A FRAMEWORK TO EXPLORE SPATIO-TEMPORAL SURVEILLANCE OF ADVERSE EVENTS FOR POST MARKET APPROVED DRUGS & VACCINES

by

Ahmed Askar
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and GeoInformation Sciences

Committee:

Andreas Züfle — Dr. Andreas Züfle, Dissertation Director

Dr. Ruixin Yang, Committee Member

Dr. Maction Komwa, Committee Member

Dr. Hong Xue, Committee Member

Dieter Pfoser — Dr. Dieter Pfoser, Department Chairperson

Donna M. Fox — Dr. Donna M. Fox, Associate Dean,
Office of Student Affairs & Special Programs,
College of Science

Dr. Fernando R. Miralles-Wilhelm, Dean,
College of Science

Date: _____ Summer Semester 2022
George Mason University
Fairfax, VA

A FRAMEWORK TO EXPLORE SPATIO-TEMPORAL SURVEILLANCE OF ADVERSE EVENTS FOR POST MARKET APPROVED DRUGS & VACCINES

by

Ahmed Askar

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and GeoInformation Sciences

Committee:

_____	Dr. Andreas Züfle, Dissertation Director
_____	Dr. Ruixin Yang, Committee Member
_____	Dr. Maction Komwa, Committee Member
_____	Dr. Hong Xue, Committee Member
_____	Dr. Dieter Pfoser, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Fernando R. Miralles-Wilhelm, Dean, College of Science

Date: _____ Summer Semester 2022
George Mason University
Fairfax, VA

A Framework to Explore Spatio-Temporal Surveillance of Adverse Events For Post Market Approved Drugs & Vaccines

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Ahmed Askar
Master of Public Health
Wright State University, 2013
Bachelor of Science
The Ohio State University, 2008

Director: Dr. Andreas Züfle, Professor
Department of Geography and GeoInformation Science

Summer Semester 2022
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my parents, my wife and my sons. In the words of Prophet Mohamed(PBUH) "O Allah, I ask You for beneficial knowledge, goodly provision and acceptable deeds"[1]

Acknowledgments

I would like to thank my dissertation chair, Dr. Andreas Züfle, and committee members for encouraging me to push through the continuous improvements to this dissertation and showing me how to do quality research. I also want to thank my wife, parents and teachers for their support to overcome the many hurdles in seeking knowledge and publishing scholarly work. I also want to thank my sons for their patience while writing this dissertation. Lastly I will like to thank my supervisors who have help funded my PhD.

Table of Contents

Page

List of Tables . . . viii
List of Figures . . . ix
Abstract . . . 0
1 Introduction . . . 1
1.1 Purpose . . . 1
1.2 Pharmacovigilance . . . 3
1.2.1 Public Health Surveillance . . . 5
1.3 Data mining . . . 5
1.3.1 Frequent Item-set Mining . . . 5
1.3.2 Topic modeling . . . 6
2 Preliminaries . . . 7
2.1 Related Work . . . 7
2.2 Pharmacovigilance . . . 9
2.2.1 Adverse Effects of Blood Thining Drugs . . . 9
2.2.2 Adverse Effects of COVID-19 Vaccines . . . 10
2.2.3 Population/Demographic Factor . . . 11
2.2.4 Environment Factor . . . 12
2.2.5 Drug Quality Factor . . . 13
2.3 Adverse Event Databases . . . 14
2.3.1 Drug Adverse Event Reporting System . . . 14
2.3.2 Vaccine Adverse Event Reporting System . . . 15
2.3.3 Limitation of Adverse Event Databases . . . 15
3 Research Methodology . . . 17
3.1 Data selection . . . 17
3.1.1 FAERS . . . 17
3.1.2 VAERS . . . 18
3.2 Data Processing . . . 19
3.3 Data Mining and Data evaluation . . . 20
3.3.1 Frequent (k) Item-set Mining . . . 20

3.3.2 Topic Modeling . . . 20
3.3.3 Similarity Test . . . 21
3.3.4 Finding Spatial Clusters . . . 22
3.3.5 Spatial Auto-correlation Test . . . 22
4 Clustering of Adverse Events for Post-Market Approved Drugs using Frequent Itemset . . . 24
4.1 Introduction . . . 25
4.2 Related Work . . . 26
4.3 Problem Definition . . . 28
4.4 Methodology . . . 31
4.4.1 Data Collection . . . 32
4.4.2 Frequent (k) Adverse Event Set Mining . . . 36
4.4.3 Similarity Measure between Sets of Adverse Events . . . 38
4.4.4 Finding Spatial Clusters . . . 42
4.4.5 Spatial Auto-correlation Test . . . 43
4.5 Results . . . 44
4.6 Conclusion . . . 49
5 Clustering of Adverse Events of Post-Market Approved Drugs using Latent Dirichlet Allocation . . . 51
5.1 Introduction . . . 52
5.2 Related Work . . . 53
5.3 Problem Definition . . . 56
5.4 Methodology . . . 59
5.4.1 Data Collection . . . 59
5.4.2 Adverse Event Topic Mining using Latent Dirichlet Allocation . . . 62
5.4.3 Similarity between Sets of Adverse Events . . . 65
5.4.4 Clustering Regions by Adverse Events . . . 66
5.4.5 Measure of Spatial Autocorrelation . . . 68
5.5 Experimental Evaluations . . . 69
5.5.1 Qualitative Analysis of Latent Topics among Adverse Events . . . 69
5.5.2 Qualitative Analysis of Adverse Event Clusters . . . 73
5.5.3 Spatial Autocorrelation . . . 73
5.6 Conclusion . . . 76
6 Clustering Adverse Events of COVID-19 Vaccines across the United States . . . 78
6.1 Introduction . . . 78
6.2 Related Work . . . 81

6.3 Problem Definition 82
6.4 Methodology 84
6.4.1 Latent Adverse Event Topic Modeling 84
6.4.2 Spatial Clustering of Vaccine Adverse Event Topics 87
6.4.3 Spatial Autocorrelation 89
6.5 Experimental Evaluation 90
6.5.1 Qualitative Analysis of Topics 91
6.5.2 Spatial Anaylsis of COVID-19 Adverse Event Topics 91
6.6 Conclusion 93
7 Final Conclusion 95
Bibliography 98

List of Tables

Table		Page
4.1	Table of Notations	28
4.2	Sample records of Adverse Event Report Database. Each Line is an Adverse Event.	29
4.3	Study areas for Drugs Rivaroxaban, Dabigatran and Apixaban for $k \in \{5, 10\}$.	34
4.4	Example k-most frequent sets of adverse events in Austria in Year 2016 for Drugs Rivaroxaban and Dabigatran. For each set the support among adverse events of that year in Austria is provided.	38
4.5	Top-5 Frequent Adverse Event between 2 countries (Germany and Austria) for Rivaroxaban in year 2014	41
4.6	Statistically significant clusters of AEs for Rivaroxaban, Dabigatran and Apixaban for Year 2014-2017 in Europe for $k \in \{5, 10\}$.	44
5.1	Table of Notations	55
5.2	Sample records of Adverse Event Report Database. Each Line corresponds to an Adverse Event.	57
5.3	Study areas for the drugs used in this study.	61
5.4	Top-10 most probably keywords for K=5 for Rivaroxaban.	66
5.5	Top-10 most probably keywords for K=5 for Dabigatran.	67
5.6	Top-10 most probably keywords for K=5 for Apixaban.	68
5.7	Statistically significant ($p < 0.05$) clusters for Rivaroxaban, Dabigatran and Apixaban for Year 2014-2017 for $k \in \{5, 10\}$.	70
6.1	Sample records of Adverse Event Report Database. Each Line is an Adverse Event.	83
6.2	Top-10 most probably adverse effects per topics across all regions and all COVID-19 vaccine brands.	90
6.3	Moran's I measure of global spatial autocorrelation for each of the $K = 10$ topics of COVID-19 adverse events.	92

List of Figures

Figure		Page
1.1	Example of Adverse Events.	2
1.2	Dissertation Motivation.	4
3.1	An Overview of the process in Knowledge Discovery in Databases	17
3.2	Route from vaccines adverse event to VAERS Report	19
4.1	Road map for spatiotemporal association model using Adverse Events Report Submission to FDA FAERS Database.	32
4.2	Study area for analyzing Top 5 & 10 unique AEs for Rivaroxaban, Apixaban and Dabigatran from year 2014 to 2017 for European countries listed with their ISO alpha-2 code	35
4.3	Adverse Events Report Submission to the FDA FAERS Database per drug per year	36
4.4	Results for Drug Rivaroxaban, $k = 5$, Year 2014. Distance matrix and hierarchical clustering (left). Adverse effect cloud weighted by frequency across the cluster of countries {at, be, ch, de, gb, gr, se} (right). Color of the text is an artifact of the python library used to generate word cloud.	45
4.5	Results for Drug Rivaroxaban, $k = 5$, Year 2016. Distance matrix and hierarchical clustering (left). Adverse effect cloud weighted by frequency across the cluster of countries {at, dk, fi, gb, ie, no, si} (right). Color of the text is an artifact of the python library used to generate word cloud.	46
4.6	Results for Drug Rivaroxaban, $k = 10$, Year 2014. Distance matrix and hierarchical clustering (left). Adverse effect cloud weighted by frequency across the cluster of countries {de, dk, es} (right). Color of the text is an artifact of the python library used to generate word cloud.	47
5.1	Road map for spatiotemporal clustering of topics generated from Latent Dirichlet Allocation using Adverse Events Report Submission to FDA FAERS Database.	60

5.2 Adverse Events Report Submission to the FDA FAERS Database per drug per year 62
5.3 LDA Topic Modeling of Events. For each adverse event a topic distribution θ is estimated and for each topic i, an adverse effect distribution φ_i is estimated. Given a topic Z generate from θ, observable adverse effects (AEs) are generated from φ_Z. 63
5.4 Coherence Scores vs Number of Topics to determine the number k of latent topics for LDA. 71
5.5 Clustering for Drug Rivaroxaban, $k = 5$, Year 2015 72
5.6 Clustering for Drug Dabigatran, $k = 5$, Year 2014 74
5.7 Clustering for Drug Apixaban, $k = 5$, Year 2017 75
6.1 COVID-19 Adverse Effect Clouds per Region. 79
6.2 LDA Topic Modeling of Adverse Events. For each adverse event a topic distribution θ is estimated and for each topic i, an adverse effect distribution φ_i is estimated. Given a topic Z generated from θ, observable adverse effects (AEs) are generated from φ_Z. 85
6.3 Pair-wise similarity matrix of latent topics of COVID-19 vaccine adverse events of counties in the United States. 88
6.4 Local Indicator of Spatial Autocorrelation (LISA). Light red areas correspond to high-high clusters. Light blue areas are low-low clusters. Dark red and dark blue areas corresponds to high-low and low-high outliers. 93

Abstract

A FRAMEWORK TO EXPLORE SPATIO-TEMPORAL SURVEILLANCE OF ADVERSE EVENTS FOR POST MARKET APPROVED DRUGS & VACCINES

Ahmed Askar, PhD

George Mason University, 2022

Dissertation Director: Dr. Andreas Züfle

Discovering all drug and vaccine side effects during the development process is impossible. This dissertation aims to propose a framework in exploring spatiotemporal adverse event surveillance models by identifying adverse effects, which co-locate together and is associated with FDA approved drugs or vaccines using spatial statistics and spatial science. This study aims to find statistically significant spatio-temporal clusters among co-occurring adverse effects. We use data obtained from the FDA's Adverse Event Reporting System (FAERS) and Vaccine Adverse Event Reporting System (VAERS) to explore the spatio-temporal distribution of combinations of adverse effects using two methods:

- Frequent Itemset Mining - to mine the most frequent sets of adverse events.
- Latent Dirichlet allocation (LDA) -to mine the most frequent group of topics related to adverse effects.

To assess the similarity of sets of adverse events or topics between spatial regions, we employ textual comparison algorithms. We apply an agglomerative hierarchical clustering approach to find clusters of regions that exhibit similar adverse events or topics. Finally, we explore the resulting clusters to discover spatial autocorrelation patterns using Global and Local Moran's I measure of spatial autocorrelation. Our approach can be applied to any product where after consumption or application results in adverse events, to study if spatially localized side-effects that may justify further investigation.

Chapter 1: Introduction

The contribution of this dissertation is the ability to mine for co-occurring adverse events and identify hidden spatial temporal patterns in adverse events associated with FDA approved drugs or vaccines. While we can use traditional pharmacovigilance or data mining techniques to uncover adverse event patterns however the use of spatial science to uncover hidden spatial temporal patterns in adverse events is limited.

The data mining algorithms in this dissertation improves upon pharmacovigilance research to mine for co-occurring adverse events and identify any spatial temporal patterns.

1.1 Purpose

The expectation of the proposed framework would help detect adverse events early to allow accurate risk assessment and appropriate response to the problem. This study will help minimize the number of adverse effects on individuals receiving the medication and lessen the potential negative impact on public health programs and alert medical professionals of the exact side effects. Discovering all side effects during the drug development process is impossible. Adverse side effects of a drug may vary over space and time due to different population demographic, environment factors, and drug quality. A major challenge in vaccine/drug development is determining possible side effects. Recent analysis found that it took a median of 4.2 years after a drug's initial approval for major safety concerns to be discovered [2]. Serious side effects could be life-threatening, which can lead to death. While less severe AEs such as rash, nausea, and fatigue might not be dangerous, however, they can lead to avoidance in taking the drug as prescribed, which can lead to a severe consequence [3].

An adverse events (AEs) is any undesirable experiences associated with the use of a

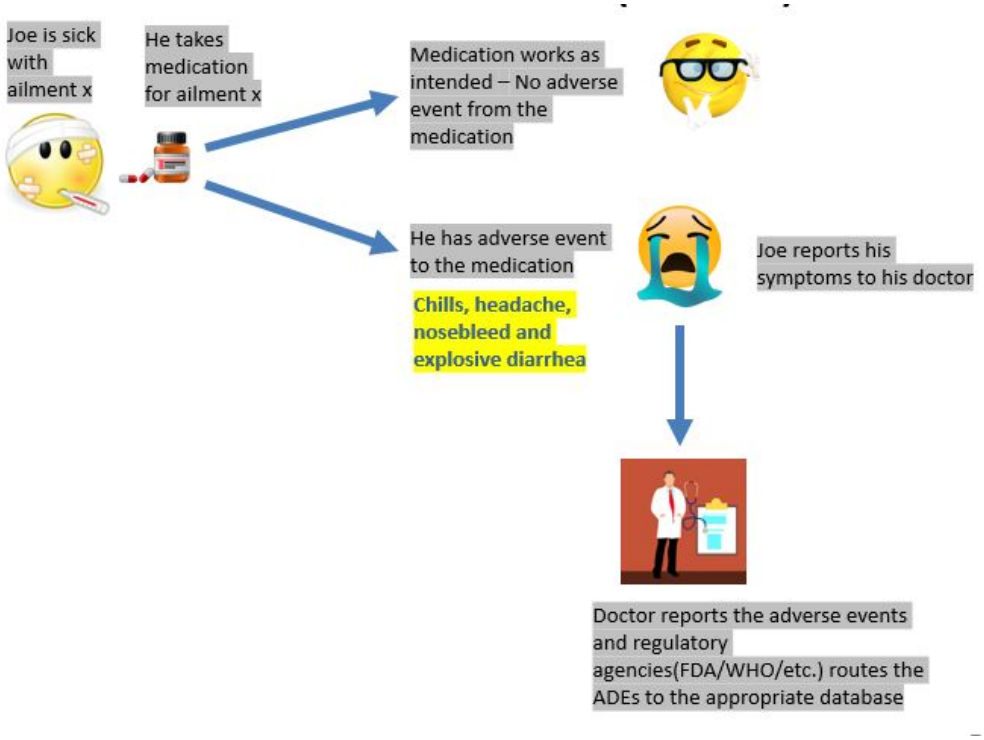


Figure 1.1: Example of Adverse Events.

medical product. Every year, many patients experience AEs from medical products , which can present in many forms. Most AEs are temporary or a nuisance such as "rash" or "nausea" however some AEs are life threatening or could cause death. Less serious adverse events might not have a substantial direct impact on the population's health. However, they may lead to noncompliance with, or interruption of, treatment, which may eventually reduce therapeutic-related benefits for the individual [4].

There are many AEs that researchers have found to be in health care settings associated with errors in prescribing or administering drugs., many of which are considered potentially preventable. Medication errors alone are estimated to account for over 7000 deaths annually in the US [5] and many more in developing nations such as Somaliland, where anyone can import cheap counterfeit or substandard medical product due to the private unregulated health care system. Pharmaceuticals complement other types of health care services to reduce morbidity and mortality rates and enhance quality of life in both developed and developing nations [6]. Medical products need to work as intended the United States Food

and Drug Administration (FDA) has developed AE self- reporting tools called MedWatch in which patients, practitioners, and drug manufacturers can report AEs including severe allergic reactions, side effects, medication errors/product use errors, product quality problems, and therapeutic failures for all FDA regulated products. See Figure 1.1 for an example of possible route of AEs from patient to reporting. These reports are available to everyone including researchers, patients and health practitioners in publicly available databases [7].

Our research question is to identify hidden spatial patterns in reported adverse events associated with FDA approved drugs or vaccines See Figure 1.2. Does adverse events repeat regardless of space or time?

- If true, then its associated to the medication because its constant.
- If false, then there are spatial or temporal underlying co-founders, which introduce or exacerbate adverse event.

Some of these possible variables are discussed in chapter 2 and would need further research along with input from medical experts to investigate the underlying spatial or temporal association.

