	1.0	3.0	0.0
Toxins, Biological	1.0	0.0	0.0
Tracheal Diseases	1.0	1.0	0.0
Transfusion Reaction	1.0	0.0	0.0
Transportation	1.0	0.0	0.0
Trauma, Nervous System	2.0	5.0	1.0
Urinary Tract Physiological Phenomena	1.0	1.0	0.0
Urogenital Neoplasms	1.0	5.0	0.0
Urologic Diseases	9.0	23.0	9.0
Vascular Diseases	11.0	53.0	11.0
Virus Diseases	5.0	19.0	7.0
Vision Disorders	1.0	3.0	1.0
Wounds, Nonpenetrating	1.0	3.0	0.0
Abdominal Injuries	0.0	1.0	1.0
Ablation Techniques	0.0	2.0	0.0
Adaptation, Psychological	0.0	1.0	0.0
Air Sacs	0.0	1.0	0.0
Alcohols	0.0	2.0	0.0
Aldehydes	0.0	1.0	0.0
Alkaloids	0.0	1.0	1.0
Americas	0.0	1.0	0.0
Amidines	0.0	1.0	0.0
Anesthesia	0.0	1.0	1.0
Animals	0.0	5.0	2.0
Ankyloglossia	0.0	1.0	0.0
Anthropology	0.0	1.0	1.0
Antigen-Presenting Cells	0.0	1.0	0.0

Antigens	0.0	1.0	0.0
Asphyxia	0.0	1.0	0.0
Autonomic Nervous System Diseases	0.0	3.0	1.0
Back Injuries	0.0	1.0	1.0
Bacterial Physiological Phenomena	0.0	1.0	0.0
Behavioral Sciences	0.0	1.0	0.0
Biliary Tract	0.0	1.0	1.0
Biological Therapy	0.0	3.0	1.0
Biopsy	0.0	2.0	0.0
Blood Coagulation Factors	0.0	1.0	0.0
Body Temperature	0.0	1.0	1.0
Breast	0.0	1.0	0.0
Cardiovascular Infections	0.0	2.0	0.0
Cardiovascular Surgical Procedures	0.0	3.0	1.0
Cartilage	0.0	2.0	1.0
Cartilage Diseases	0.0	1.0	0.0
Cat Diseases	0.0	1.0	0.0
Catheterization	0.0	2.0	0.0
Cattle Diseases	0.0	1.0	0.0
Cautery	0.0	1.0	0.0
Cell Count	0.0	2.0	1.0
Cellular Structures	0.0	1.0	1.0
Central Nervous System	0.0	4.0	1.0
Clinical Laboratory Techniques	12.0	22.0	9.0
Clinical Protocols	0.0	1.0	0.0
Communicable Diseases	0.0	1.0	0.0
Communication	0.0	1.0	0.0
Conjunctival Diseases	0.0	2.0	0.0
Connective Tissue	0.0	3.0	1.0
Constriction	0.0	1.0	0.0

Cross Infection	0.0	2.0	0.0
Curriculum	0.0	1.0	0.0
Cysts	0.0	6.0	0.0
Cytological Techniques	0.0	3.0	1.0
DNA Viruses	0.0	2.0	2.0
Dental Health Surveys	0.0	1.0	0.0
Diabetes Mellitus	0.0	8.0	2.0
Digestive System Diseases	0.0	1.0	0.0
Digestive System Fistula	0.0	3.0	0.0
Digestive System Neoplasms	0.0	9.0	1.0
Ear	0.0	1.0	0.0
Economics	0.0	1.0	0.0
Emotions	0.0	2.0	2.0
Endocrine Surgical Procedures	0.0	2.0	0.0
Endospore-Forming Bacteria	0.0	1.0	0.0
Enzyme Precursors	0.0	1.0	0.0
Epithelial Cells	0.0	1.0	0.0
Europe	0.0	1.0	0.0
Euryarchaeota	0.0	1.0	0.0
Evaluation Studies as Topic	0.0	1.0	0.0
Eye Abnormalities	0.0	8.0	2.0
Eye Diseases	0.0	1.0	0.0
Eye Infections	0.0	5.0	0.0
Eye Injuries	0.0	1.0	0.0
Eye Manifestations	0.0	1.0	0.0
Firmicutes	0.0	3.0	0.0
Fish Diseases	0.0	1.0	0.0
Fractures, Bone	0.0	7.0	8.0
Fused-Ring Compounds	0.0	1.0	2.0
Genetic Structures	0.0	2.0	0.0