1.2 Pharmacovigilance

The field of pharmacovigilance aims at understanding the occurrence of adverse effects of drugs [8, 9]. Beyond understanding the adverse effects of single drugs, Zitnik, Agrawal, and Leskovec have studied the problem of modeling polypharmacy adverse effects, that is, adverse effects resulting from the interaction of multiple drugs. These important existing works provide solutions to finding significant links between specific drugs and specific adverse effects. However, these studies do not give any consideration to the spatial locations of these adverse effects. Could some patterns between drugs and adverse effects be explained by the spatial distribution of reported adverse effect records? Is it possible that some links

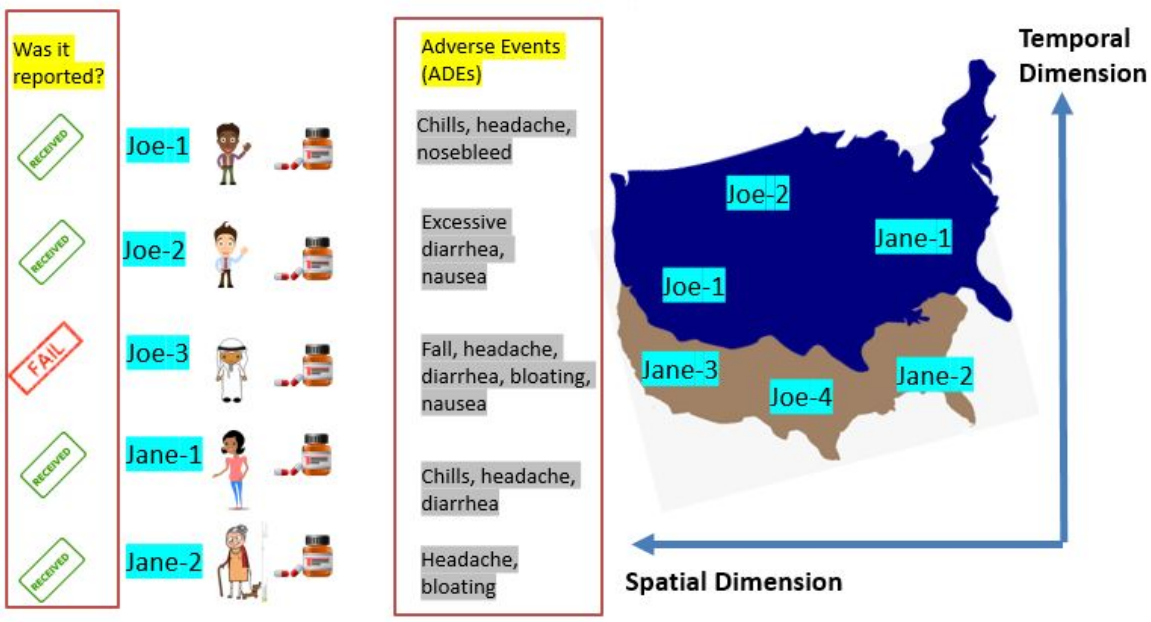


Figure 1.2: Dissertation Motivation.

between drugs and adverse effects are only observed in a specific region or during a certain time? Existing research leaves such questions largely unanswered. Fortunately, large databases of adverse events, such as the FDA's Adverse Event Reporting System (FAERS) and Vaccine Adverse Event Reporting System (VAERS) database are becoming increasingly available and enrich adverse events with both spatial and temporal information.

From complementary perspective, existing work has shown that adverse effects of a single drug or multiple combination of drugs may vary over space and time due to racial and ethnic disparities [10, 11, 12], environment [13, 14], and drug quality [15]. While these studies describe specific cases and specific drugs, there is no data-driven approach to identify such variations automatically.

1.2.1 Public Health Surveillance

Disease surveillance has been a critical ingredient in public health well over half a century. Spatial and Spatio-temporal analysis in Geographical information systems (GIS) can provide essential tools in assessment, prediction, and mitigation of disease, where the place can be considered as a proxy for the interaction between genetic factors, lifestyle and environment [16]. GIS is a series of tools for the acquisition, storage, retrieval, analysis, and display of spatial data and, coupled with data mining techniques becomes even more useful in terms of amount of methods that can be deployed. The importance of spatial and spatiotemporal data mining is growing with the increased incidence and availability of public health datasets [17].

1.3 Data mining

Data mining or knowledge discovery from a database (KDD) is used to determine useful, implicit and hidden patterns in a large dataset which was previously unknown [18]. For an overview of the process in Knowledge Discovery in Databases see Figure 3.1. It used in a wide range of disciplines such as public health [19].

1.3.1 Frequent Item-set Mining

Market Basket Analysis is one of many data mining techniques used usually by large retailers to uncover associations between items [20]. It works by looking for combinations of items in each transaction that frequently occur together. It allows market researchers to identify relationships between the items that customers buy, such as milk, bread and diapers [21]. Frequent Item-set and association rules mining are not restricted to market basket analysis, but instead, they can be applied in other settings [22]. We leverage frequent itemset mining to represent the adverse events reported in a spatial region as a set of mined frequent adverse effects. Then, we use this representation to cluster regions having similar frequent adverse effects for the same drug at the same time.

1.3.2 Topic modeling

Topic modeling is an unsupervised learning technique to discover underlying themes of a collection of documents. Latent Dirichlet Allocation (LDA) is one of the more common topic modeling techniques in the literature [23]. LDA assumes an underlying generative probabilistic model that produces the words of a text document given a mixture of k latent topics. Each topic is characterized by a distribution of words. While the traditional application for LDA is modeling of topics among news articles and microblogs [24], it has been used to model the latent topics of points of interest such as restaurants [25]. In the context of pharmacovigilance, LDA has been to find potentially unsafe dietary supplements [26], but without the consideration of the spatial distribution of latent topics among adverse effects. We leverage LDA to find underlying topics of adverse effects reported in a spatial region as a set of latent topics. We then employ this latent feature representation to find spatial clusters of regions that exhibit similar latent adverse effect topics.

Chapter 2: Preliminaries

In this chapter, we present the existing adverse event databases, current spatio-temporal concepts exploited by our algorithms and our approach to combine frequent item-set mining, latent dirichlet allocation with spatial science.

2.1 Related Work

Data mining or knowledge discovery from a database (KDD) is used to determine useful, implicit and hidden patterns in a large dataset which was previously unknown. It used in a wide range of disciplines related to data mining in public health [27]. Market Basket Analysis is one of many data mining techniques used, usually by large retailers to uncover associations between items [20]. It works by looking for combinations of elements that occur together frequently in transactions. It allows retailers to identify relationships between the things that customers buy. Frequent Item-set and association rules mining is not restricted to market basket analysis, but instead, they can be applied in other settings such as in pharmacovigilance [22]. Another algorithm used in this dissertation is Latent Dirichlet Allocation (LDA) Topic modeling is an unsupervised learning technique to discover underlying themes of a collection of documents. Latent Dirichlet Allocation (LDA) is one of the more common topic modeling techniques in the literature [23]. In the context of pharmacovigilance, LDA has been used to find potentially unsafe dietary supplements [26], but without the consideration of the spatial distribution of latent topics among adverse effects. To the best of our knowledge, no other work tries to use machine learning and spatial science to cluster adverse events however there are other existing approaches to study adverse events. Kreimeyer et al used probabilistic methods to identify duplicate cases in spontaneous adverse event reporting systems [28]. Botsis et al used text mining to extract

features from vaccine safety reports [29]. Ball and Botsis used network analysis to improve pattern recognition among adverse events [30]. Visual analytic and maps have been used to understand clusters of diseases such as COVID-19. The works of [31, 32] map the change in mobility and adherence to social distancing guildelines across the US. Elarde et al analysis provides a comprehensive understanding of mobility change in response to the COVID-19 pandemic. While Goa et al provided an interactive web-based mapping platform that provided timely quantitative information on how people in United States reacted to the social distancing guidelines. Agent based models have been used to do simulation models to understand clusters (hot spots) of disease cases and transmissions. Examples of diseases simulations that are data-driven include [33, 34, 35, 36]. Hinch et al looked at the unprecedented restrictions on social and economic activity during COVID-19 and simulated using an agent-based simulation of the epidemic including detailed age stratification and realistic social networks [33]. Pesavento et al found out that many agent based models lack realistic representations of human mobility so they proposed LDA to coupled foot-traffic data to develop a realistic model of human mobility in an agent based model(abm) [34]. Zufle et al looked at the complexity of human behavior using abm and how their location and co-location are affected. They captured the innate needs of a human-like population and explore how such needs shape social constructs such as friendship and wealth. Their model looked at agents social networks, which in turn affected the places the agents visited [35]. Kim et al addressed studying location-based social networks (LBSNs). In their paper, they addressed the missing comprehensive data in social network studies and which affects the causal links as to why movement happens in the first place [36]. The spatial computing community has done tremendous effort towards understanding the spread of COVID-19 by mining large sets of human mobility data related to the pandemic, evident by two workshops that have been organized on this topic. SpatialEpi'21 [37] and COVID-Workshop'20 [38] and two Special Issues of the ACM SIGSPATIAL Newsletter on this topic [39, 40]. There are other spatial problems where spatial science and machine learning where used to find

some solutions for problems related to foot-traffic during the pandemic using tensor factorization [41], house pricing prediction using recommendation systems [42], prediction of emotions using spatial statistics [43], and prediction of intensification of tropical storms using tensor factorization [44]. Public health surveillance data is the base of effective public health practice, Qazi et al [45] presented a large dataset of tweets discussing COVID-19 tags in ACM SIGSPATIAL Newsletter [39] for researchers. Health surveillance has improved in recent years due availability of mobile phones and mobile applications for digital contact tracing [46] and reporting medication errors [47]. Pharmacovigilance is a branch of public health surveillance and is the detection, assessment, understanding, and prevention of adverse effects [8, 9, 48].

2.2 Pharmacovigilance

The field of pharmacovigilance aims at understanding the occurrence of adverse effects of drugs [8, 9]. Specifically for vaccines, there is evidence that stress may have an amplifying effect on immune response and adverse events [49]. However, such aspects of understanding the interactions between drugs and other external factors are out of scope of this work. In this dissertation, we investigate the effect of location on adverse effects of blood thinners drugs in chapter 4, 5 and COVID-19 vaccines in chapter 6. Our approach in chapter 5 and 6 has been published in peered review journals [50, 51]. While location may be a proxy of other factors (such as stress), this work does not provide or imply any causality between location and adverse events. Yet, we hope that an understanding of the spatial distribution and autocorrleation of adverse events may help experts discover such causalities.

2.2.1 Adverse Effects of Blood Thining Drugs

In chapter 4 and 5, we explore spatial temporal clusters of a three FDA approved drugs for post market AEs patterns and trends. We didn't use concomitant drugs of FAERS dataset and used AEs reports associated to a single drug for Dabigatran, Rivaroxaban

and Apixaban. Existing work has shown that adverse effects of a single drug or multiple combination of drugs may vary over space and time due to racial and ethnic disparities [10, 11, 12], environment [13, 14], and drug quality [15]. In the United States, these three drugs appear to have similar effectiveness [52], although Apixaban may be associated with a lower bleeding risk and Rivaroxaban may be associated with an elevated bleeding risk. A similar study in Norway reached similar findings, showing that Dabigatran and Apixaban were both associated with significantly lower risk of major bleeding compared with Rivaroxaban [53]. While these studies investigated the differences of adverse effects across different drugs, these works did not consider spatial or temporal properties of the data. Combined with our knowledge that adverse affects vary across populations and space [10, 11, 12], we investigate if we can identify spatial clusters of regions that exhibit similar adverse effects using two data mining approaches i.e Frequent Item-set mining and Topic mining using latent dirichlet allocation.

2.2.2 Adverse Effects of COVID-19 Vaccines

In chapter 6, we explore spatial temporal clusters of three COVID-19 vaccines Janssen, Moderna, and Pfizer for AEs patterns and trends. Vaccines are, without any doubt, a paramount weapon to fight deadly diseases evident by the fact that "In 1900, for every 1,000 babies born in the United States, 100 would die before their first birthday, often due to infectious diseases" [54]. Furthermore, vaccines not only protect those receiving the vaccines but also vulnerable groups around them, such as new born babies, who may not be able to receive a vaccine [55]. Understanding and mitigating these adverse events will not only improve the well-being of those receiving the vaccines, but will also decrease fear of vaccines that leads to high vaccine hesitancy as observed during the COVID-19 pandemic [56]. We investigate if we can identify spatial clusters of regions that exhibit similar adverse effects using latent dirichlet allocation.

2.2.3 Population/Demographic Factor

Despite efforts to increase diversity in recruitment, clinical trials are getting less diverse over past two decades in the US due many factors and constraints on the volunteers. During the drug development and drug approval process, the potential harm and benefit of the drug is weighted because people may respond to treatments differently. Discovering side effects in medication will continue to be more challenging and costly due to less diversity clinical trials without a true sample size that is a reflective of the human fabric [57].

Albuterol is less effective in African American and Puerto Rican children compared with European American and Mexican children. Mak et al whole-genome sequencing study revealed some clues on reduced albuterol response associated with African-American and Puerto Rican children as it does for European American or Mexican children due to genes involved in immune response, lung capacity, and response to blockers. This study revealed new risk markers for children who would not respond as initially intended to albuterol and other current first-line anti-asthma drugs and help guide new development of new therapies specific to over come related adverse events [58].

An HIV Drug Abacavir causes potentially life-threatening adverse events if HLA-B*5701 gene is present 3 however the same genotype has been found to be resistant to miliaria.4 Caucasians have higher prevalence rates of HLA-B*57:01 (4–8%) than African-Americans, Asians, and Hispanics (0.2–4%) [59, 60].

Another example of adverse event associated with communication is the language barriers and understanding of hospital discharge instructions. The demographic of this adverse event may come from all racial group however the common factor is the language barrier between the patient and health care staff or discharge instructions. Karliner et al compared spanish-speaking, chinese-speaking, and english-speaking patients admitted to 2 urban hospitals between 2005 and 2008. This study found that the understanding of appointment type and medications after discharge was low with limited English-proficient patients demonstrating the potential medication use errors [61].

2.2.4 Environment Factor

Drug – Environment interaction

Adverse event can also be associated with environmental factors affecting the human body while taking the medication. Ketoconazole (Nizoral) and Delavirdine (Rescriptor) require an acidic environment to be absorbed. Solubility is drastically reduced in medication, which may raise stomach PH levels [62].

Drug – Food interaction

Genser's study looked at foods and drugs, when taken simultaneously, can alter the body's ability to utilize a particular food or drug, or cause serious side effects. Some medication require certain nutrition avoidance which can lead to serious consequences such as reduced absorption of certain oral antibiotics and predisposes the patient to treatment failure. Certain food inhibits enzymes in the gut which may lead to a significant change in oral bio-availability of drugs such as grape fruit juice which is selective intestinal CYP3A4 inhibitor. The overall exposure of some drugs can be increased by more than fivefold when taken with grapefruit juice and increase the risk of adverse effects [63]. Wayfarin, which is a blood thinner was found that food (mostly vegetables) normally high in vitamin K interferes with the effectiveness and safety of Wayfarin therapy [64].

Drug – Disease interaction

Drug-disease interactions are situations where the benefit/risk ratio of drugs for specific populations may have a negative effect on patients' comorbidities. A black box label is added to the labeling of drugs or drug products by the regulatory agency such as the FDA, when serious adverse reactions or special problems occur, particularly those that may lead to death or serious injury [65]. Cardiovascular diseases have the most drug-disease interactions. In elderly patients, 15%–16% of the patients had at least one drug-disease interaction al [66]. Black label recommendations support both pharmacists and physicians

by signalling clinically relevant drug-disease interactions at point of care, thereby improving medication safety [67].

Drug – Drug interaction

Almost 80% of Americans over the age of 60 are taking multiple drugs, and the occurrence drug to drug interaction and intensity of an adverse events increases with the number of drug combinations. Clinical trials typically focus on a single drug and rarely on drug to drug combinations due to the sheer amount of multiple drug combination of all FDA approved drugs. Such an enormous number of drug combinations and possible AEs from drug to drug interaction make it challenging and is needed for patient safety [68]. Zitnik et al use convolutional network to predict for multiple drug to drug AEs. Using neural network they construct for protein to protein interactions, drug to protein target interactions and drug to drug interactions to predict for AEs for multiple drug combination, which have not been used yet with patients [48].

2.2.5 Drug Quality Factor

In 2012, 753 patients in 20 states were diagnosed with a fungal infection after receiving steroid injections manufactured by New England Compounding Pharmacy for back pain, and this resulted in many patients being hospitalized and dying of fungal infection. This was due to the unsanitary manufacturing process by the compounding pharmacy and during an FDA investigation - an investigator described it as "fungal zoo". This outbreak was the largest public health crisis ever caused by a contaminated pharmaceutical drug/injection. Drug quality issues such as this one are not common however possible – it can come from the manufacturer or during the supply chain. These can also be detrimental to the patient however deadly, its not systematic [69].

2.3 Adverse Event Databases

Spatio-temporal pharmacovigilance is the detection, assessment, understanding, and prevention of adverse effects in the temporal and spatial dimensions. We used spatial data mining techniques along with tools from Geographic Information Systems (GIS) to mine information from FAERS Adverse event FDA database [8, 9]. Adverse Event (AEs) or Adverse Drug Reaction (ADRs) is any undesirable experience associated with the use of a medical product. The Food and Drug Administration (FDA) has developed AE self-reporting tools called MedWatch in which patients, practitioners, and drug manufacturers can report adverse events including severe allergic reactions, side effects, medication errors/product use errors, product quality problems, and therapeutic failures for all FDA regulated products. This information is available to researchers, patients, and practitioners in publicly available databases (FAERS) [7].

2.3.1 Drug Adverse Event Reporting System

FDA Adverse Event Reporting System (FAERS) is a database designed to support the FDA's post- marketing safety surveillance program and contains adverse event reports from a medication error and product quality complaints reported to the FDA. FAERS began on September 10, 2012, and replaced the Adverse Event Reporting System also known as Legacy AERS. Each extract covers reports received by FAERS during one quarter of the Year. Adverse events in FAERS are coded using terms in the Medical Dictionary for Regulatory Activities (MedDRA) terminology to standardized medical terminology so it will facilitate sharing of information by regulatory authorities, pharmaceutical companies, clinical research organizations, and health care professionals, which will allow better global protection of public health [70]. Reporting systems such as FAERS are critical tools for monitoring the safety and the standards. FDA uses FAERS for investigating new safety concerns, which might be related to an FDA regulated product or evaluating a manufacturer's compliance with reporting regulations and responding to outside requests for

information [7].

2.3.2 Vaccine Adverse Event Reporting System

Vaccine Adverse Event Reporting System (VAERS) is a database designed to support national early warning system to detect possible safety problems in U.S.-licensed vaccines. VAERS's co-managed by the CDC and the FDA. It similiar to FAERS in accepting reports of adverse events. Such as FAERS anyone can report an adverse event to VAERS. Healthcare professionals are required to report certain adverse events and manufacturers are required to report all adverse events that come to their attention to VAERS and FAERS for their intended products. These databases are not designed to determine if a vaccine or drugs caused a health problem, but is especially useful for detecting unusual or unexpected patterns of adverse event reporting that might indicate a possible safety problem which might lead to additional work and evaluation for further investigation of a possible safety concern [71].

2.3.3 Limitation of Adverse Event Databases

VAERS and FAERS share many limitations, both these systems are a passive reporting system, meaning it relies on individuals to send in reports of their experiences to CDC and FDA, which means that there will be under-reporting. The degree of under-reporting for both databases varies. More serious adverse events are going to be reported compare to less serious events. Reports vary in quality and completeness because reports are accepted from various sources such as patients or health care providers and reports vary in quality and completeness [71].

Information sharing is constraint by the HIPPA law. HIPAA (Health Insurance Portability and Accountability Act of 1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information. The law has emerged into greater importance in recently due to the health data breaches caused by cyberattacks and ransomware attacks on health insurers and providers. This law also puts a constraint on

researchers to safeguard PII that could potentially identify a specific individual. Any information that can be used to distinguish one person from another can be considered as PII. In the spatial data mining setting, a person's location is also data, which can be classified as PII and must be de-identified [72]. De-identification is the process used to prevent a person's identity from being connected with information. HIPPA law stats that zip-code data can only be used if zip-code has more than 20,000, zip-code with a population less than 20,000 has to go through another round of de-identification where zip-code are grouped by the first 3 numbers [73].

Chapter 3: Research Methodology

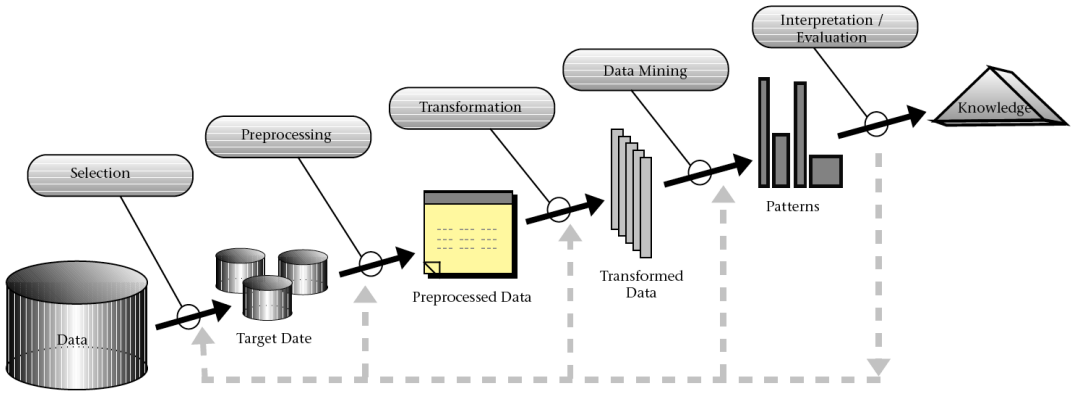


Figure 3.1: An Overview of the process in Knowledge Discovery in Databases

3.1 Data selection

3.1.1 FAERS

Adverse reports were downloaded from the FAERS public-facing database. FAERS public facing database is updated quarterly and was downloaded using Python scripts. Quarterly data files are in zip format available at fda.gov [74] or could be downloaded using OpenFDA API [75]. The quarterly data files include:

- demographic and administrative information and the initial report image ID number (if available);
- drug information from the case reports;

- reaction information from the reports;
- patient outcome information from the reports;
- information on the source of the reports;
- a "README" file containing a description of the files.

Spatial attributes are limited to the country level for the public-facing FAERS database due to HIPAA regulations. Due to data aggregation, all data were summarized at the country level to protect PII information and this also effects data variability and results. In this case, using United States as our study area was not plausible, we used European countries as a good backup due to to the richness in the data for adverse events and spatiotemporal attributes along with closeness in regulatory policies because of the EU overarching regulatory arm.

3.1.2 VAERS

Vaccine Adverse reports were downloaded from VAERS public-facing data portal. VAERS public facing database is updated weekly and are in zip format and were downloaded using Python scripts. Data files include:

- VAERS DATA, which includes demographic and administrative information and patient outcome information from the reports;
- VAERS Vaccine, which includes vaccine information from the case reports;
- VAERS Symptoms, which includes reaction information from the reports;

Spatial attributes are limited to the US State level for the public-facing VAERS database due to HIPAA regulations. Due to data aggregation, all adverse events data were summarized at the US State level to protect PII information. In this case, we used United States as our study area even though we knew that we will lose a lot of the data variability if the data was summarized at a lower hierarchical geographic entity such US County [71]. For an overview of the process and routes of adverse events in VAERS see Figure 3.2.

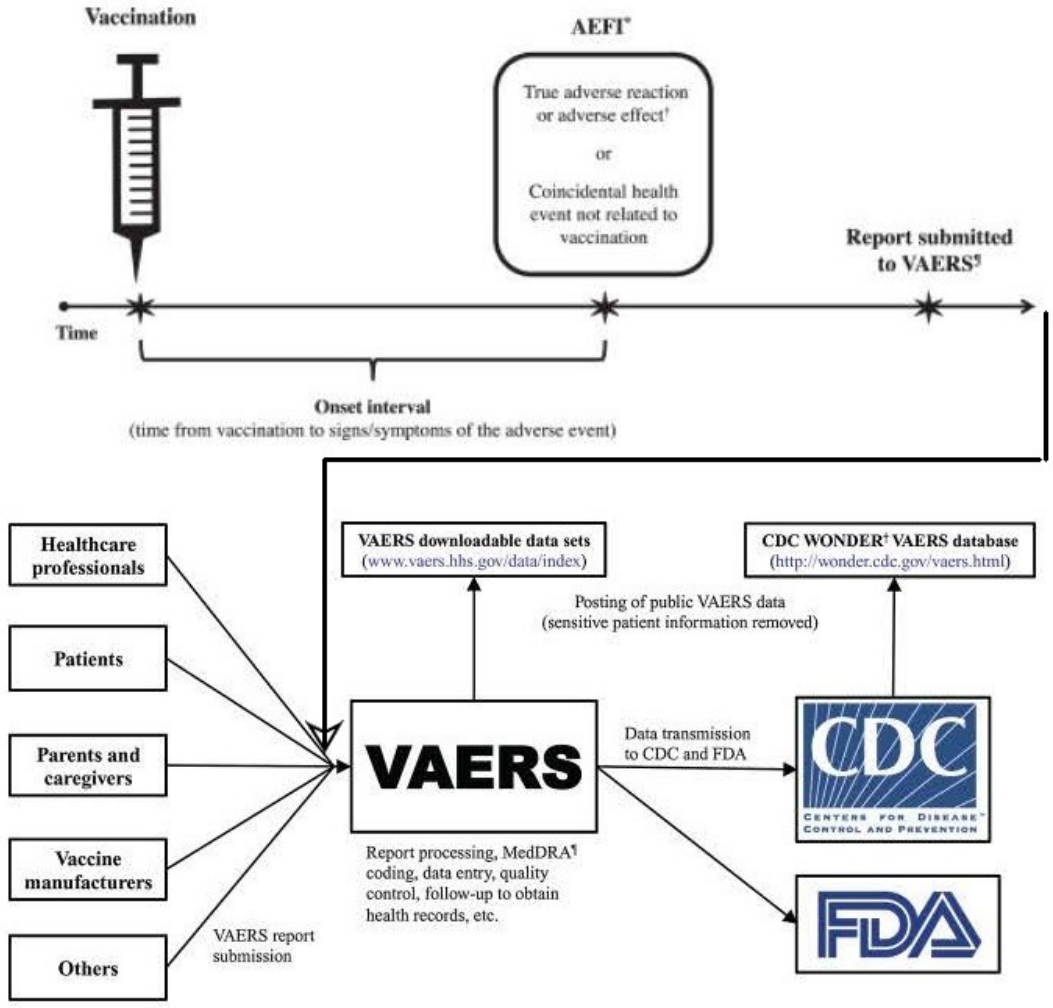


Figure 3.2: Route from vaccines adverse event to VAERS Report

3.2 Data Processing

Python programming language was used for all the data analysis and hosted in a Jupyter Notebook environment for reproducibility. The analysis process was relatively straightforward because of the ample amount of resources that are available for Python and the

following libraries. Pandas python library was used to carry out entire data analysis workflow in Python [76]. Mlxtend python library was used for frequent itemset data mining, gensim was used for topic modelling, numpy for mathematical calculations, geopandas for handling spatial attributes, and PYSAL for conducting spatial auto-correlation and statistical tests [77, 78, 79, 80, 81]. Spatial Data transformation is spatially enabling data with for mapping function that establishes a spatial correspondence between spatial geometry data points or polygons including spatial projections so spatial data aligns with each other.

3.3 Data Mining and Data evaluation

3.3.1 Frequent (k) Item-set Mining

To find interesting spatial rules and patterns in geographic space, we used Apriori algorithm to get sets of top (k) frequent item-sets. Frequent Item-set uses Apriori Algorithm and was introduced to find frequent groupings of items in a database containing baskets/records of items [82]. A priori Algorithm works by eliminating most large sets of candidates by looking first at smaller sets, and it recognizes that a large set cannot be frequent unless all its subsets are frequent [83]. The apriori algorithm generates a set of candidate item-sets. The transaction data set will then be scanned to see which sets meet the minimum support level (minsup). Itemset that doesn't meet the minimum support level will get tossed out. The remaining sets will then be combined to make item-sets with additional elements. Again, the transaction data set will be scanned, and item-sets, which does not meet the minimum support level will get tossed. The process will be repeated until all sets are tossed out or top (k) frequent itemset are met [83].

3.3.2 Topic Modeling

Topic modeling is an unsupervised learning technique to discover underlying themes of a collection of documents. Latent Dirichlet Allocation (LDA) is one of the more common topic modeling techniques in the literature. LDA assumes that adverse events in our database

contains a mix of topics that are found throughout the entire dataset [23]. There are many choices of topic modeling algorithms such as Latent Semantic Allocation (LSA) and Hierarchical Dirichlet Process (HDP). LSA assumes that words that are close in meaning will occur in similar pieces of text [84]. A challenge of mining adverse events is the potentially large number of different adverse effects in the database, which could appear anywhere in the database. Similarly HDP assumes all data set shares similarity or a base distribution [85]. Hence LSA and HDP were not a good choice for our adverse event dataset. In future research, other topic modeling algorithm will be explored. The FAERS Adverse Event Databases uses MedDRA codes [86] and terminology to standardize adverse effects such as using “pyrexia” instead of “heightened temperature” of “fever”. Yet, the number of possible adverse effects is too large and the resulting feature space of using bag-of-words semantics to represent adverse effects is too high dimensional. To address this issue, we acknowledge that adverse effects are symptoms of unknown (latent) underlying causes. While one way of identifying causes is involving a medical expert, we propose a data-driven approach to identify underlying topics among adverse events using topic modeling that we interpret as causes. For that, we employ Latent Dirichlet Allocation (LDA) [23] – a generative probabilistic model which assumes that each adverse event is a mixture of underlying (latent) topics, and each topic has a (latent) distribution of more and less likely adverse effects.

3.3.3 Similarity Test

We measure similarity between adverse effects and of spatial regions, we use text similarity measures. The FAERS ADEs databases uses MedDRA codes and terminology to standardized AEs terms such as using "nausea" instead of "feeling queasy". These terms follow medical nomenclature such as the term "hemoglobin","hemophilia", "hemorrhage", "hemorrhoids" all relate to blood due to the prefix "Hemo", which relates to blood. Therefore we had to use a similarity algorithm, which uses pattern matching as compare to string matching to score similarity.

The Ratcliff/Obershelp Pattern Recognition algorithm also known as Gestalt Pattern Matching was introduced in 1983 by Ratcliff and Obershelp [87]. It computes the similarity between two sets of Top-k most frequent adverse effects by finding the longest contiguous common matching sequence or part of the string and repeatedly, matching characters in the unmatched region on either side of the longest common part of the string.

We measure similarity between LDA topics by using Euclidean distance. Sets of adverse events corresponding to two regions at time. We describe each such set as the mean of latent features within the set, and measure the Euclidean distance in the latent feature space.

3.3.4 Finding Spatial Clusters

Distances generated from the similarity Index Coefficients are used to find clusters. The closer the similarity coefficients, the more robust the clusters generated from hierarchical agglomerative clustering. Hierarchical agglomerative clustering was used in our study because it treats each observation as a separate cluster and builds the cluster from a bottom-up approach. The closer the similarity scores of each state to another state, the closer the top (k) frequent itemset of the selected pharmaceutical drug adverse event. Our approach is that adverse events of each country will resemble more the countries that are closer in geographic space unless there are other underlying factors such as supply chain issues, environmental, health insurance coverage, and economic factors.

3.3.5 Spatial Auto-correlation Test

First Law of Geography states that everything is related to everything else. But near things are more similar than distant things [88]. We will test if the clustering output from hierarchical agglomerative clustering has any spatial tendencies. The output of hierarchical agglomerative clustering is categorical since it's the grouping of countries or US states by the similarity of coefficients of top (k) adverse events.

This chapter describes the roadmap that were utilized in this dissertation in detail. Each sub-sections explains the method used for combining spatial similarity search with

topic modeling and mining itemsets, and the database our approached were applied to. These algorithms are explained further in Chapter 4, 5, and 6 in the form of individual papers.

Chapter 4: Clustering of Adverse Events for Post-Market Approved Drugs using Frequent Itemset

Abstract

Adverse side effects of a drug may vary over space and time due to different population, environment, and drug quality. Discovering all side effects during the development process is impossible. Once approved and available to the public, regulators rely on a combination of surveillance, reporting (by doctors and patients), and data mining to discover any post-market issues in approved pharmaceutical drugs. Our goal of this study is to find statistically significant spatio-temporal clusters among co-occurring adverse effects of U.S Food and Drug Administration (FDA) approved drugs. We use data obtained from FDA's Adverse Event Reporting System (FAERS) to explore the spatio-temporal distribution of combinations of adverse effects. This is done by computing, for each spatial region and for each year, the top k most frequent sets of adverse events using a frequent itemset mining approach. To assess the similarity of sets of adverse events between spatial regions, we employ Gestalt Pattern Matching between the textual representation of reported adverse effects. To find clusters of regions that exhibit similar adverse events we apply an agglomerative hierarchical clustering approach. Finally, we explore the resulting clusters of similar adverse events to discover patterns of spatial autocorrelation using Moran's I measure of spatial autocorrelation. In our experimental evaluation, we use adverse event records in Europe for three pharmaceutical drugs between 2014 and 2017. Our result show that the vast majority of mined clusters of regions having similar adverse events did not exhibit significant spatial auto-correlation, indicating that the adverse events within a clusters are not the result of spatial patterns or local effects. For a small number of clusters, we found significant spatial autocorrelation but after applying Bonferroni correction to account for

the large number of tested hypotheses, we found no significant and interesting cases of spatial autocorrelation for three drugs studied. Yet, we note that our approach can be applied on other drugs to explore if other drugs may exhibit spatially localized side-effects that may justify further investigation. We conclude our work by discussing future directions to discover spatial trends among adverse events that this study was not able to find.

4.1 Introduction

Public health surveillance is the base of effective public health practice [89], and it has been put in the spotlight due to the coronavirus disease epidemic, which has impacted aspects of public health governance in its response and recovery [90]. Pharmacovigilance is a branch of public health surveillance and is the detection, assessment, understanding, and prevention of adverse effects [8, 9]. We propose an approach to find spatial autocorrelation among adverse effects using data mining and spatial statistics to support pharmacovigilance.

Adverse Events (AEs) are any undesirable experiences associated with the use of a medical product. The United States Food and Drug Administration (FDA) has developed an Adverse Event (AE) self-reporting tool called MedWatch. Patients, practitioners, and drug manufacturers can all report adverse events to FAERS. It is not limited to only allergic reactions but also allows to report issues such as product use errors, product quality problems, and therapeutic failures can all be reported via MedWatch. This information is available to everyone including researchers, patients and health practitioners [7]. Currently, in FAERS, there are 24 million reports and growing.

Information sharing is constraint by the HIPPA law. HIPAA (Health Insurance Portability and Accountability Act of 1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information. The law has emerged into greater importance recently due to the health data breaches caused by cyber-attacks and ransom-ware attacks on health insurers and providers. This law also puts a constraint on researchers to safeguard personally identifiable information (PII) that could potentially

identify a specific individual. Any information that can be used to distinguish one person from another can be considered as PII. In the spatial data mining setting, a person's location is also data, which can be classified as PII and must be de-identified [70].

AEs in the FAERS database are coded using terms in the Medical Dictionary for Regulatory Activities (MedDRA) terminology to standardized medical terminology. Standardizing AE keywords helps facilitate the sharing of information by regulatory authorities, pharmaceutical companies, clinical research organizations and health care professionals and allows for better global protection of public health [86]. Reporting systems such as FAERS are critical tools for monitoring the safety, efficacy, and quality standards of approved pharmaceutical drugs. FDA uses FAERS for postmarket surveillance of approved drugs when investigating safety concerns [7].

A major challenge in vaccine/drug development is determining possible side effects. Recent analysis found that it took a median of 4.2 years after a drug's initial approval for major safety concerns to be discovered [2]. Serious side effects could be life-threatening, which can lead to death. While less severe AEs such as rash, nausea, and fatigue might not be dangerous, however, they can lead to avoidance in taking the drug as prescribed, which can lead to a severe consequence [3]. Our motivation for this study is to identify co-occurin AEs. For this purpose, we first survey existing work in Section 4.2 and define the problem of spatio-temporal clustering of adverse events in Section 4.3. Then, we propose our approach for clustering regions having similar adverse event sets in Section 4.4. We apply our approach to three common anticoagulant drugs in Section 4.4, 4.5 and conclude in Section 4.6.

4.2 Related Work

Data mining or knowledge discovery from a database (KDD) is used to determine useful, implicit and hidden patterns in a large dataset which was previously unknown. It used in a wide range of disciplines such as public health [19]. Market Basket Analysis is one

of many data mining techniques used usually by large retailers to uncover associations between items [20]. It works by looking for combinations of items in each transaction that frequently occur together. It allows market researchers to identify relationships between the items that customers buy, such as milk, bread and diapers [21]. Frequent Item-set and association rules mining are not restricted to market basket analysis, but instead, they can be applied in other settings [22]. We leverage frequent itemset mining to represent the adverse events reported in a spatial region as a set of mined frequent adverse effects. Then, we use this representation to cluster regions having similar frequent adverse effects for the same drug at the same time.