Genitalia	0.0	2.0	1.0
Glymphatic System	0.0	1.0	0.0
Gonadal Disorders	0.0	5.0	0.0
Grandparents	0.0	1.0	0.0
Gubernaculum	0.0	1.0	0.0
Health	0.0	1.0	1.0
Heart	0.0	3.0	0.0
History	0.0	1.0	0.0
Hormones	0.0	3.0	0.0
Housing	0.0	1.0	0.0
Immune System Diseases	0.0	1.0	0.0
Immunocompetence	0.0	1.0	0.0
Immunologic Techniques	0.0	2.0	0.0
Informatics	0.0	1.0	0.0
Information Centers	0.0	1.0	0.0
Jaw Diseases	0.0	2.0	0.0
Joint Dislocations	0.0	1.0	0.0
Lacerations	0.0	1.0	0.0
Laryngeal Diseases	0.0	4.0	3.0
Larynx	0.0	1.0	0.0
Leisure Activities	0.0	1.0	1.0
Lens Diseases	0.0	1.0	0.0
Ligaments	0.0	1.0	0.0
Lymph Node Excision	0.0	1.0	0.0
Lymphoid Tissue	0.0	2.0	0.0
Macrocyclic Compounds	0.0	2.0	0.0
Manufactured Materials	0.0	1.0	1.0
Membranes	0.0	3.0	0.0
Microbiota	0.0	1.0	0.0
Minimally Invasive Surgical Procedures	0.0	3.0	0.0

Multiple Trauma	0.0	2.0	0.0
Muscles	0.0	1.0	0.0
Muscular Diseases	0.0	13.0	2.0
Musculoskeletal Abnormalities	0.0	1.0	0.0
Musculoskeletal Physiological Phenomena	0.0	4.0	1.0
Nails	0.0	1.0	0.0
Neoplastic Syndromes, Hereditary	0.0	2.0	0.0
Nervous System Diseases	0.0	1.0	0.0
Nervous System Malformations	0.0	1.0	0.0
Neurodegenerative Diseases	0.0	3.0	1.0
Neurodevelopmental Disorders	0.0	2.0	0.0
Neurosecretory Systems	0.0	1.0	0.0
Nucleosides	0.0	1.0	0.0
Nutrition Disorders	0.0	4.0	0.0
Obstetric Surgical Procedures	0.0	2.0	0.0
Ocular Hypertension	0.0	3.0	0.0
Ocular Motility Disorders	0.0	3.0	0.0
Oncogenic Viruses	0.0	1.0	0.0
Ophthalmologic Surgical Procedures	0.0	1.0	0.0
Optic Nerve Diseases	0.0	2.0	0.0
Optical Devices	0.0	1.0	0.0
Optical Phenomena	0.0	1.0	0.0
Organization and Administration	0.0	2.0	0.0
Orthopedic Procedures	0.0	1.0	0.0
Otorhinolaryngologic Neoplasms	0.0	4.0	0.0
Pain Management	0.0	1.0	0.0
Pancreas	0.0	1.0	0.0
Pancreatic Diseases	0.0	5.0	0.0
Paraneoplastic Syndromes	0.0	1.0	0.0
Parasitic Diseases, Animal	0.0	1.0	0.0