Almost 80% of Americans over the age of 60 are taking multiple drugs, and the occurrence and intensity of an adverse events increases with the number of drug combinations. Clinical trials typically focus on a single drug and rarely on drug to drug combinations due to the sheer amount of multiple drug combination of all FDA approved drugs. Such an enormous number of drug combinations and possible AEs from drug to drug interaction make it challenging and is needed for patient safety [68]. Zitnik et al use convolutional network to predict for multiple drug to drug AEs. Using neural network they construct for protein to protein interactions, drug to protein target interactions and drug to drug interactions to predict for AEs for multiple drug combination, which have not been used yet with patients [48]. Adverse side effects of a single drug or multiple combination of drugs may vary over space and time due to different population dynamic [10, 11, 12], environment [13, 14], or drug quality [15, 91]. In this paper, we explore spatial temporal clusters of a three FDA approved drugs for post market AEs patterns and trends. We didn't used concomitant drugs of FAERS dataset and used AEs reports associated to a single drug for Dabigatran, Rivaroxaban and Apixaban. Previous cohorts studies in the US and in Norway compared these dabigatran, rivaroxaban, and apixaban. The cohort study in US concluded that these three drugs appear to have similar effectiveness, although apixaban may be associated with a lower bleeding risk and rivaroxaban may be associated with an elevated bleeding risk. The Cohort in Norway reached similar findings; dabigatran and apixaban were both associated

Table 4.1: Table of Notations

Variable	Description
$\mathcal{A}$	The domain of all adverse effects
$A \subseteq \mathcal{A}$	A set of adverse effects
$\mathcal{S}$	The domain of all spatial regions
s	A single spatial region
$\mathcal{T}$	The domain of all discrete time intervals (commonly: years)
$t \in \mathcal{T}$	A time interval
$\mathcal{D}$	The domain of pharmaceutical drugs
$d \in \mathcal{D}$	A pharmaceutical drug
$\mathcal{DB}$	A database of Adverse Events
$(t, s, A, d) \in \mathcal{DB}$	An adverse event (AE).
$\mathcal{DB}_t$	Adverse events reported during time t
$\mathcal{DB}_{s,t}$	Adverse events reported in region s during time t
$TopkFAE(DB_{s,t})$	Top-k frequent adverse effects among adverse events in $DB_{s,t}$
$Gestalt(TopkFAE(DB_1), TopkFAE(DB_2))$	Gestalt similarity between Top-k frequent adverse events
$dist(TopkFAE(DB_1), TopkFAE(DB_2))$	Distance function between Top-k frequent adverse events
$LCSS(s_1, s_2)$	The longest common subsequence between two strings

with significantly lower risk of major bleeding compared with rivaroxaban [53, 52]. Our goal is to add to this literature using spatial datamining on FAERS report and reporting any significant colocated AEs.

4.3 Problem Definition

This section formally defines the problem of spatio-temporal clustering of adverse events. A summary of all notations used in this work is found in Table 4.1. First, we provide a definition of adverse effects and events.

Table 4.2: Sample records of Adverse Event Report Database. Each Line is an Adverse Event.

Adverse Event ID	Set of Adverse Effects	Location	Event Time	Drug
109947323	Abdominal pain, Abdominal pain upper, Constipation, Diarrhoea, Headache, Heart rate increased, Nausea, Pain in extremity, Vertigo, Vomiting	Germany (DE)	9/24/2014	Rivaroxaban
106823542	Duodenal ulcer haemorrhage, Gastric ulcer haemorrhage, Shock haemorrhagic	Netherlands (NL)	11/2/2014	Rivaroxaban
109449521	Death	United Kingdom (GB)	2/12/2015	Rivaroxaban
120813061	Asthenia, Haemorrhage	Croatia (HR)	1/16/2016	Rivaroxaban
145539611	Purpura, Skin exfoliation, Skin lesion	United Kingdom (GB)	12/28/2017	Rivaroxaban

Definition 1 (Adverse Effect)**.** *An Adverse Effect is a textual representation of an undesirable experiences associated with the use of a medical product. We let $\mathcal{A}$ denote the set of all adverse events.*

Data such as collected in the FAERS database is a collection of records each associated with a set of adverse effects, a specific pharmaceutical drug, a location, and time. We call such as record an Adverse Event (AE), formally defined as follows:

Definition 2 (Adverse Event Database)**.** *Let $\mathcal{A}$ denote a set of adverse effects, let $\mathcal{S}$ denote a set of spatial regions, let $\mathcal{T}$ denote a set of time intervals (such as years), and let $\mathcal{D}$ denote a set of drugs. An Adverse Event Report Database $\mathcal{DB}$ is a collection of adverse event reports (t, s, A, d), where $t \in \mathcal{T}$ is a point in time, $s \in \mathcal{S}$ is a spatial region, $A \subseteq \mathcal{A}$ is a set of adverse effects, and $d \in \mathcal{D}$ is the drug for which the adverse effects are reported.*

We note that a single adverse event may report multiple adverse effects. As an example,

Table 4.2 shows exemplary adverse events from the FAERS database. The first line in Table 4.2 implies that "Abdominal pain, Abdominal pain upper, Constipation, Diarrhoea, Headache, Heart rate increased, Nausea, Pain in extremity, Vertigo and Vomiting" are adverse effects that occurred on on 9/24/2014 in Germany for drug Rivaroxaban.

Our goal is to find clusters of locations that, at a given time, exhibit similar adverse events. Towards this goal, we group adverse events by region and time. Further, we abstractly define a similarity measure between the adverse events at a given region at a given time.

Definition 3 (Spatio-Temporal Adverse Events)**.** *Let $\mathcal{DB}$ be an adverse event report database, let $s' \in \mathcal{S}$ be a spatial region, and let $t' \in \mathcal{T}$ be a time interval. We define*

$$\mathcal{DB}_{s',t'} := \{(t, s, A, d) \in \mathcal{DB} | t = t' \wedge s = s'\}$$

as the set of all adverse events reported at time t' at location s'. For two spatial regions s_1 and s_2, we let

$$dist(\mathcal{DB}_{s_1,t}, \mathcal{DB}_{s_2,t}) \mapsto [0, 1]$$

denote an abstract distance function between two sets of adverse events.

We propose a concrete implementation of $dist()$ in Section 4.4.3. Given a distance function to assess the adverse event similarity of two regions at the same time and for a given drug, we define a spatial adverse event clustering as follows:

Definition 4 (Spatial Adverse Event Clustering)**.** *Let $\mathcal{DB}$ be an adverse event report database, let $dist()$ be a distance function to measure dissimilarity among sets of adverse events. Further, let*

$$\mathcal{DB}_t := \{x \in \mathcal{DB} | x.t = t\}$$

denote the set of all adverse events reported at time t. A spatial clustering $\mathcal{C}(\mathcal{DB}_t) = \{C_1, ..., C_K\}$ is a partition of regions S such that $\forall i : C_i \subseteq \mathcal{S}$ and $\forall i \neq j : C_i \cap C_j = \emptyset$. As element $C \in \mathcal{C}$ of a spatial adverse event clustering is called a spatial adverse event cluster.

Problem Statement 1 (Spatial Adverse Event Clustering)**.** *Our first goal is to find a good spatial adverse event clustering, that is, a clustering such that the similarity of regions in the same cluster is maximized, while the similarity of regions across clusters is minimized.*

Once a good spatial adverse event clustering is found, our second goal is to explore the spatial auto-correlation of these cluster to answer the question if some regions exhibit similar adverse effects at certain times, or if adverse effects are independent of space and time.

Problem Statement 2 (Spatial Auto-correlation of Clusters)**.** *Our second goal is to explore the spatial auto-correlation among spatial adverse event clusters.*

4.4 Methodology

This section describes our approach to find spatio-temporal clusters of AEs and to find those having significant spatial autocorrelation. And overview of this section is found in Figure 4.1. First, we obtained the data from the FAERS Database using the openFDA API [75, 74] as described in Section 4.4.1. To concisely represent the AEs reported in a region, Section 4.4.2 presents our approach to extract frequent sets of adverse effects from AEs reported in a region. Our approach leverages an itemset-mining approach similar to the Apriori algorithm [83] in which we treat AEs as transactions, and adverse effects as items within a transaction. Having each region (at a specified time) represented by local sets of frequent sets of adverse effects, we propose a similarity measure between regions in Section 4.4.3. This similarity measure uses Gestalt Pattern Matching [87] to estimate the set-similarity between the frequent sets of adverse effects. Using this similarity measure to to define the abstract function $dist()$ in Definition 3, we proceed to present our region clustering approach in Section 4.4.4 to group regions by similar frequent sets of AEs and to solve Problem Statement 1. To find clusters that exhibit significant spatial auto-correlation, we apply Moran's I test statistic to measure spatial auto-correlation as described in Section 4.4.5. All of our algorithms used for our data analysis have been implemented in

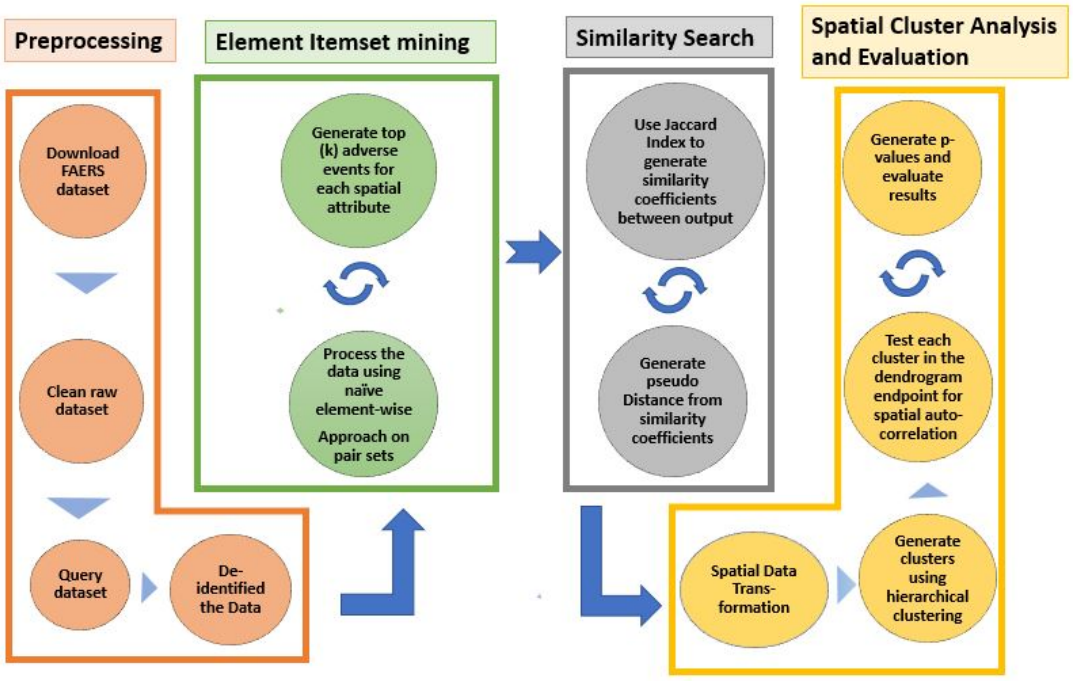


Figure 4.1: Road map for spatiotemporal association model using Adverse Events Report Submission to FDA FAERS Database.

Python 3.7 and are published on GitHub as a Jupyter Notebook for reproducibility at https://github.com/ahmedaskar64/Spatio-Temporal-Clusters-AEs-Post-Market.

4.4.1 Data Collection

As of 6/01/2022, there are 24 million AEs report for countless drug combinations in the FAERS database, we implemented a data crawler in Python to download these reports from the FAERS public-facing database published quarterly as a zip file available at fda.gov [74] or by OpenFDA API [75]. FAERS data comes in multiple files with a primary key linking all files. The files include a demographic and administrative information file; drug information of the case reported; reaction information from the reported case along with patient outcome information and the source of the report. Spatial attributes are aggregated to the country level for the public-facing FAERS database due to HIPAA regulations. Due to

this aggregation, we focus on clustering of adverse effects using a study area of European countries as shown in Figure 4.2 and in Table 4.3 as a proof of concept due to data availability for a pharmaceutical drug ($d \in \mathcal{D}$) and Top-k selection. We note that our solutions could also be applied other study areas, such as the US on state or zip-code level, if such data is available. We used Europe for our study area due to the richness in the data for spatiotemporal attributes along with closeness in regulatory policies across Europe due to EU overarching regulatory arm.

Once the data was collected, we joined all datasets by their primary key to obtain one large data table, including for each case report a spatial attribute and a timestamp of the observed AE. While spatial attributes are aggregated at the country level case report have detailed timestamp of the event date. We group this dataset by year between 2014 to 2017. Data varied for different countries depending on the pharmaceutical drug availability, prescribers' preferences, pharmaceuticals marketing, supply chain, etc. In the exploratory data selection phase, we were limited to select pharmaceutical drugs which were used consistently throughout Europe and didn't have temporal usage variability as well. We looked for a pharmaceutical drug that had general AE rates across Europe, so our results were not skewed. Rivaroxaban was selected as our pharmaceutical study drug due to the spatiotemporal variance of reported AEs across Europe along with it being one of top ten most reported pharmaceutical drug in FAERS Database. Since Rivaroxaban is an anticoagulant drug, we used other similar anticoagulant drugs such as Dabigatran and Apixaban as our other drugs in our study. There is ample about of literature comparing these three drugs [53, 52]. Rivaroxban and Dabigatran received approval to market from European Medicines Agency (EMA) in 2008 [92, 93] and Apixaban in 2011 [94]. It takes a median of 4.2 years after a drug's initial approval for major safety concerns to be discovered [2]. Figure 4.3a, 4.3b, 4.3c is a count of the total AEs submissions to FAERS for Europe. Report submission increased few years after the initial approval.

Since many patients might take combination of other medications, we were only interested in the AE reports submitted of patients with only using one drug at a time of report

Table 4.3: Study areas for Drugs Rivaroxaban, Dabigatran and Apixaban for $k \in \{5, 10\}$.

Countries in the study area	# of countries	Drug Name	k
Austria(AT), Belgium(BE), Germany(DE), Denmark(DK), Spain(ES), Finland(FI), France(FR), United Kingdom(GB), Greece(GR), Hungary(HU), Ireland(IE), Italy(IT), Netherlands(NL), Norway(NO), Poland(PL), Portugal(PT), Sweden(SE), Turkey(TR)	18	Dabigatran	5
Austria(AT), Belgium(BE), Germany(DE), Denmark(DK), Spain(ES), Finland(FI), France(FR), United Kingdom(GB), Italy(IT), Netherlands(NL), Poland(PL), Portugal(PT), Sweden(SE)	13	Dabigatran	10
Austria(AT), Germany(DE), Spain(ES), France(FR), United Kingdom(GB), Greece(GR), Italy(IT), Norway(NO), Sweden(SE)	9	Apixaban	5
Austria(AT), Germany(DE), Spain(ES), France(FR), United Kingdom(GB), Greece(GR), Italy(IT), Norway(NO), Sweden(SE)	9	Apixaban	10
Austria(AT), Belgium(BE), Switzerland(CH), Germany(DE), Denmark(DK), Spain(ES), Finland(FI), France(FR), United Kingdom(GB), Greece(GR), Hungary(HU), Ireland(IE), Italy(IT), Netherlands(NL), Norway(NO), Poland(PL), Sweden(SE), Slovenia(SI), Turkey(TR)	19	Rivaroxaban	5
Austria(AT), Belgium(BE), Switzerland(CH), Germany(DE), Denmark(DK), Spain(ES), Finland(FI), France(FR), United Kingdom(GB), Greece(GR), Hungary(HU), Ireland(IE), Italy(IT), Netherlands(NL), Norway(NO), Poland(PL), Sweden(SE), Slovenia(SI), Turkey(TR)	19	Rivaroxaban	10

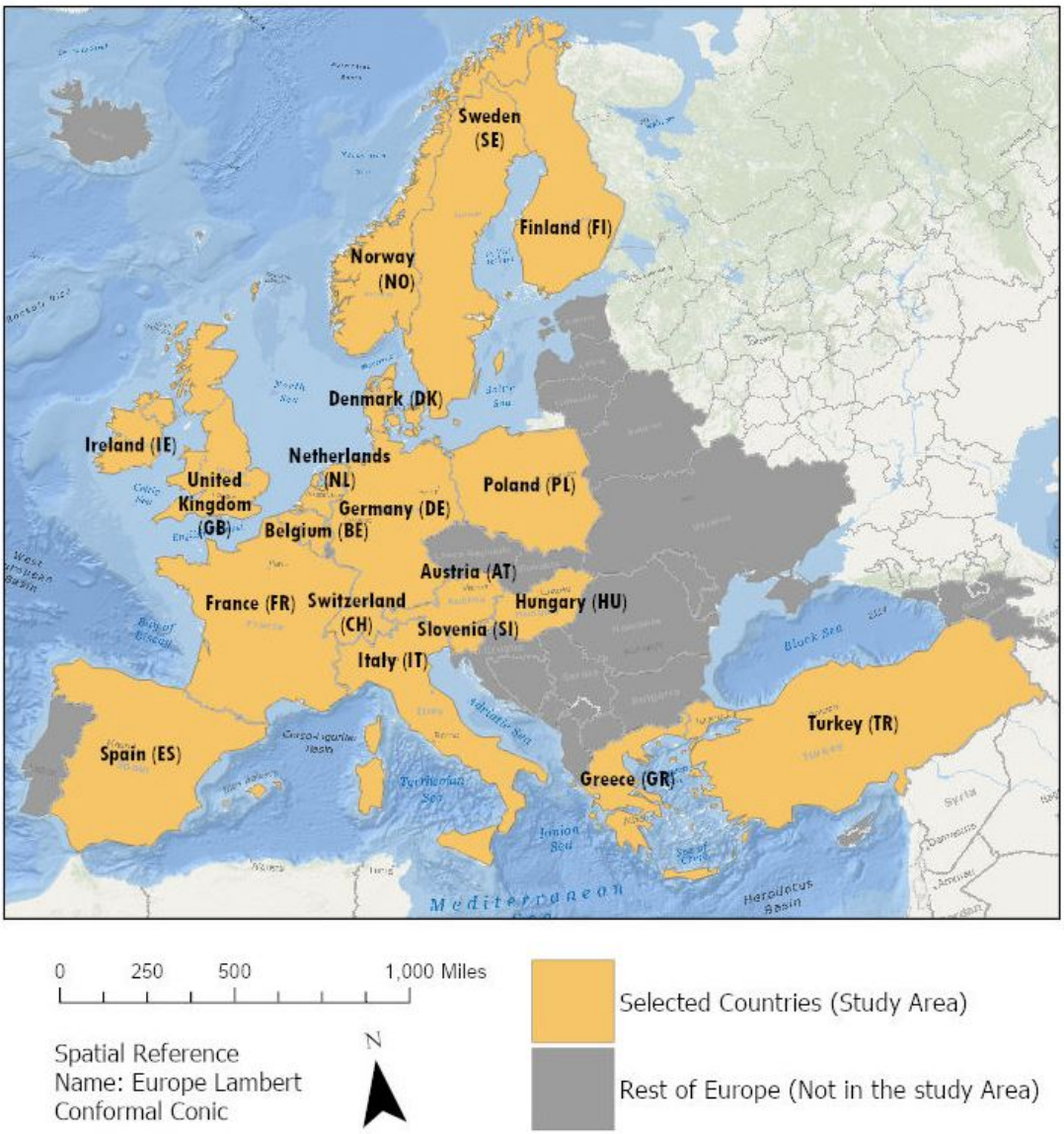
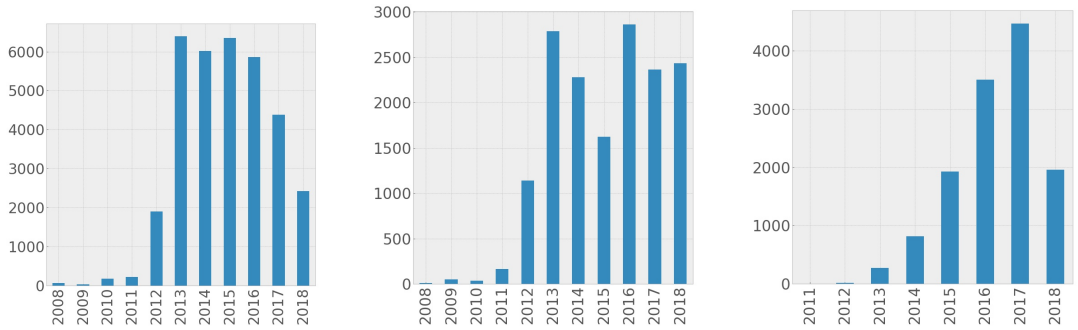


Figure 4.2: Study area for analyzing Top 5 & 10 unique AEs for Rivaroxaban, Apixaban and Dabigatran from year 2014 to 2017 for European countries listed with their ISO alpha-2 code

submission as compare to concurrent medication use , which might pose added challenge of not knowing which medication to subscribe the reported AEs. While in this work, we only consider usage of a single drug, we note that our work can be easily extended to compare



(a) Xarelto (b) Pradaxa (c) Eliquis

Figure 4.3: Adverse Events Report Submission to the FDA FAERS Database per drug per year

concurrent drug usage to explore the effects of polypharmacy AEs [48].

Once data is collected, our goal is to cluster countries by their AEs set similarity in Section 4.4.4. To accomplish this task, we need:

- 1) an efficient way to extract the most frequent sets of adverse effects from large collections of AEs reported in each region in Section 4.4.2
- and 2) a distance (or dissimilarity) measure between sets of AEs to quantify similarity for clustering as proposed in Section 4.4.3.

4.4.2 Frequent (k) Adverse Event Set Mining

We propose to represent a potentially large set of AEs in a region by the Top-k most frequent adverse effects reported in that region. We define the set of Top-k Frequent AEs as follows:

Definition 5 (Top-k Frequent Adverse Event)**.** *Let $\mathcal{DB}$ be an adverse event(AE) report database, then $\mathcal{DB}_{\mathcal{R},\mathcal{Y}}$ denote the database having adverse events(AEs) only in region $\mathcal{R}$ and in year $\mathcal{Y}$. Let $TopkFAE(DB_{R,Y})$ denote the Top-k most frequent sets of AEs in $DB_{R,Y}$ defined as follows:*

$$TopkFAE(DB_{R,Y}) := arg\,max^{k}_{A \subseteq \mathcal{A}} |\{AE \in DB_{R,Y} | A \in AE\}|, \tag{4.1}$$

where $arg\,max^{k}_{a \in args} f(x)$ *returns the set of* k *arguments that maximize function* $f(x)$ *(in this case, the count of sets of adverse effects* $A \subseteq \mathcal{A}$ *among all AEs* $DB_{R,Y}$*), and* $|\cdot|$ *is the set-cardinality function that returns the support (number of occurrences) of a set of AEs.*

A naive implementation requires to enumerate all (combinatorically many) subsets of $\mathcal{A}$, to find the subsets having the k highest support. To mine this set of Top-k most frequent adverse effects more efficiently, we leverage a variant of the classic Apriori algorithm [83], which mines Top-k frequent itemsets in transaction databases. The classic Apriori algorithm returns all itemsets having sufficiently large support. The variant we use [95] instead returns the Top-k most frequent itemsets. This approach scans the database and computes the support of every item and sorts items by their frequency in descending order. Then it iteratively scans database again to construct a frequent pattern tree, update the frequency count and use this to prune the tree to avoid evaluating itemsets whose support must be lower than the support of at least k other itemsets.

The choice of parameter k is important. If k is chosen too low, most countries may have similar Top-k AEs. Thus, it will no longer possible to discriminate differences if k is chosen too low. If k is chosen too large, then a large set of rare AEs may create noise that may cause regions to appear to more dissimilar. Our experiments have shown that values of $k \in \{5, 10\}$ provide a good balance of discrimination and resistance to noise. While our study area is Europe, countries available for analysis change due to data availability. For $k = 5$ we require at least five unique adverse effects each year and for $k = 10$ we require at least ten unique adverse events. Given the three selected pharmaceutical drug and $k \in \{5, 10\}$ for our framework model run, Table 4.3 shows the available countries for our analysis. Table 4.4 shows an output of mined adverse effects for Austria using Equation 4.1, with $k = 10$ for year 2016. Adverse effects are ordered by frequency (relative number of adverse events that the adverse effect appears in). Note that in this example, there are sets

Table 4.4: Example k-most frequent sets of adverse events in Austria in Year 2016 for Drugs Rivaroxaban and Dabigatran. For each set the support among adverse events of that year in Austria is provided.

k=10 Itemsets|min support	Year	Drug
({'Death'}, 0.125), ({'Cerebral haemorrhage'}, 0.125), ({'Cerebrovascular accident'}, 0.125), ({'Haemorrhage'}, 0.0937), ({'Pain'}, 0.0937), ({'Atrial thrombosis'}, 0.0937), ({'Arterial thrombosis'}, 0.0625), ({'Acute coronary syndrome'}, 0.0625), ({'Pain', 'Arterial thrombosis'}, 0.0625), ({'Acute coronary syndrome', 'Arterial thrombosis'}, 0.0625)	2016	Rivaroxaban
({'Gastrointestinal haemorrhage'}, 0.333), ({'Acute kidney injury'}, .333), ({'Cholecystitis'}, 0.167), ({'Haematochezia'}, 0.167), ({'Hemiparesis'}, 0.167), ({'Hepatic failure'},0.167), ({'Overdose'}, 0.167), ({'Hepatic failure', 'Gastrointestinal haemorrhage'}, 0.167), ({'Haematochezia', 'Gastrointestinal haemorrhage'}, 0.167), ({'Cholecystitis', 'Acute kidney injury'}, 0.167)	2016	Dabigatran

of singular (having exactly one element) and multiple frequent adverse effects. In general, for $k \geq 10$, an itemset may include multiple adverse effects. We used the Pandas Python library [96], numpy, and mlxtend Python library to carry out the data analysis [77, 97].

4.4.3 Similarity Measure between Sets of Adverse Events

To measure similarity between the Top-k most frequent adverse effects $TopkFAE(DB_{R_1,Y})$ and $TopkFAE(DB_{R_2,Y})$ of two countries R_1 and R_2 in a year Y, we use text similarity measures. The FAERS AEs databases uses MedDRA codes and terminology to standardized AEs terms such as using "nausea" instead of "feeling queasy". These terms follow medical nomenclature such as the term "hemoglobin","hemophilia", "hemorrhage", "hemorrhoids" all relate to blood due to the prefix "Hemo", which relates to blood. Therefore we had to use a similarity algorithm, which uses pattern matching as compare to string matching to score similarity.

The Ratcliff/Obershelp Pattern Recognition algorithm also known as Gestalt Pattern Matching was introduced in 1983 by Ratcliff and Obershelp [87]. It computes the similarity between two sets of Top-k most frequent adverse effects $TopkFAE(DB_{R_1,Y})$ and $TopkFAE_(DB_{R_2,Y})$ by finding the longest contiguous common matching sequence or part of the string and repeatedly, matching characters in the unmatched region on either side of the longest common part of the string.

$$Gestalt(s_1, s_2) = \frac{2 \cdot LCSS(s_1, s_2)}{|s_1| + |s_2|} \tag{4.2}$$

where in $LCSS(s_1, s_2)$ is the longest common subsequence [98].

In our analysis, we used Gestalt Pattern Matching to compare the AEs sets from the results of the Top-k to one another in our study area.

$$Gestalt(TopkFAE(DB_{R_1,Y}), TopkFAE(DB_{R_2,Y})) =$$

$$\frac{2 \cdot LCSS(TopkFAE(DB_{R_1,Y}, TopkFAE(DB_{R_2,Y})}{|TopkFAE(DB_{R_1,Y}| + |TopkFAE(DB_{R_2,Y}|} \tag{4.3}$$

Equation 4.4.3 yields a similarity score in the range from 0 to 1. The closer to 1, the more similar set of AEs are from Top-k output.

Next, we use the Gestalt Pattern Matching similarity function in Equation 4.4.3 to define a dual distance function by substracting from one in Equation 4.4.3.

$$dist(TopkFAE(DB_{R_1,Y}), TopkFAE(DB_{R_2,Y})) =$$

$$1 - Gestalt(TopkFAE(DB_{R_1,Y}), TopkFAE(DB_{R_2,Y})) \tag{4.4}$$

In each step of the similarity computation, each drug Top-k AEs ($TopkFAE(DB_{R_1,Y})$) is paired with another set from another country Top-k event($TopkFAE(DB_{R_2,Y})$). We

compute the similarity of AEs output against countries in our set for that calendar year. In another words, similarity is computed spatially and filtered temporally.

	1	2	3	4	5	6	7	8		9	10	11	12	13	14	15	16	17	18	19
S1	c	e	r	e	b	r	a	l		i	s	c	h	a	e	m	i	a		
S2	c	e	r	e	b	r	a	l		h	a	e	m	o	r	r	h	a	g	e

Example 1. *Another way of describing 'Cerebral haemorrhage' is 'Cerebral ischaemia'. These terms are semantically similar however a simple test of textual equality would not be able to capture such partial similarity as compare to Gestalt pattern, which scores the similarity at 0.72 as follows. We first obtain the length of these two strings as* $|s_1| = 17$ *and* $s_2 = 19$. *The longest common subsequence (LCSS) is "cerebralhaema" having a length of 13 characters. Thus, we get* $LCSS(\text{'Cerebral haemorrhage','Cerebral ischaemia'}) = 13$. *And we can compute the Gestalt matching similarity (Equation 4.4.3) as:*

$$Gestalt(\text{"}cerebralischaemia\text{"}, \text{"}cerebralhaemorrhage\text{"}) =$$

$$\frac{2 \times 13}{17 + 19} = 0.72$$

To obtain a distance instead of a similarity measure, we use Equation 4.4.3 as follows:

$$dist(\text{"}cerebralischaemia\text{"}, \text{"}cerebralhaemorrhage\text{"}) =$$

$$1 - Gestalt(\text{"}cerebralischaemia\text{"}, \text{"}cerebralhaemorrhage\text{"}) = 0.28$$

As an another example, let's look at an example of concrete Top-5 sets of AEs for two countries.

Example 2. *Table 4.5 shows Austria and Germany Top-5 frequently mined AEs in year 2014 for drug-Rivaroxaban. Equation 4.4.3 Gestalt Pattern Matching is computed between countries for that same calendar year and the same drug.*

The strings $TopkFAE(DB_{Austria,2014})$ *and* $TopkFAE(DB_{Germany,2014})$ *are obtained through concatenation of the adverse effects shown in Table 4.5. We first obtain the length of these two strings as* $|TopkFAE(DB_{Austria,2014})| = 92$ *and* $|TopkFAE(DB_{Germany,2014})| = 84$. *The LCSS between the two string is 'Cerebralaraer infctieos aie', having a length of 27 characters. Thus, we get* $LCSS(DB_{Germany,2014}, DB_{Austria,2014})$*=27, then we compute the Gestalt matching similarity (Equation 4.4.3) as:*

$$Gestalt(TopkFAE(DB_{Austria,2014}), TopkFAE(DB_{Germany,2014})) = \frac{2 \cdot 27}{92 + 84} = 0.307$$

Table 4.5: Top-5 Frequent Adverse Event between 2 countries (Germany and Austria) for Rivaroxaban in year 2014

Drug	Country	Top (5) AEs for Yr 2014
Rivaroxaban	Germany	'Fall', 'Anaemia', 'Cerebral haemorrhage', 'Cerebral infarction', 'Cerebrovascular accident'
Rivaroxaban	Austria	'Cerebrovascular accident', 'Drug ineffective', 'Haemoptysis', 'Anaemia', 'Diarrhoea'

4.4.4 Finding Spatial Clusters

The similarity scores in Equation 4.4.3 will be used as a distance variable in the clustering analysis. For this purpose, we propose to use a hierarchical agglomerative clustering approach [99]. The advantage of such approach is that we neither have to guess the number of clusters as often needed for partitioning clustering approaches [100] and we don't have to define a density threshold as required by density-based clustering algorithms [101] Distances generated from the similarity score in Equation 4.4.3 are used to find clusters. The closer the similarity coefficients, the more robust the clusters generated from hierarchical agglomerative clustering.

Hierarchical agglomerative clustering was used in our study because it treats each observation as a separate cluster and builds the cluster from a bottom-up approach. Hierarchical agglomerative clustering builds a dendrogram, which is a hierarchical structure that cluster points by their distances using complete linkage method to build the dendrogram [99]. Given a dendrogram, there are many approaches to finding clusters using different clustering techniques to compute the distance between points and clusters. We used the complete linkage method, which defines the distance between two clusters as the furthest neighbor pair-wise distance of points across the clusters. It tends to produce more compact clusters with similar diameters and can have a limitation for outliers. It is a method that assures that all items in a cluster are at a minimal distance from one another [102].

Our approach only considers the Top-k frequent itemsets of a region for clusters, but does not consider the spatial distance of regions. Therefore, in the next step, we investigate whether the adverse effect-based clusters exhibit significant spatial autocorrelation. There are many factors spatial neighbors can influence one another such as similarity in demographics, socioeconomic and public health policies. These factors and others such as pharmacogenetics, supply chain, distribution and health disparities contribute how AE present themselves in space and time. These clusters would help gauge if these AEs are random or can be explained by current research.

4.4.5 Spatial Auto-correlation Test

Tobler's First Law states that everything is related to everything else. But near things are more similar than distant things [88]. Our hypothesis states that Top-k frequently mined AEs will resemble other AEs items that are closer in geographic space or their spatial neighbors. We used the geopandas library for handling spatial attributes and Pysal library for spatial autocorrelation test [80, 103]. We tested if the clustering output from hierarchical agglomerative clustering were spatially autocorrelated. The output of hierarchical agglomerative clustering is categorical and is the cluster groupings of countries using hierarchical clustering analysis using the similarity coefficients *dist* calculated for Top-k AEs of our chosen pharmaceutical drug. A commonly used statistic that describes spatial autocorrelation is Moran's I. Moran's I is a correlation coefficient that measures the overall spatial autocorrelation or the correlation among values of a given variable in dependence on the relative locations between the spatial units [104, 103].

We used Moran I to test for autocorelation at two ends of statistical test tail. Moran I two tail test for similarity of a given value to its neighbors resulting in postive autocorrelation (a clustering effect) or dissimilarity of a given value to its neighbors concludes in negative autocorrelation(a dispersion effect). Queen contingency neighborhood is used in our analysis and neighborhood is considered if they share boundary. Given a set of clustering labels from the hierarchical clustering output in the previous section, it evaluates whether these clustering labels assigned to countries using the Similarity Distance (*dist*) have pattern expressed as clustering, random or dispersion. Along with Global Moran I computation using Pysal [103], Z score and P value are also calculated using the expected and observed Moran I values indicating statistical significance. For each cluster identified in hierarchical clustering output, a value of 1 is given and all others a value of 0. We test if these clusters are spatially autocorrelated.

Table 4.6: Statistically significant clusters of AEs for Rivaroxaban, Dabigatran and Apixaban for Year 2014-2017 in Europe for $k \in \{5, 10\}$.

Fig #	Spatial Pattern	P-value	P-value w Bonferroni correction	Moran's I	List of Countries (two-letter ISO code)	# of countries	Year	Drug Name	k AEs
	Clustered	0.07142	0.4207	0.35714	hu, ie, pl, pt	4	2014	Dabigatran	5
	Clustered	0.07257	0.4274	0.375	at, be, de, dk, es, fi, fr, gr, ie, it, nl, no	12	2017	Dabigatran	5
	Clustered	0.0664	0.3911	0.45	dk, es, fi, fr, gb, it, nl, se	8	2015	Dabigatran	10
	Dispersed	0.09587	0.5647	-0.55556	be, fr, it, pt	4	2016	Dabigatran	10
	Clustered	0.0664	0.3911	0.45	de, dk, es, fi, fr, gb, it, se	8	2017	Dabigatran	10
Fig. 4.4	Clustered	0.03818	0.4311	0.43452	at, be, ch, de, gb, gr, se	7	2014	Rivaroxaban	5
Fig. 4.5	Dispersed	0.02559	0.2889	-0.58333	at, dk, fi, gb, ie, no, si	7	2016	Rivaroxaban	5
	Dispersed	0.08266	0.9332	-0.46154	at, dk, fi, gr, ie, si	6	2015	Rivaroxaban	10
Fig. 4.6	Clustered	0.00176	0.0199	0.60417	de, dk, es	3	2016	Rivaroxaban	10
	Clustered	0.01181	0.2007	0.75	at, de, es	3	2016	Apixaban	10

4.5 Results

For our experimental evaluation we collected data from the FAERS database as described in Section 4.4.1, mined frequent adverse effect sets for each European country for each each year as described in Section 4.4.2, clustered the resulting sets of frequent adverse effects as described in Section 4.4.4 using the similarity measure described in Section 4.4.3, and tested for significant spatial auto-correlation of the resulting adverse effect clusters as described in Section 4.4.5. We repeated this process for for Adverse Events in the FAERS database for three anticoagulant medications - Apixaban, Rivaroxaban and Dabigatran grouped by four years (2014-2017) and using values for $k \in \{5, 10\}$. Our null hypothesis states that there are no spatial clusters and AEs appear at random. The results of our experimental evaluation for *selected* years, drugs, and clusters can be found in Table 4.6. This table shows only and

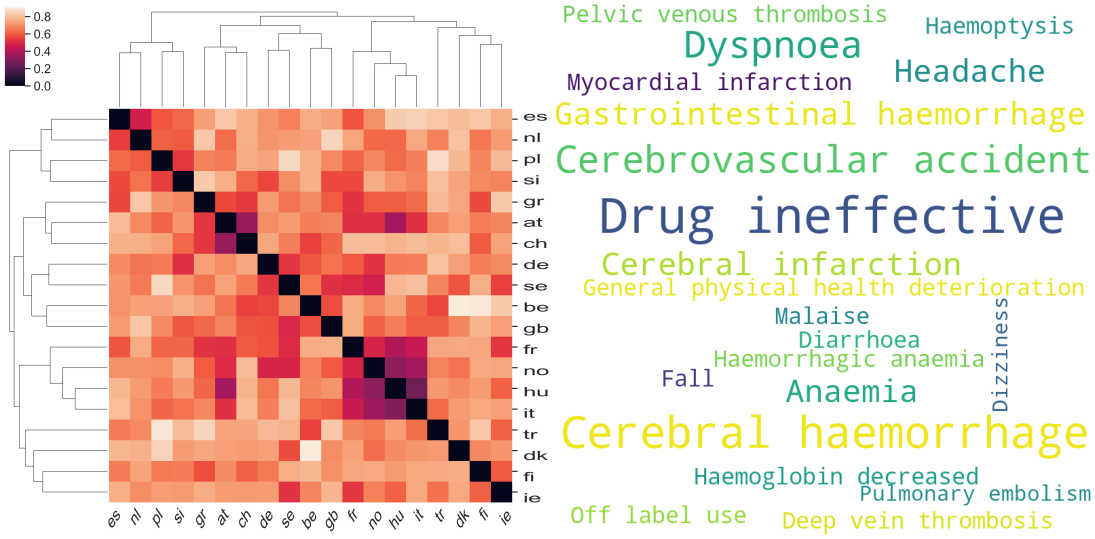


Figure 4.4: Results for Drug Rivaroxaban, $k = 5$, Year 2014. Distance matrix and hierarchical clustering (left). Adverse effect cloud weighted by frequency across the cluster of countries {at, be, ch, de, gb, gr, se} (right). Color of the text is an artifact of the python library used to generate word cloud.

all clusters having P-values less than 0.1 *before* Bonferroni correction [105]. Thus, for any combination of drug, parameter k, year, and country not shown in Table 4.6 the spatial pattern of the cluster including this country is not significantly clustered or dispersed. For three clusters using Drug Rivaroxaban we provide detailed results in Figures 4.4-4.6.