Parathyroid Diseases	0.0	1.0	0.0
Pathological Conditions, Anatomical	0.0	21.0	3.0
Pelvic Infection	0.0	1.0	0.0
Periodontics	0.0	1.0	0.0
Peripheral Nervous System	0.0	3.0	1.0
Peritoneal Diseases	0.0	4.0	0.0
Pharmaceutical Preparations	0.0	1.0	0.0
Pharmacologic Actions	0.0	8.0	1.0
Pigments, Biological	0.0	4.0	0.0
Pituitary Diseases	0.0	3.0	2.0
Plants	0.0	6.0	3.0
Polymers	0.0	1.0	1.0
Polysaccharides	0.0	1.0	1.0
Precancerous Conditions	0.0	2.0	0.0
Pregnancy Complications	0.0	3.0	1.0
Pregnancy, puerperium and perinatal conditions	0.0	1.0	0.0
Prostheses and Implants	0.0	1.0	0.0
Prosthesis Implantation	0.0	1.0	0.0
Psychology, Applied	0.0	1.0	0.0
Psychology, Social	0.0	1.0	0.0
Psychotherapy	0.0	1.0	0.0
Public Health Dentistry	0.0	1.0	0.0
Publishing	0.0	1.0	0.0
Punctures	0.0	1.0	0.0
Pupil Disorders	0.0	2.0	0.0
Purpura, Thrombocytopenic	0.0	1.0	0.0
Quality Assurance, Health Care	0.0	1.0	0.0
Quality of Health Care	0.0	2.0	2.0
RNA Viruses	0.0	3.0	1.0
Radiotherapy	0.0	1.0	0.0

Refractive Errors	0.0	1.0	0.0
Regeneration	0.0	1.0	0.0
Renal Replacement Therapy	0.0	1.0	0.0
Reproductive Tract Infections	0.0	1.0	0.0
Respiratory Hypersensitivity	0.0	2.0	2.0
Respiratory Physiological Phenomena	0.0	1.0	2.0
Respiratory System	0.0	1.0	0.0
Respiratory Tract Fistula	0.0	2.0	0.0
Respiratory Tract Neoplasms	0.0	5.0	0.0
Retinal Diseases	0.0	6.0	0.0
Retreatment	0.0	1.0	0.0
Rupture	0.0	1.0	1.0
Schizophrenia Spectrum and Other Psychotic Disorders	0.0	2.0	0.0
Sexual Dysfunctions, Psychological	0.0	2.0	0.0
Sexually Transmitted Diseases	0.0	1.0	1.0
Sheep Diseases	0.0	1.0	0.0
Skeleton	0.0	8.0	4.0
Skin	0.0	1.0	0.0
Sleep Wake Disorders	0.0	3.0	1.0
Smart Materials	0.0	1.0	0.0
Social Control, Formal	0.0	2.0	0.0
Socioeconomic Factors	0.0	1.0	0.0
Soft Tissue Injuries	0.0	1.0	0.0
Somatoform Disorders	0.0	1.0	0.0
Sorption Detoxification	0.0	1.0	0.0
Spinal Cord Injuries	0.0	1.0	0.0
Sprains and Strains	0.0	1.0	0.0
Surgical Equipment	0.0	1.0	0.0
Surgical and medical procedures	0.0	1.0	0.0
Survivors	0.0	1.0	1.0

Swine Diseases	0.0	1.0	0.0
Temporomandibular Joint Disorders	0.0	2.0	0.0
Tendon Injuries	0.0	1.0	0.0
Thoracic Diseases	0.0	1.0	0.0
Thoracic Injuries	0.0	3.0	0.0
Thoracic Surgical Procedures	0.0	2.0	1.0
Tooth Diseases	0.0	4.0	0.0
Tooth Injuries	0.0	1.0	0.0
Transplantation	0.0	2.0	0.0
Transplantation Immunology	0.0	1.0	0.0
Triazenes	0.0	1.0	0.0
Urinary Tract	0.0	2.0	1.0
Urogenital Surgical Procedures	0.0	2.0	0.0
Uveal Diseases	0.0	2.0	0.0
Vascular System Injuries	0.0	1.0	0.0
Vector Borne Diseases	0.0	4.0	0.0
Wound Infection	0.0	1.0	1.0
Wounds and Injuries	0.0	1.0	0.0
Adaptation, Biological	0.0	0.0	1.0
Amides	0.0	0.0	1.0
Anthropometry	0.0	0.0	1.0
Anxiety Disorders	0.0	0.0	1.0
Arm Injuries	0.0	0.0	1.0
Bone Marrow Cells	0.0	0.0	1.0
Catheter-Related Infections	0.0	0.0	1.0
Crush Injuries	0.0	0.0	1.0
Decompression, Surgical	0.0	0.0	1.0
Equipment and Supplies	0.0	0.0	1.0
Extravasation of Diagnostic and Therapeutic Materials	0.0	0.0	1.0
Hamartoma	0.0	0.0	1.0

Health Facilities	0.0	0.0	2.0
Health Personnel	0.0	0.0	1.0
Heat Stress Disorders	0.0	0.0	1.0
Hematopoietic System	0.0	0.0	1.0
Hip Injuries	0.0	0.0	1.0
Homeostasis	0.0	0.0	1.0
Jaw	0.0	0.0	1.0
Laboratories	0.0	0.0	1.0
Liver	0.0	0.0	1.0
Membrane Lipids	0.0	0.0	1.0
Microsurgery	0.0	0.0	1.0
Nebulizers and Vaporizers	0.0	0.0	1.0
Neoplasms, Second Primary	0.0	0.0	1.0
Neuroimaging	0.0	0.0	1.0
Occupational Groups	0.0	0.0	1.0
Physical Stimulation	0.0	0.0	1.0
Psychological Techniques	0.0	0.0	1.0
Shoulder Injuries	0.0	0.0	1.0
Skin Physiological Phenomena	0.0	0.0	1.0
Urinary Tract Infections	0.0	0.0	1.0
Urogenital Abnormalities	0.0	0.0	1.0

D. Rates of all the Parent Events

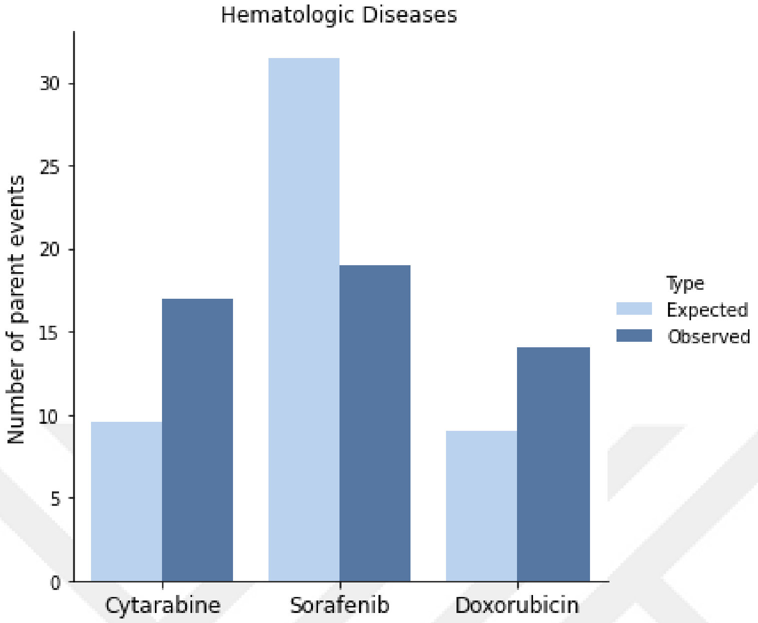


Figure D.1. The expected and observed rates of the "Hematologic Diseases" parent event.