Qualitative Analysis

Figure 4.4 shows a cluster of countries having similar adverse events that is spatially autocorrelated in Year 2014 for $k = 5$, including Austria, Belgium, Switzerland, Germany, England, Greece and Sweden. The left of Figure 4.4 provides the distance matrix between all countries and indicates how the hierarchical clustering aggregated countries from Greece (Column/Line 5) to Great Britain (Column/Line 11). We also see (considering the shading of the submatrix at Column/Line 4-6) that countries Greece, Austria, and Switzerland have very similar adverse events. The cluster containing Countries Belgium, Great Britain,

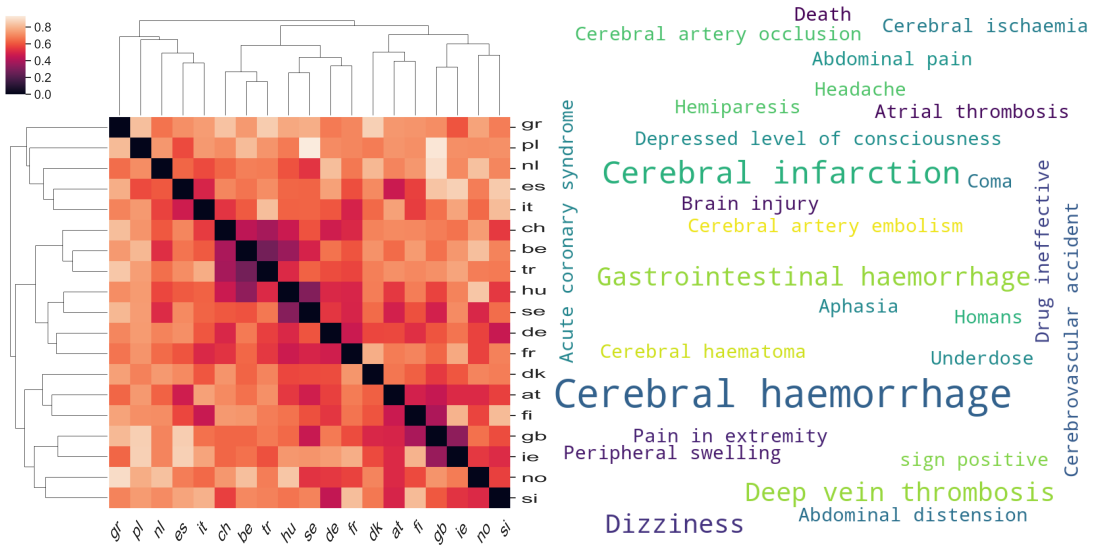


Figure 4.5: Results for Drug Rivaroxaban, $k = 5$, Year 2016. Distance matrix and hierarchical clustering (left). Adverse effect cloud weighted by frequency across the cluster of countries {at, dk, fi, gb, ie, no, si} (right). Color of the text is an artifact of the python library used to generate word cloud.

Sweden, and Germany was merged into this cluster to the high complete-link similarity, meaning that there no pairs of countries in this cluster having low similarity (we reiterate that we used complete-link clustering, see Section 4.4.4). We also observe a dense cluster of very similar adverse events between countries France, Norway, Hungary, and Italy, which was not included in Table 4.6 due to insignificant spatial auto-correlation. The right of Figure 4.4 shows a word cloud of all adverse effects among adverse events in this cluster. We see that a main adverse effect is “Drug Ineffective”which is less frequent in other clusters (see Figures 4.5 and 4.6).

The left of Figure 4.5 shows the distance matrix two years later for the same drug (Rivaroxaban) and parameter k (5). Interestingly, we observe very different adverse event clusters. Austria now clusters with Turkey, France, Iceland, and Finland (a cluster which does not have significant spatial auto-correlation) and Germany clusters with Denmark and Spain. The only cluster having significant auto-correlation in this case includes countries

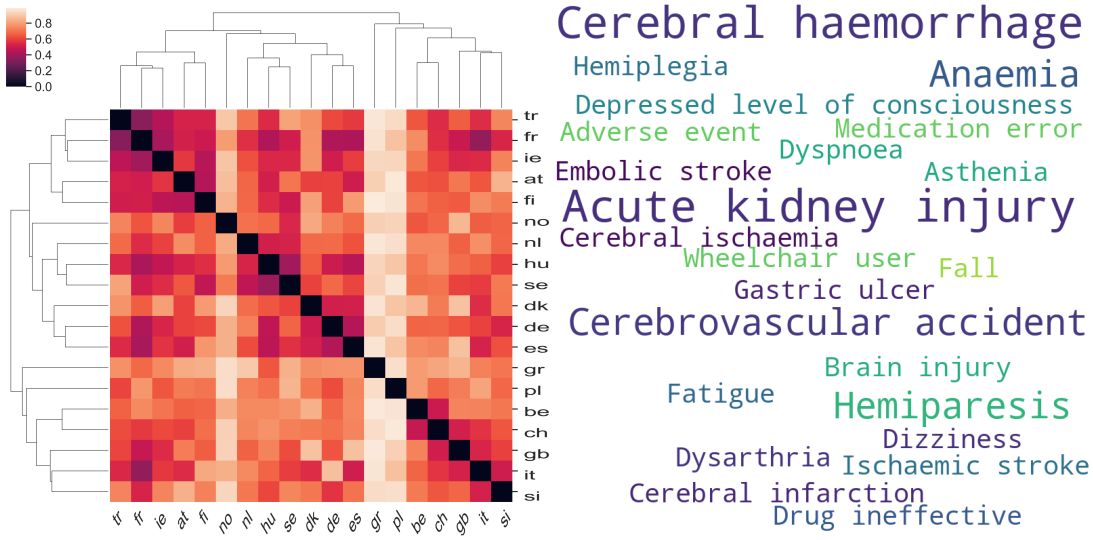


Figure 4.6: Results for Drug Rivaroxaban, $k = 10$, Year 2014. Distance matrix and hierarchical clustering (left). Adverse effect cloud weighted by frequency across the cluster of countries {de, dk, es} (right). Color of the text is an artifact of the python library used to generate word cloud.

Austria, Denmark, Finnland, Great Britain, Ireland, Norway, and Slovenia and exhibits negative spatial auto-correlation, including countries from Britain, Northern Europe and Southern Europe. Figure 4.5 shows the the most frequently reported adverse events in this cluster, having only the adverse effect “Cerebral haemorrhage” which is a common adverse effect in all clusters, as a very common adverse effect having a broad variety of adverse effects with low frequency.

The case studied in Figure 4.6 is similar to the case of Figure 4.4 except for considering the $k = 10$ most frequent sets of adverse effects rather than having $k = 5$. Due to having different representations of countries, the similarity matrix has changed as well. The only cluster having significant spatial auto-correlation in this case includes countries Germany, Denmark, and Spain. And interesting observation made to the right of Figure 4.6 is that for this small cluster of three countries, the reported adverse effects are much more severy, including "Acute kidney injury" and "Cerebrovascular accident" (commonly referred to as a "stroke") as the most common adverse effect.

Summarizing our observation, we observe interesting clusters in terms of adverse effects, including clusters of countries having mainly mild adverse effect (such as "drug ineffective") and clusters of countries have severe adverse affects (such as "Acute kidney injury"). However, it is difficult to visually identify any spatial or temporal patterns, leading us to our quantitative analysis.

Quantitative Analysis

First, we observe seven settings which exhibit a significant (at a level of significance of 0.1) positive auto-correlation and three settings having significant negative auto-correlation. But accounting for the large number of hypothesis carried out by our experiments (for each $k \in \{5, 10\}$, for each of the three drugs, for each of the four years, and for each resulting cluster), only one of these clusters remains significant after Bonferroni correction. For example, for the cluster shown in Figure 4.4, assuming a random pattern, we would expect with a probability of at least 43.31% to find a p-value this low by coincidence in at least one of our hypothesis tests. We conclude that, after Bonferroni correction, most clusters do *not* allow us to confidently reject the null hypothesis that adverse effects appear without a spatial pattern. The one case that remains significant after Bonferroni correction is shown in Figure 4.6 including only the three countries Germany, Denmark, and Spain. Intuitively, this may not sound like a strong spatial cluster. The reason that Morran's I test statistic reports this cluster as spatial-autocorrelated is that in this case, most adjacent countries

share the same values (value 1 for being part of the cluster and 0 for not being part of the cluster). Since Spain is located in a dead-end of Europe having only two adjacent countries (France and Portugal) out of which only one country was considered (Portugal was excluded for the case of $k = 10$ due to insufficient data, see Table 4.3). Thus, the inclusion of Spain only yields one "discordant" pair of spatial neighbors, while Germany and Denmark are adjacent. We conclude that this cluster is an artifact of having for countries for the $k = 10$ case rather than an interesting spatial cluster.

We conclude that after Bonferroni correction to account for the multiple hypothesis testing problem [106], none of the clusters of countries having similar adverse events exhibit a significant and interesting spatial auto-correlation.

4.6 Conclusion

We proposed a first approach to investigate the spatial auto-correlation of adverse events of drugs. We first applied an unsupervised clustering approach to group countries having similar adverse events. For this purpose, we employed a agglomerative hierarchical clustering approach using a distance measure based on Gestalt Pattern Matching to asses the similarity of the most frequent sets of adverse effects mined from each country. For each resulting cluster of countries having similar adverse events, we used Moran's I statistic to test for spatial auto-correlation between countries in the same cluster. While some of the resulting clusters initially showed significant p-values, we had to account for multiple hypothesis testing. After Bonferroni correction to account for the expected number of false positives resulting from all our tested hypothesis, none of the resulting clusters remained significant and interesting using the adjusted p-values.

We conclude that our proposed approach was not able to significant spatial auto-correlation among adverse events of drugs. However, there are many directions of future work to refine this first approach. A first direction is to consider a different representation of the adverse events of each country. In this work, we chose to mine the most frequent sets of adverse effects of the country. A drawback of this approach is that it is not able

to directly understand the similarity between adverse effects, such as, for example, descriptions of the same adverse effect in different languages. To allow an algorithm to learn and understand such similarity, a possible direction is to use a topic modeling approach, for example using Latent Dirichlet Allocation [23], to understand latent topics among adverse effects. A second direction is to use a different distance measure between representations of the adverse events of countries. While we employed a text-similarity measure using Gestalt Pattern Matching, different measures can be employed to better assess which countries are similar. Third, a specialised clustering algorithm may be required. In this work, we employed a complete-link hierarchical agglomerative clustering approach. But using different representations of the adverse events of countries and using different distance metrics, it may be possible to develop clustering algorithms better able to leverage the semantic similarity of different adverse effects. Fourth, a new approach to consider the time of adverse events may be useful to better understand the adverse effects of countries. In our case, we grouped adverse effects by year without looking at similarity between years. But there may be cases where a version of a drug that exhibits certain adverse effects may be released in one country first, and another country later. By considering temporal lag, such patterns could be discerned. Fifth, and finally, a local measure of spatial auto-correlation such as Anselin's Local Indicator of Spatial Association [107] may be used.

We hope our first approach at mining publicly available adverse event databases, such as the FAERS database, to improve our understanding of the spatio-temporal change of the adverse effects of a drug.

Chapter 5: Clustering of Adverse Events of Post-Market Approved Drugs using Latent Dirichlet Allocation

Abstract

Adverse side effects of a drug may vary over space and time due to different populations, environments, and drug quality. Discovering all side effects during the development process is impossible. Once a drug is approved, observed adverse effects are reported by doctors and patients and made available in the Adverse Event Reporting System provided by the U.S. Food and Drug Administration . Mining such records of reported adverse effects, this study proposes a spatial clustering approach to identify regions that exhibit similar adverse effects. We apply a topic modeling approach on textual representations of reported adverse effects using Latent Dirichlet Allocation. By describing a spatial region as a mixture of the resulting latent topics, we find clusters of regions that exhibit similar (topics of) adverse events for the same drug using Hierarchical Agglomerative Clustering. We investigate the resulting clusters for spatial autocorrelation to test the hypothesis that certain (topics of) adverse effects may occur only in certain spatial regions using Moran's I measure of spatial autocorrelation.

Our experimental evaluation exemplary applies our proposed framework to a number of blood-thinning drugs, showing that some drugs exhibit more coherent textual topics among their reported adverse effects than other drugs, but showing no significant spatial autocorrelation of these topics. Our approach can be applied to other drugs or vaccines to study if spatially localized adverse effects may justify further investigation.

5.1 Introduction

Public health surveillance is the base of effective public health practice [89]. Pharmacovigilance is a branch of public health surveillance and is the detection, assessment, understanding, and prevention of adverse effects [8, 9, 48].

Adverse Events (AEs) are any undesirable experiences associated with the use of a medical product. Recent analysis found that it took a median of 4.2 years after a drug's initial approval for major safety concerns to be discovered [2]. Serious side effects could be life-threatening, which can lead to death. While less severe AEs such as rash, nausea, and fatigue might not be dangerous, however, they can lead to avoidance in taking the drug as prescribed, which can lead to a severe consequence [3]. The United States Food and Drug Administration (FDA) uses numerous tools including literature review and surveillance databases to spot potential safety concerns. However, there is no current focus on spatiotemporal aspects of adverse events and their co-occurrence in space and time [108].

To monitor and track Adverse Events, the FDA has developed an Adverse Event self-reporting tool called MedWatch [109]. Patients, practitioners, and drug manufacturers can all report adverse events to FDA Adverse Event Reporting System (FAERS) database via MedWatch. It is not limited to only allergic reactions but also allows to report issues such as product use errors, product quality problems, and therapeutic failures can all be reported via MedWatch. This information is available to everyone including researchers, patients and health practitioners [7]. AEs in the FAERS database are coded using terms in the Medical Dictionary for Regulatory Activities (MedDRA) terminology to standardized medical terminology [110]. Standardizing AE keywords helps facilitate the sharing of information by regulatory authorities, pharmaceutical companies, clinical research organizations and health care professionals and allows for better global protection of public health [86]. Reporting systems such as FAERS are critical tools for monitoring the safety, efficacy, and quality standards of approved pharmaceutical drugs. FDA uses FAERS to study AEs for postmarket surveillance of approved drugs. when investigating safety concerns [7].

We propose an approach to find semantic clusters among adverse effects for a specific drug using topic modeling. We explore if such semantic clusters exhibit significant spatial autocorrelation indicating a clustering in geospace that may justify further investigation to understand causality. For this purpose, we first survey existing work in Section 5.2 and define the problem of spatio-temporal clustering of adverse events in Section 5.3. Then, we propose our approach for clustering regions having similar adverse event sets in Section 5.4. We apply our approach to three common anticoagulant drugs and shared our results in 5.5 and conclude in Section 5.6.

5.2 Related Work

This section surveys related work in pharmacovigilance and related work of using latent topics modeling of text documents.

Pharmacovigilance

The field of pharmacovigilance aims at understanding the occurrence of adverse effects of drugs [8, 9]. Beyond understanding the adverse effects of single drugs, Zitnik, Agrawal, and Leskovec have studied the problem of modeling polypharmacy adverse effects, that is, adverse effects resulting from the interaction of multiple drugs. These important existing works provide solutions to finding significant links between specific drugs and specific adverse effects. However, these studies do not give any consideration to the spatial locations of these adverse effects. Could some patterns between drugs and adverse effects be explained by the spatial distribution of reported adverse effect records? Is it possible that some links between drugs and adverse effects are only observed in a specific region or during a certain time? Existing research leaves such questions largely unanswered. Fortunately, large databases of adverse events, such as the FDA FAERS database are becoming increasingly available and enrich adverse events with both spatial and temporal information.

From complementary perspective, existing work has shown that adverse effects of a single drug or multiple combination of drugs may vary over space and time due to racial

and ethnic disparities [10, 11, 12], environment [13, 14], and drug quality [15]. While these studies describe specific cases and specific drugs, there is no data-driven approach to identify such variations automatically.

Our proposed approach augments data-driven pharmacovigilance with spatial information and provides a framework of finding spatial clusters of regions that exhibit semantically similar adverse effects during a specified period.

Topic Modeling

Topic modeling is an unsupervised learning technique to discover underlying themes of a collection of documents. Latent Dirichlet Allocation (LDA) is one of the more common topic modeling techniques in the literature [23]. LDA assumes an underlying generative probabilistic model that produces the words of a text document given a mixture of k latent topics. Each topic is characterized by a distribution of words. While the traditional application for LDA is modeling of topics among news articles and microblogs [24], it has been used to model the latent topics of points of interest such as restaurants [25]. In the context of pharmacovigilance, LDA has been to find potentially unsafe dietary supplements [26], but without the consideration of the spatial distribution of latent topics among adverse effects. We leverage LDA to find underlying topics of adverse effects reported in a spatial region as a set of latent topics. We then employ this latent feature representation to find spatial clusters of regions that exhibit similar latent adverse effect topics.