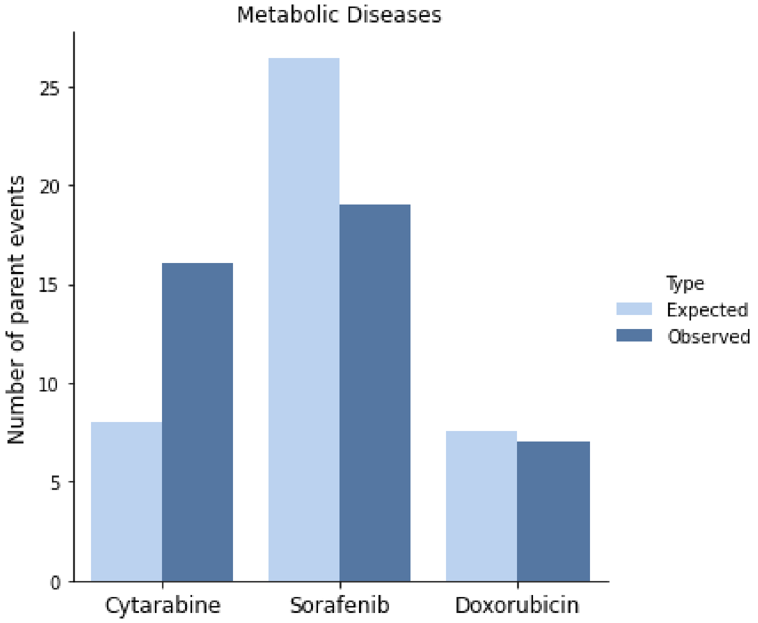


Figure D.2. The expected and observed rates of the "Metabolic Diseases" parent event.

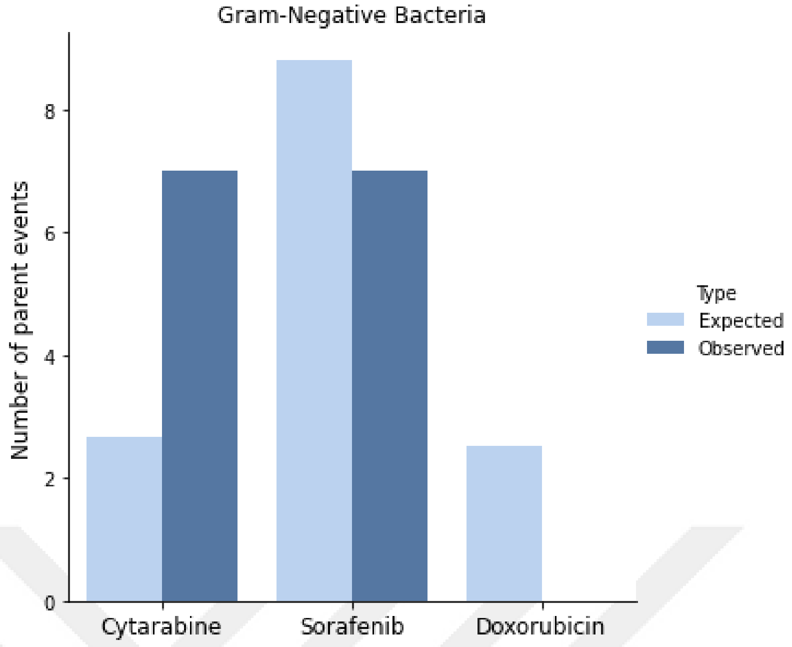


Figure D.3. The expected and observed rates of the "Gram-Negative Bacteria" parent event.

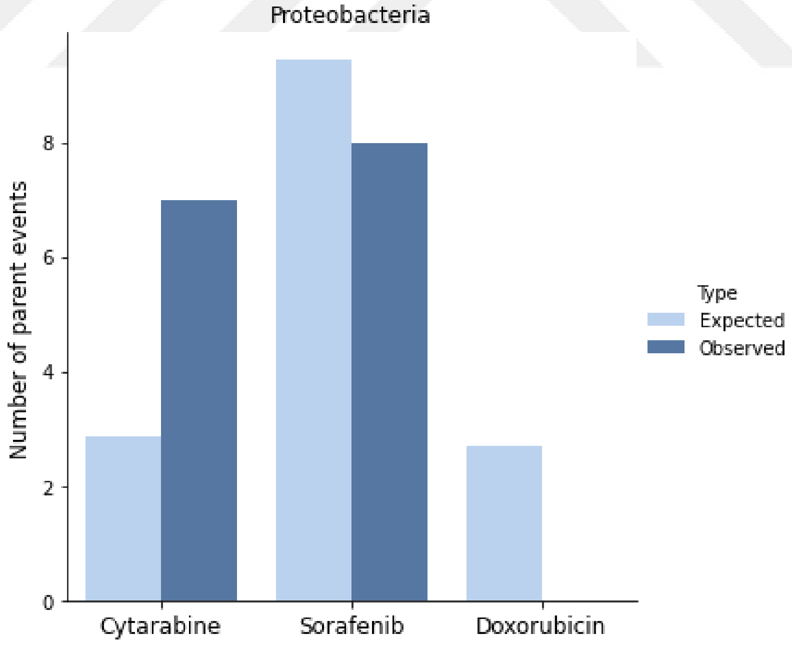


Figure D.4. The expected and observed rates of the "Proteobacteria" parent event.

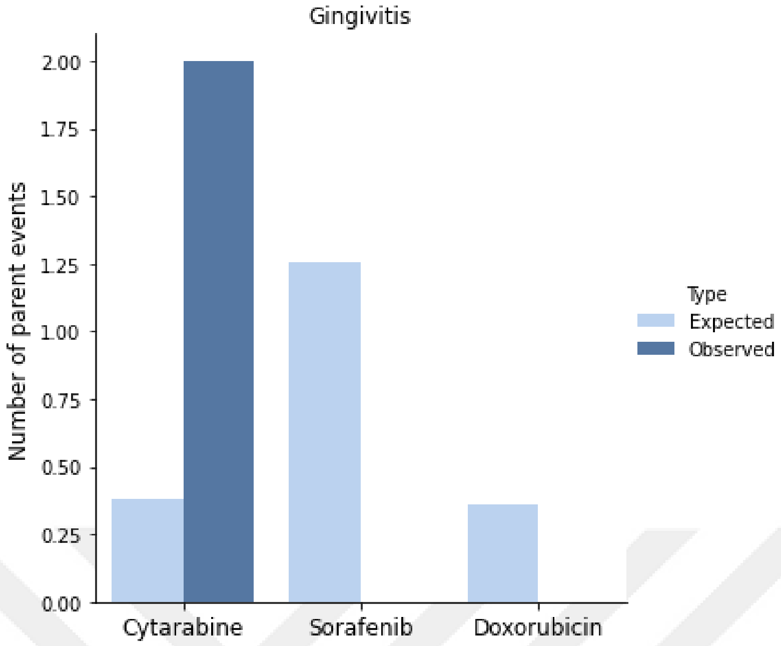


Figure D.5. The expected and observed rates of the “Gingivitis” parent event.

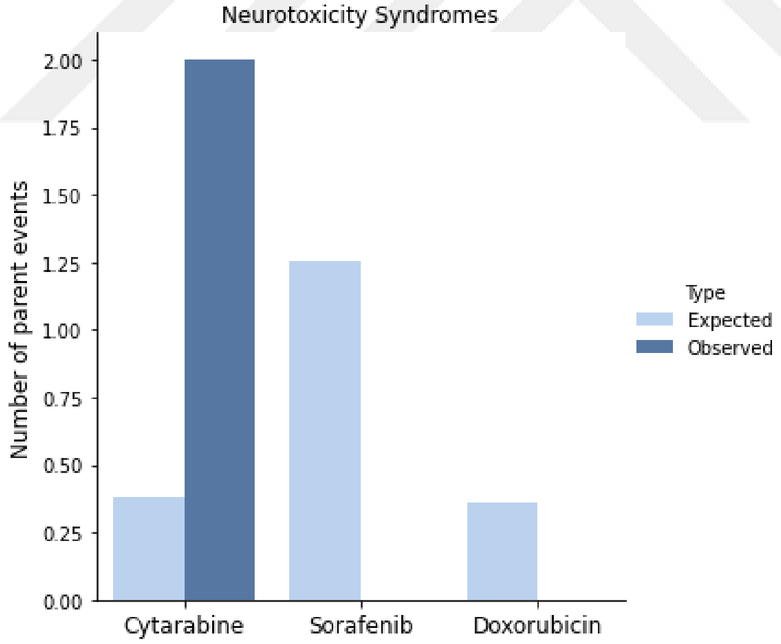


Figure D.6. The expected and observed rates of the “Neurotoxicity Syndromes” parent event.