Adverse Effects of Blood Thining Drugs

In our experimental evaluation, we chose to investigate spatio-temporal clusters of three blood thinning drugs, namely Dabigatran, Rivaroxaban and Apixaban, due to the wide availability of data on these drugs and their adverse events. Previous studies have shown that within the United States, these three drugs appear to have similar effectiveness [52], although Apixaban may be associated with a lower bleeding risk and Rivaroxaban may be

Table 5.1: Table of Notations

Notation	Description
$\mathcal{A}$	The domain of all adverse effects
$N = \|\mathcal{A}\|$	The number of all advese effects
$A \subseteq \mathcal{A}$	A set of adverse effects
$\mathcal{S}$	The domain of all spatial regions
$\mathcal{T}$	The domain of all discrete time intervals
$\mathcal{D}$	The domain of pharmaceutical drugs
$\mathcal{DB}$	A database of Adverse Events
$M = \|\mathcal{DB}\|$	The number of adverse events in $\mathcal{DB}$
(t, s, A, d)	An adverse event in $\mathcal{DB}$.
$\mathcal{DB}_t$	Adverse events reported during time t
$\mathcal{DB}_{s,t}$	Adverse events reported in region s during time t
K	Number of latent topics used by LDA
α, β	Prior Dirichlet distribution of topics and adverse events within a topic
θ	Topic distribution of adverse events
φ_i	Adverse Effect distribution of topic $1 \leq i \leq K$
Z	A topic $1 \leq Z \leq K$ chosen from θ

associated with an elevated bleeding risk. A similar study in Norway reached similar findings, showing that Dabigatran and Apixaban were both associated with significantly lower risk of major bleeding compared with Rivaroxaban [53]. While these studies investigated the differences of adverse effects across different drugs, these works did not consider spatial or temporal properties of the data. Combined with our knowledge that adverse affects vary across populations and space [10, 11, 12], we investigate if we can identify spatial clusters of regions that exhibit similar adverse effects. We hope that our proposed techniques will be found useful to find links, not only between drugs, but also between regions and drugs to enable spatio-temporal pharmavigilance to find significant links between regions and reported adverse effects.

5.3 Problem Definition

This section formally defines the problem of spatio-temporal clustering of adverse events. A summary of all notations used in this work is found in Table 5.1. First, we provide a definition of adverse effects and events. Intuitively, an adverse effect is a single undesireable effect (such as "bleeding" or "pain"). An adverse event is a report of one or more adverse effects associated with a drug, a time, and a location. Formally:

Definition 6 (Adverse Effect)**.** *An Adverse Effect is a textual representation of an undesirable experiences associated with the use of a medical product. We let $\mathcal{A}$ denote the set of all adverse effects and define $N := |\mathcal{A}|$ as the number of all adverse effects.*

Data such as collected in the FAERS database is a collection of records each associated with a set of adverse effects, a specific pharmaceutical drug, a location, and time. We call such as record an Adverse Event (AE), formally defined as follows:

Definition 7 (Adverse Event Database)**.** *Let $\mathcal{A}$ denote a set of adverse effects, let $\mathcal{S}$ denote a set of spatial regions, let $\mathcal{T}$ denote a set of time intervals (such as years), and let $\mathcal{D}$ denote a set of drugs. An Adverse Event Report Database $\mathcal{DB}$ is a collection of adverse event reports (t, s, A, d), where $t \in \mathcal{T}$ is a point in time, $s \in \mathcal{S}$ is a spatial region, $A \subseteq \mathcal{A}$ is a set of adverse effects, and $d \in \mathcal{D}$ is the drug for which the adverse effects are reported. We let $M := |\mathcal{DB}|$ denote the number of adverse events and emphasize that a single adverse event may report multiple adverse effects.*

As an example, Table 5.2 shows exemplary adverse events from the FAERS database. The first line in Table 5.2 implies that "Abdominal pain, Abdominal pain upper, Constipation, Diarrhoea, Headache, Heart rate increased, Nausea, Pain in extremity, Vertigo and Vomiting" are adverse effects that occurred on on 9/24/2014 in Germany for drug Rivaroxaban.

Our goal is to find clusters of locations that, at a given time, exhibit similar adverse events. Towards this goal, we group adverse events by region and time.

Table 5.2: Sample records of Adverse Event Report Database. Each Line corresponds to an Adverse Event.

Adverse Event ID	Set of Adverse Effects	Location	Event Time	Drug
109947323	Abdominal pain, Abdominal pain upper, Constipation, Diarrhoea, Headache, Heart rate increased, Nausea, Pain in extremity, Vertigo, Vomiting	Germany (DE)	9/24/2014	Rivaroxaban
106823542	Duodenal ulcer haemorrhage, Gastric ulcer haemorrhage, Shock haemorrhagic	Netherlands (NL)	11/2/2014	Rivaroxaban
109449521	Death	United Kingdom (GB)	2/12/2015	Rivaroxaban
120813061	Asthenia, Haemorrhage	Croatia (HR)	1/16/2016	Rivaroxaban
145539611	Purpura, Skin exfoliation, Skin lesion	United Kingdom (GB)	12/28/2017	Rivaroxaban

Definition 8 (Spatio-Temporal Adverse Events)**.** *Let $\mathcal{DB}$ be an adverse event report database, let $s' \in \mathcal{S}$ be a spatial region, and let $t' \in \mathcal{T}$ be a time interval. We define*

$$\mathcal{DB}_{s',t'} := \{(t, s, A, d) \in \mathcal{DB} | t = t' \wedge s = s'\}$$

as the set of all adverse events reported at time t' at location s'. For two spatial regions s_1 and s_2, we let

$$dist(\mathcal{DB}_{s_1,t}, \mathcal{DB}_{s_2,t}) \mapsto [0, 1]$$

denote an abstract distance function between two sets of adverse events.

We propose a concrete implementation of $dist()$ in Section 4.4.3 based upon latent topic similarity of the adverse effects in each spatial region. Given a distance function to assess

the adverse event similarity of two regions at the same time and for a given drug, we define a spatial adverse event clustering as follows:

Definition 9 (Spatial Adverse Event Clustering)**.** *Let $\mathcal{DB}$ be an adverse event report database, let dist() be a distance function to measure dissimilarity among sets of adverse events. Further, let*

$$\mathcal{DB}_t := \{x \in \mathcal{DB} | x.t = t\}$$

denote the set of all adverse events reported at time t. A spatial clustering $\mathcal{C}(\mathcal{DB}_t) = \{C_1, ..., C_K\}$ is a partition of regions S such that $\forall i : C_i \subseteq \mathcal{S}$ and $\forall i \neq j : C_i \cap C_j = \emptyset$. As element $C \in \mathcal{C}$ of a spatial adverse event clustering is called a spatial adverse event cluster.

Problem Statement 3 (Spatial Adverse Event Clustering)**.** *Our first goal is to find a good spatial adverse event clustering $\mathcal{C}(\mathcal{DB}_t)$, that is, a clustering such that the distance $dist(\mathcal{DB}_{s_1,t} \in C_i, \mathcal{DB}_{s_2,t} \in C_i)$ of regions in the same cluster is minimize, while the distance $dist(\mathcal{DB}_{s_1,t} \in C_i, \mathcal{DB}_{s_2,t} \in C_j)$ of regions across clusters $C_i \neq C_j$ is maximized.*

Once a good spatial adverse event clustering is found, our second goal is to explore the spatial autocorrelation of these clusters to answer the question if some regions exhibit similar adverse effects at certain times, or if adverse effects are independent of space and time.

Problem Statement 4 (Spatial autocorrelation of Clusters)**.** *Given a spatial adverse event clustering $\mathcal{C}(\mathcal{DB}_t)$, our second goal is to explore the spatial autocorrelation among regions in the same cluster $C_i \in \mathcal{C}(\mathcal{DB}_t)$.*

The clustering step is needed to map spatial regions to a numeric values of their (one hot encoded) cluster membership. These values can then be investigated for spatial autocorrelation.

5.4 Methodology

This section describes our approach to find spatio-temporal clusters of adverse events and to find those having significant spatial autocorrelation. And overview of this section is found in Figure 5.1. First, we obtained the data from the FAERS Database using the openFDA API [75, 74] as described in Section 4.4.1. To concisely represent the adverse events reported in a region, Section 5.4.2 presents our Latent Dirichlet Allocation (LDA) based approach to extract latent topics from adverse events. Having each region (at a specified time) represented by a set of latent features corresponding to the strength of respective topics among the reported adverse effects, we define a distance function to measure dissimilarity between the adverse events of regions in Section 4.4.3 and leverage this distance function to cluster the adverse events among regions in Section 4.4.4. We test each of the resulting clusters for significant spatial autocorrelation using Moran's I test statistic to measure spatial autocorrelation as described in Section 4.4.5. All of our algorithms used for our data analysis have been implemented in Python 3.7 and are published on GitHub as a Jupyter Notebook for reproducibility at https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Clusters-LDA.

5.4.1 Data Collection

As of 12/01/2020, there are 19 million adverse event reports for countless drug combinations in the FAERS database. We implemented a data crawler in Python to download these reports from the FAERS public-facing database published quarterly as a zip file available at fda.gov [74] or by invoking the OpenFDA API [75]. FAERS data comes in multiple files with a primary key linking all files. The files include a demographic and administrative information file; drug information of the case reported; reaction information from the reported case along with patient outcome information and the source of the report. Unfortunately, spatial information is aggregated to the country level for the public-facing FAERS database due to HIPAA regulations. Due to this aggregation, we focus on clustering of adverse effects using European countries having at least 30 adverse events per drug as shown in Table 5.3 for the three drugs used in our experimental evaluation.

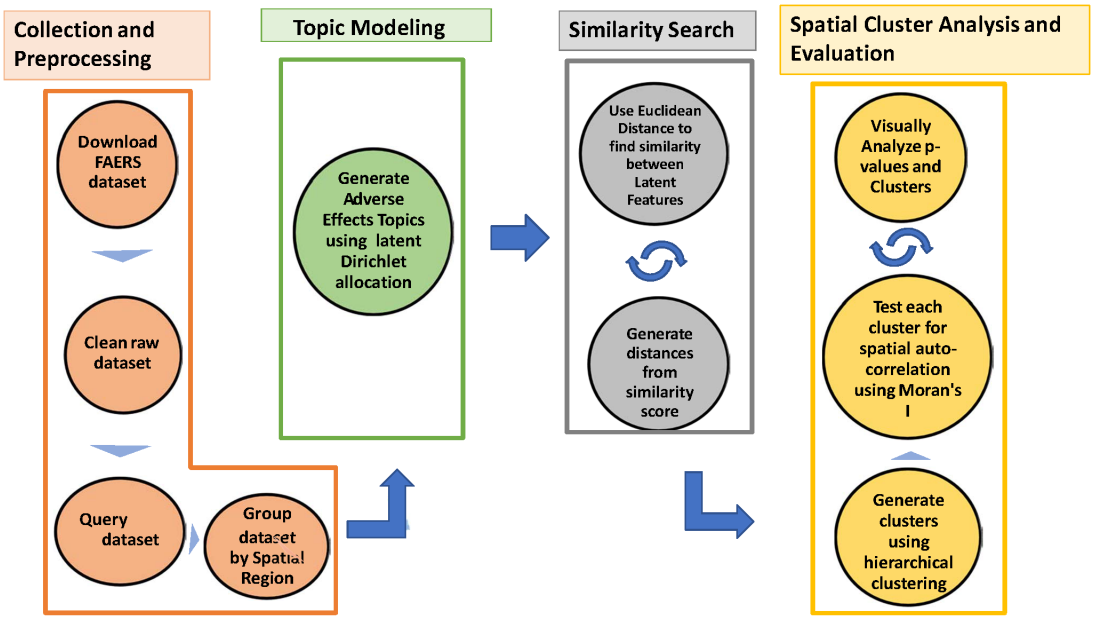


Figure 5.1: Road map for spatiotemporal clustering of topics generated from Latent Dirichlet Allocation using Adverse Events Report Submission to FDA FAERS Database.

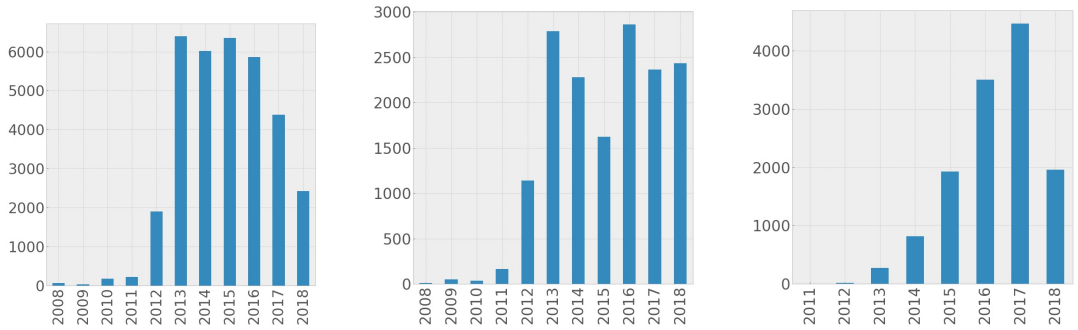
We note that our solutions can also be applied to other study areas, such as the United States on state or county level for policy makers having such data and for the general public when such data becomes publicly available. Once the data was collected, we joined all datasets by their primary key to obtain one large data table, including for each case report a spatial attribute and a timestamp of the observed adverse events. We group this dataset by year between 2014 to 2017. Data varied for different countries depending on the pharmaceutical drug availability, prescribers' preferences, pharmaceuticals marketing, supply chain, etc. In the exploratory data selection phase, we were limited to select pharmaceutical drugs which were used consistently within this study period and used broadly across Europe. Rivaroxaban was selected as our pharmaceutical study drug due to a wide distribution of adverse events across Europe along with it being one of top ten most reported pharmaceutical drug in the FAERS Database. Since Rivaroxaban is an anticoagulant drug, we used other similar anticoagulant drugs such as Dabigatran and Apixaban as our other drugs in our

Table 5.3: Study areas for the drugs used in this study.

Countries in the study area	# of Countries	Drug Name
AT, BE, DE, EE, ES, FR, GB, GR, IE, IT, NL, NO, PL, PT, SE, SI	16	Dabigatran
AT, DE, ES, FR, GB, GR, IT, NO, SE	9	Apixaban
AT, BE, BG, BY, CH, CZ, DE, DK, EE, ES, FL, FR, GB, GR, HR, HU, IE, IT, NL, NO, PL, PT, SE, SI, SK, TR	26	Rivaroxaban

study. There is ample about of literature comparing these three drugs [53, 52]. Rivaroxban and Dabigatran received approval to market from European Medicines Agency (EMA)in 2008 [92, 93] and Apixaban in 2011 [94]. It takes a median of 4.2 years after a drug's initial approval for major safety concerns to be discovered [2]. Report submission increases few years after the initial approval. For reference, Figure 5.2 shows the temporal distribution of adverse events between 2008 and 2018 to justify our choice of using 2013 to 2017 as our study period. Many reported adverse events in the FAERS data report the concurrent usage of multiple drugs (phramacovigilance). For our study, we did not consider any such adverse events to avoid confusion caused by adverse effects of other drugs. We note that our work can be easily extended to compare concurrent drug usage to explore the effects of polypharmacy AEs [48]. Once data is collected for a drug, we group adverse events by year and country to obtain spatio-temporal adverse events $\mathcal{DB}_{s,t}$. A Python script to obtain data through the OpenFDA API [75] and to apply the preprocessing steps as described in this section can be found at https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Clusters-LDA.

Given spatio-temporal adverse events, we next propose an latent topic modeling approach to understand similarity among different adverse effects among adverse events in Section 5.4.2. We then use the latent topics among adverse events to asses the similarity of adverse events between two regions as described in Section 4.4.3 which is used to cluster regions as described in Section 4.4.4. Finally, we investigate the spatial autocorrelation of the resulting clusters as described in Section 4.4.5.



(a) Rivaroxaban (b) Dabigatran (c) Apixaban

Figure 5.2: Adverse Events Report Submission to the FDA FAERS Database per drug per year

5.4.2 Adverse Event Topic Mining using Latent Dirichlet Allocation

A challenge of mining adverse events is the potentially large number of different adverse effects. The FAERS Adverse Event Databases use MedDRA codes [86] and terminology to standardize adverse effects such as using "nausea" instead of "feeling queasy". Yet, the number of possible adverse effects is too large and the resulting feature space of using bag-of-words semantics to represent adverse effects is too high dimensional. To address this issue, we acknowledge that adverse effects are symptoms of unknown (latent) underlying causes. While one way of identifying causes is involving a medical expert, we propose a data-driven approach to identify underlying topics among adverse events using topic modeling that we interpret as causes. For that, we employ Latent Dirichlet Allocation (LDA) [23] – a generative probabilistic model which assumes that each adverse event is a mixture of underlying (latent) topics, and each topic has a (latent) distribution of more and less likely adverse effects.

A graphical representation of our LDA model using plate notation is shown in Fig. 5.3. A vector α of length K is used to parameterize the *a priori* distribution of topics. The parameter K corresponds to the number of latent topics used to model adverse events.

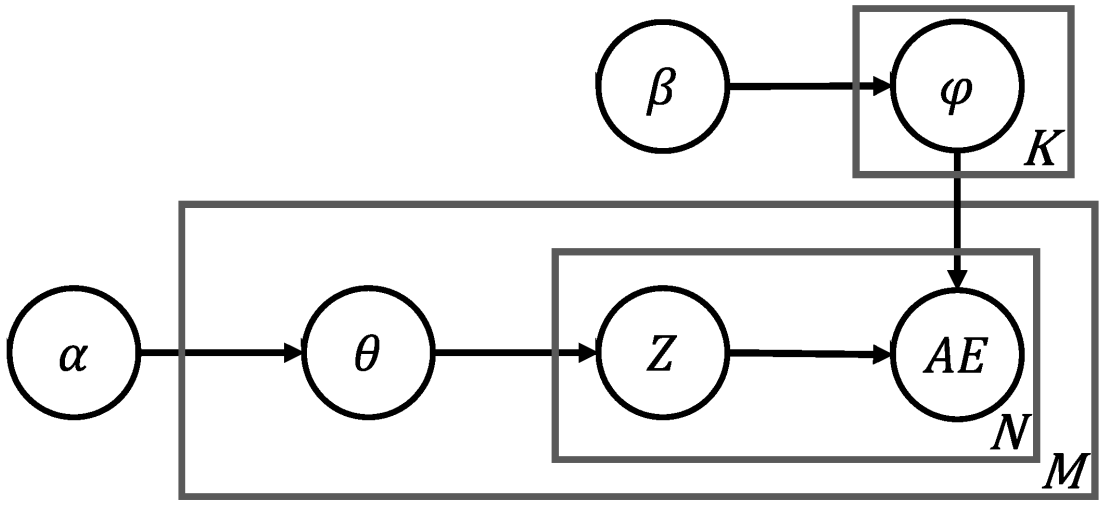


Figure 5.3: LDA Topic Modeling of Events. For each adverse event a topic distribution θ is estimated and for each topic i, an adverse effect distribution φ_i is estimated. Given a topic Z generate from θ, observable adverse effects (AEs) are generated from φ_Z.