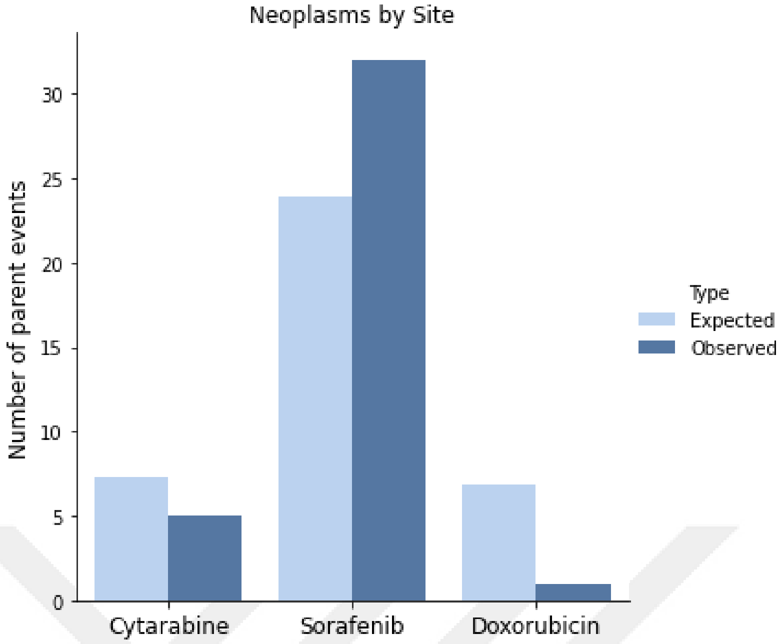


Figure D.7. The expected and observed rates of the "Neoplasms by Site" parent event.

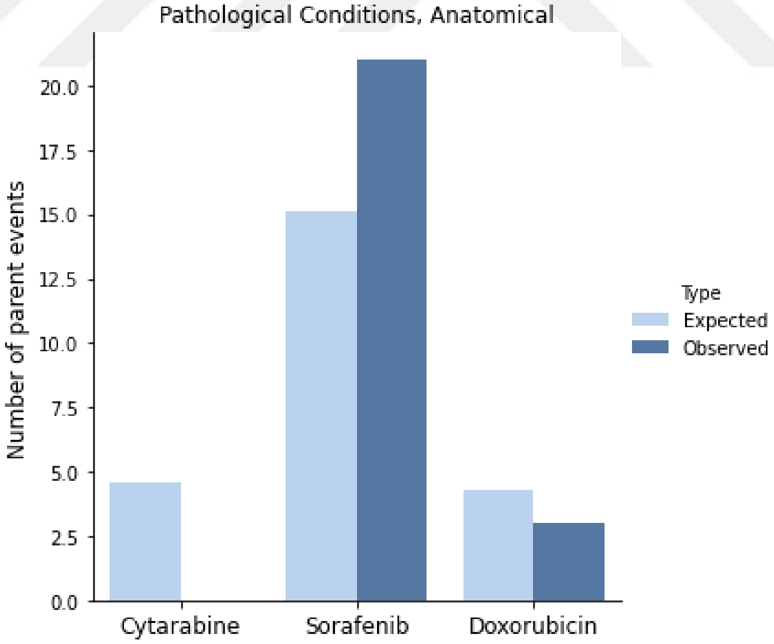


Figure D.8. The expected and observed rates of the "Pathological Conditions, Anatomical" parent event.

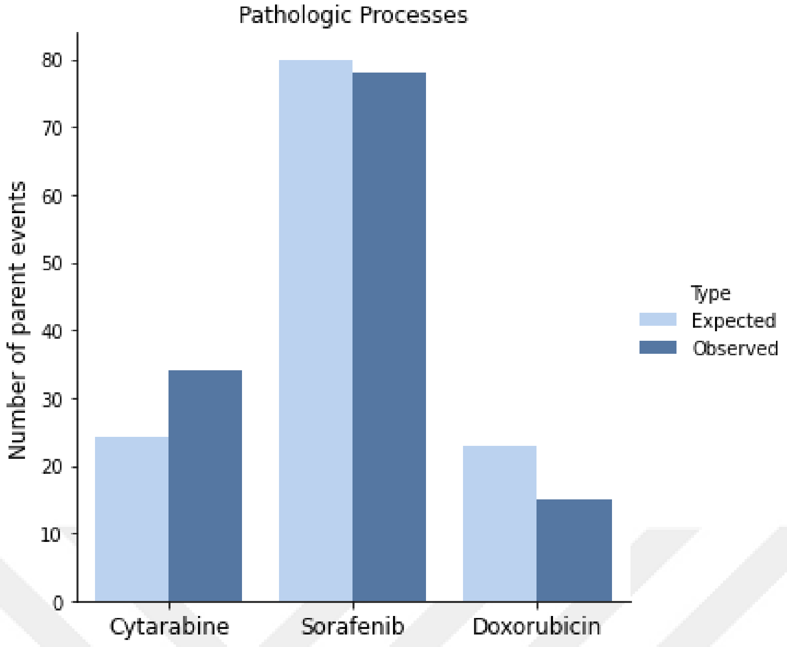


Figure D.9. The expected and observed rates of the "Pathologic Processes" parent event.

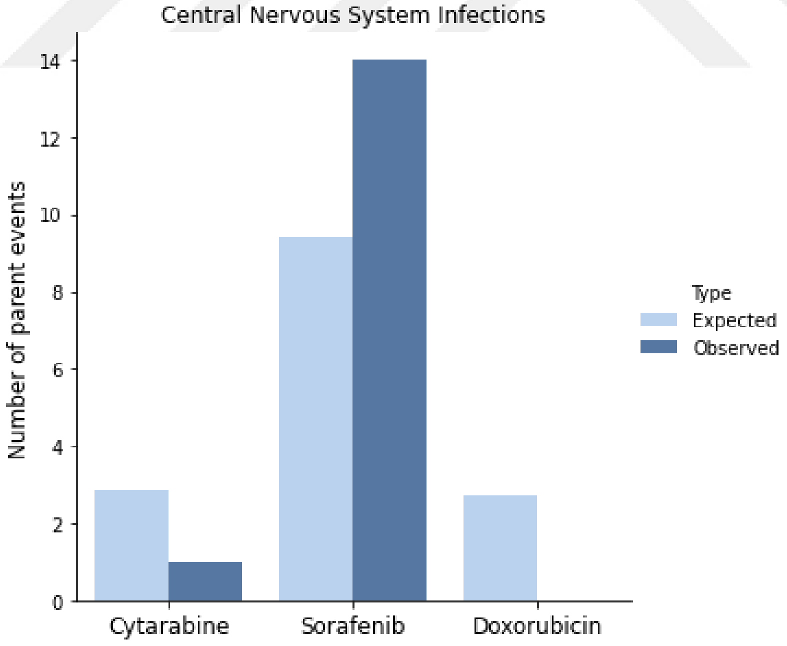


Figure D.10. The expected and observed rates of the "Central Nervous System Infections" parent event.

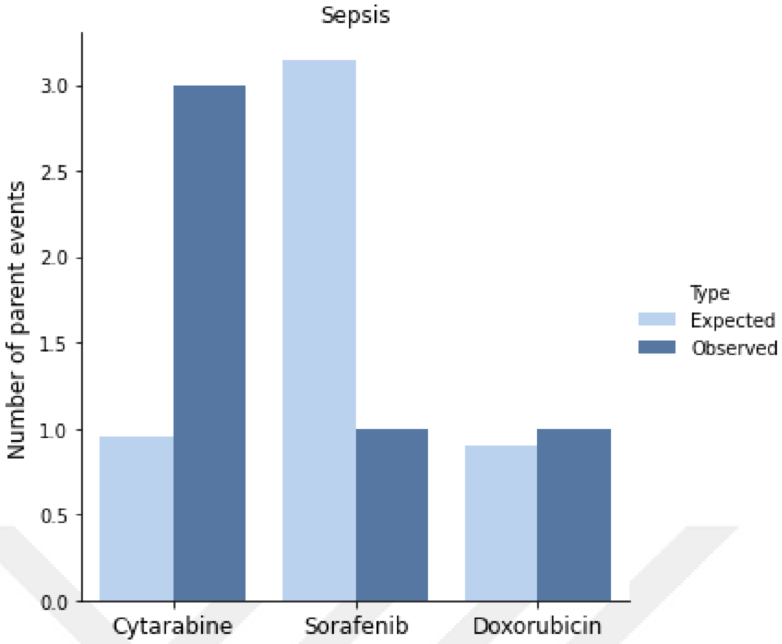


Figure D.11. The expected and observed rates of the “Sepsis” parent event.