When an adverse event is created, we assume that its topics are chosen following a *Dirichlet distribution* having parameter α which we use to obtain a topic distribution θ for each of our $M =$ adverse events. Thus, the large plate in Fig. 5.3 corresponds to a set of M adverse events, each having a topic distribution θ drawn randomly (and Dirichlet distributed) from α.

For each topic, the prior parameter β is used to generate the distribution of words within a topic. Thus, we assume that a topic generates adverse effects following a Dirichlet distribution having a vector β of length $|\mathcal{A}|$ as parameter, where $\mathcal{A}$ is the set of observed adverse effects (c.f. Definition 6). For each of our K topics, a resulting vector $\varphi_i, 1 \leq i \leq K$ stores the adverse effect distribution of topic K.

To generate the adverse effects of an adverse event, a topic is chosen randomly from the topic distribution θ and, given this topic, a number of N_i adverse effects are generated randomly from the adverse effect distribution φ – where N_i is assumed to be independent from the chosen topic and uniformly distributed. In Fig. 5.3, the node AE denotes the

(observable) set of all $N = \sum_i N_i$ adverse effects, and Z is a function that maps each word to the topic that generated it. The reason for choosing a Dirichlet distribution rather than a more straightforward uniform or multinomial distribution for the topic and word priors is inspired by research showing that the distribution of words in text can be better approximated using a Dirichlet distribution [111].

To infer the topics of our adverse event database $\mathcal{DB}$, we employ a generative process. Given the observed adverse effects, LDA optimizes the latent variables to maximize the likelihood of matching the observed adverse events and corresponding adverse effects. This generative process works as follows. Adverse events are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all N adverse effects. LDA assumes the following generative process for database $\mathcal{DB}$ consisting of M adverse events, each having a number of N_i adverse effects.

- For each adverse event choose a topic distribution $\theta_m \sim Dir(\alpha), 1 \leq m \leq M$, where $Dir(\alpha)$ is a Dirichlet distribution with prior α. In our experiments, we initially assume each topic to have uniform prior probabilities, having $\alpha_i = \alpha_j$ for $1 \leq i, j \leq K$. This apriori distribution is adapted using Bayesian inference [23] to maximize the likelihood of generating the observed keywords.

- For each topic, choose an adverse effect distribution $\varphi_i \sim Dir(\beta)$, where $1 \leq i \leq K$. For our experiments, we assume each adverse effect to have the same prior probability N^{-1}.

- For each adverse effect ae in adverse event j:

  1. Choose a topic $z \sim Multinomial(\theta_j)$ from the topic distribution of j, and
  2. Choose a word $w \sim Multinomial(\varphi_z)$ from the adverse effect φ_z of topic z.

  Here, $Multinomial(x)$ corresponds to a multinomial distribution drawing from a stochastic vector x.

To describe each adverse event in a latent topic space, we use the adverse event specific topic distributions θ_m which describe each adverse event m as a set of K latent features corresponding to the weight of the respective latent topic. While this topic modeling does not provide us with any semantic of the underlying topics, we know that adverse events having similar latent features also exhibit similar adverse effects. Based on the similarity of latent topics of individual adverse events we define a similarity measure to assess the similarity between the adverse events within a region in Section 4.4.3. Given this measure of adverse event similarity between regions we propose a hierarchical agglomerative clustering approach to find regions that exhibit similar adverse events in Section 4.4.4 and test these clusters for spatial autocorrelation using Moran's I in Section 4.4.5.

5.4.3 Similarity between Sets of Adverse Events

Definition 9 requires a distance function $dist(\mathcal{DB}_{s_1,t}, \mathcal{DB}_{s_2,t})$ between sets of adverse events $\mathcal{DB}_{s_1,t}$ and $\mathcal{DB}_{s_2,t}$ corresponding to two regions s_1 and s_2 at time t. We describe each such set as the mean of latent features within the set, and measure the Euclidean distance in the latent feature space, formally:

Definition 10 (Spatio-Temporal Adverse Event Distance)**.** *Let $\mathcal{DB}$ be an adverse event data, let $\mathcal{DB}_{s_1,t}, \mathcal{DB}_{s_2,t} \subseteq \mathcal{DB}$, let K be a positive integer and let $\theta(ae)$ denote the latent topic distribution of an adverse event $ae \in \mathcal{DB}$ using the LDA model described in Section 5.4.2, then:*

$$dist(\mathcal{DB}_{s_1,t}, \mathcal{DB}_{s_2,t}) := \left\| \frac{\sum_{\mathcal{DB}_{s_1,t}} \theta(ae)}{|\mathcal{DB}_{s_1,t}|} - \frac{\sum_{\mathcal{DB}_{s_2,t}} \theta(ae)}{|\mathcal{DB}_{s_2,t}|} \right\|_2,$$

where $\|.\|_2$ is the Euclidean norm.

This distance function in Equation 10 allows us to identify clusters of spatial regions $s \in \mathcal{S}$ that exhibit similar adverse event topics as described in the following.

Table 5.4: Top-10 most probably keywords for K=5 for Rivaroxaban.

Topic	Keywords (Probabilities in %)
1	(1.6)"cerebrovascular-accident", (1.3)"ischaemic-stroke", (0.9)"cerebral-haemorrhage", (0.8)"aphasia", (0.8)"muscular-weakness", (0.7)"urinary-tract-infection", (0.7)"death", (0.7)"asthenia", (0.6)"haematuria", (0.6)"subdural-haematoma"
2	(2.7)"drug-ineffective", (1.9)"deep-vein-thrombosis", (1.6)"dizziness", (1.5)"dyspnoea", (1.3)"cerebral-haemorrhage", (1.3)"pulmonary-embolism", (1.2)"headache", (1.2)"cerebral-infarction", (1.2)"anaemia", (1.2)"gastrointestinal-haemorrhage"
3	(1.6)"dizziness", (1.6)"cerebral-haemorrhage", (1.5)"headache", (1.5)"fall", (1.5)"gastrointestinal-haemorrhage", (1.4)"dyspnoea", (1.2)"cerebrovascular-accident", (1.1)"deep-vein-thrombosis", (1.1)"pulmonary-embolism", (1.0)"upper-gastrointestinal-haemorrhage"
4	(4.7)"drug-ineffective", (3.6)"anaemia", (3.5)"ischaemic-stroke", (2.3)"deep-vein-thrombosis", (2.1)"pulmonary-embolism", (1.7)"cerebral-infarction", (1.5)"cerebral-haemorrhage", (1.2)"asthenia",(1.2)"fall", (1.2)"dyspnoea"
5	(0.8)"drug-ineffective-for-unapproved-indication", (0.7)"deep-vein-thrombosis", (0.5)"refractory-cytopenia-with-unilineage-dysplasia", (0.5)"arterial-rupture", (0.5)"gamma-glutamyltransferase-abnormal", (0.5)"pulmonary-haemorrhage", (0.5)"cholestatic-liver-injury", (0.3)"death")"budd-chiari-syndrome", (0.3)"cerebral-venous-thrombosis"

5.4.4 Clustering Regions by Adverse Events

To find clusters among regions having similar topics of adverse events, we employ a hierarchical agglomerative clustering approach [99]. The advantage of such approach is that

Table 5.5: Top-10 most probably keywords for K=5 for Dabigatran.

Topic	Keywords (Probabilities in %)
1	(8.1)"ischaemic-stroke", (5.0)"cerebrovascular-accident", (3.8)"cerebral-infarction", (2.5)"hemiparesis", (2.0)"fall", (1.9)"haemorrhage", (1.4)"cerebral-haemorrhage", (1.2)"aphasia", (1.2)"dysarthria"
2	(4.5)"ischaemic-stroke", (3.7)"gastrointestinal-haemorrhage", (2.1)"fall", (2.1)"death", (2.1)"haemorrhage", (1.6)"renal-failure", (1.6)"atrial-fibrillation", (1.6)"atrial-thrombosis", (1.3)"cerebrovascular-accident", (1.1)"haemoglobin-decreased"
3	(2.4)"gastroenteritis", (1.7)"cerebrovascular-accident", (1.7)"syncope", (1.7)"tubulointerstitial-nephritis", (0.9)"injury", (0.9)"haematuria", (0.9)"cholecystitis", (0.9)"renal-failure", (0.9)"subarachnoid-haemorrhage", (0.9)"acute-kidney-injury"
4	(2.6)"gastrointestinal-haemorrhage", (2.6)"rectal-haemorrhage", (2.6)"cerebral-haemorrhage", (1.9)"anaemia", (1.6)"melaena", (1.6)"cardiac-failure", (1.6)"weight-decreased", (1.6)"fall", (1.3)"dyspnoea", (1.3)"abdominal-pain-upper"
5	(5.5)"drug-ineffective", (3.8)"ischaemic-stroke", (2.5)"anaemia", (2.1)"cerebrovascular-accident", (2.0)"melaena", (1.7)"cerebral-haematoma", (1.5)"rectal-haemorrhage", (1.4)"death", (1.2)"gastrointestinal-haemorrhage", (1.2)"haemorrhage-intracranial"

we neither have to guess the number of clusters as often needed for partitioning clustering approaches [100] nor have to define a density threshold as required by density-based clustering algorithms [101]. To merge clusters, we employ complete linkage, which defines the distance between two clusters of regions as the maximum pair-wise distance of regions among the clusters.

We retain all clusters (of all sizes) corresponding to all nodes in the dendrogram excluding clusters of size one and excluding the root of the dendrogram that contains all regions. As an example, Figure 5.6 shows the similarity matrix between $n = 16$ countries in Europe for drug Dabigatran in Year 2014 using $k = 5$ latent topics. The corresponding dendrogram is shown above and to the left of this matrix. We consider each node of the dendrogram

Table 5.6: Top-10 most probably keywords for K=5 for Apixaban.

Topic	Keywords (Probabilities in %)
1	(5.1)"anaemia", (2.2)"asthenia", (2.2)"haematuria", (1.7)"off-label-use", (1.7)"head-injury", (1.7)"transfusion", (1.3)"medication-error", (1.3)"rectal-haemorrhage", (0.9)"ischaemic-stroke", (0.9)"surgery"
2	(4.3)"fall", (1.9)"cerebral-haemorrhage", (1.5)"haematoma", (1.3)"cerebrovascular-accident", (1.3)"subdural-haematoma", (1.3)"transfusion", (1.2)"anaemia", (1.1)"haemorrhage", (1.1)"prescribed-underdose", (1.1)"ischaemic-stroke"
3	(3.1)"fall", (2.1)"cerebral-haemorrhage", (2.1)"off-label-use", (1.8)"drug-ineffective", (1.6)"rectal-haemorrhage", (1.3)"pruritus", (1.3)"transfusion", (1.3)"head-injury", (1.3)"haematuria", (1.1)"death"
4	(2.1)"fall", (2.1)"haematoma", (2.1)"anaemia", (1.5)"transfusion", (1.2)"haemorrhage", (1.2)"cerebral-haemorrhage", (1.2)"epistaxis", (1.2)"melaena", (1.2)"myocardial-infarction", (1.2)"limb-injury"
5	(1.9)"acute-kidney-injury", (1.8)"fall", (1.3)"haemorrhage-intracranial", (1.3)"drug-ineffective", (1.3)"gastrointestinal-haemorrhage", (1.3)"cerebral-haemorrhage", (0.7)"cognitive-disorder", (0.7)"depressed-level-of-consciousness", (0.7)"cardioversion", (0.7)"hemiparesis"

(and countries in the corresponding subtree) as a cluster. This yields a total of $n-2$ clusters of size between 2 and $n-1$. Next, we investigate if the resulting clusters exhibit significant spatio autocorrelation.

5.4.5 Measure of Spatial Autocorrelation

Given a cluster of regions that exhibit similar topics of adverse events, we employ Moran's I measure of spatial autocorrelation [112]. Moran's I statistic tests if a variable measured

on spatial regions exhibits a significant spatial autocorrelation, either positive (clustered) or negative (dispersed). To measure the spatial autocorrelation of clusters obtained as described in Section 4.4.4, we encode each individual cluster membership into a binary variable. Thus, for a cluster C, the cluster membership variable of a region r is set to 1 if $r \in C$ and 0 otherwise. Given such encoding, Moran's I evaluates if the regions of a cluster exhibit a spatially clustered, random or dispersed pattern. Moran's I requires an adjacency metric on regions for which we use Queen Contiguity, that is, two regions are considered adjacent if they share boundary. We directly report Moran's I test statistic whose range is in $[-1, -1]$, ranging from strongly dispersed (close to -1) to strongly clustered (close to 1). We also report the p-value of the null-hypothesis that the regions are distributed randomly without any spatial pattern by transforming Moran's I values to z-values and employing a two-tailed z-test [113]. The resulting p-values indicate whether a cluster of regions having similar topics of averse events are significantly spatially clustered or dispersed. We used the geopandas library for handling spatial attributes and Pysal library for Moran's I test of spatial autocorrelation [80, 103].

5.5 Experimental Evaluations

For our experimental evaluation we collected data from the FAERS database as described in Section 4.4.1 for three anticoagulant medications - Apixaban, Rivaroxaban and Dabigatran grouped by four years (2014-2017). To investigate the latent topics found by our topic modeling approach in Section 5.5.1, we first present a qualitative evaluation of the resulting topics in Section 5.5.1. Then, Section 5.5.3 presents the results of the clustering approach described in Sections 4.4.4 and investigates the spatial autocorrelation among this clusters as described in Section 4.4.5 .

5.5.1 Qualitative Analysis of Latent Topics among Adverse Events

Figure 5.4 shows the topic coherence [114] of the latent topics modeled by our LDA approach described in Section 5.4.2. We observe a high cluster coherence for drugs Dabigatran and

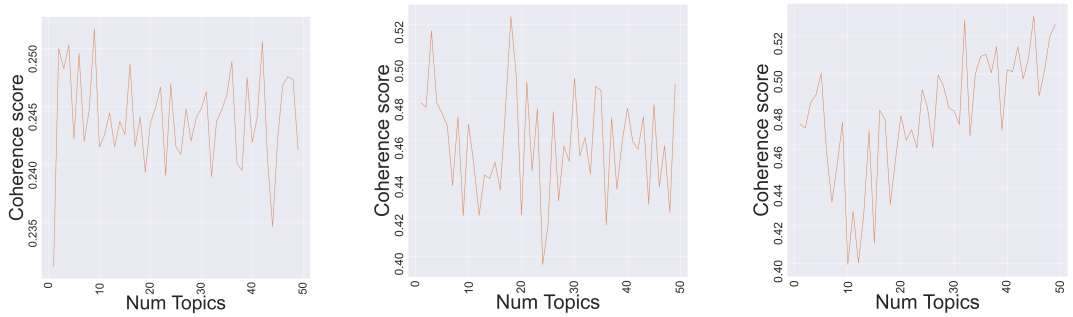
Table 5.7: Statistically significant ($p < 0.05$) clusters for Rivaroxaban, Dabigatran and Apixaban for Year 2014-2017 for $k \in \{5, 10\}$.

Fig	Spatial Pattern	Pvalue	Zscore	MoranI	Countries	# of countries in cluster	Year	Drug Name	Topic #
5.5	Clustered	0.041	2.04	0.287	es, fr, it	3	2015	Rivaroxaban	5
	Dispersed	0.0321	-2.14	-1	at, fr, se	3	2014	Apixaban	5
5.7	Cluster	0.0003	3.63	1.20	no, se	2	2017	Apixaban	10
	Clustered	0.0109	2.54	0.653	es, gb, ie, nl, pt	5	2014	Dabigatran	5
5.6	Cluster	0.0192	2.34	0.605	es, gb, gr, ie, nl, pt	6	2014	Dabigatran	5
	Clustered	0.0404	2.04	0.402	be, nl	2	2015	Dabigatran	5
	Clustered	0.0344	2.11	0.515	ee, gb, ie, it	4	2017	Dabigatran	5
	Clustered	0.0001	3.85	0.816	es, pt	2	2014	Dabigatran	5
	Clustered	0.0042	2.86	0.682	gb, ie, nl	3	2014	Dabigatran	5
	Clustered	0.0001	5.19	1.12	gb, ie	2	2017	Dabigatran	5
	Clustered	0.0404	2.04	0.402	be, nl	2	2015	Dabigatran	10
	Clustered	0.0048	2.81	0.670	no, se, si	3	2016	Dabigatran	10

Apixaban for less than $k = 10$ latent topics. For Drug Rivaroxaban, we observe a lower cluster coherence, indicating a large number of adverse effects (keywords) that cannot be assigned to any topic. Given these observations, we use $k \in \{5, 10\}$ for our experiments.

For $k = 5$ latent topics, Tables 5.4-5.6 show the top-10 most probable keywords for each of the $\varphi_i, 1 \leq i \leq 5$ word distributions and for each of the three considered drugs.

Of all the drugs in our study, Drug Rivaroxaban (Table 5.4) shows the least clear topics. The first topic mainly relates to injuries of the brain, as the top-5 most common adverse effects are: cerebrovascular accident, ischaemic stroke, cerebral haemorrhage, and aphasia. Topics #2-#4 appear to relate to less severe adverse effects such as ineffective drug, thrombosis, and dizziness, but also involve brain-related adverse effects with high probabilities. Topic #5 is an outlier topic covering many adverse effects with very low weights which explains the low coherence of this model (see Figure 5.4), Overall, the adverse effects that



(a) Rivaroxaban (b) Dabigatran (c) Apixaban

Figure 5.4: Coherence Scores vs Number of Topics to determine the number k of latent topics for LDA.

we see most dominantly, with high probabilities in most topics, refer to ineffective drugs.

For Drug Dabigatran (Table 5.5) the first topic relates to adverse effects of the brain caused by bleeding or lack of blood supply. The second topic also related to strokes but also including side effects in other parts of the body. Topics #3 and #4 pertain to adverse events of the stomach while Topic #5 appears to refer to other side effects, including drug inefficacy but also death. We also note that strokes appear to be a common adverse effect of Dabigatran, which appears in most most topics, often with high probability.

For Drug Apixaban (Table 5.6), we observe one topic related to Anaemia (lack of red blood cells) and multiple topics related to (with different weights) fall and bleeding (haemorrhage and haematoma). We observe that "fall" appears to be a common side effect of Apixaban.

We note that this qualitative study should not be used to assess, in any way, the quality of drugs. There may be other causal factors such as different patient groups (having different age and pre-conditions) for different drugs. What we can observe, is that this qualitative interpretability of the resulting adverse event topics shows that LDA is able to discriminate topics among adverse events for all drugs. This is a promising result, showing that given a

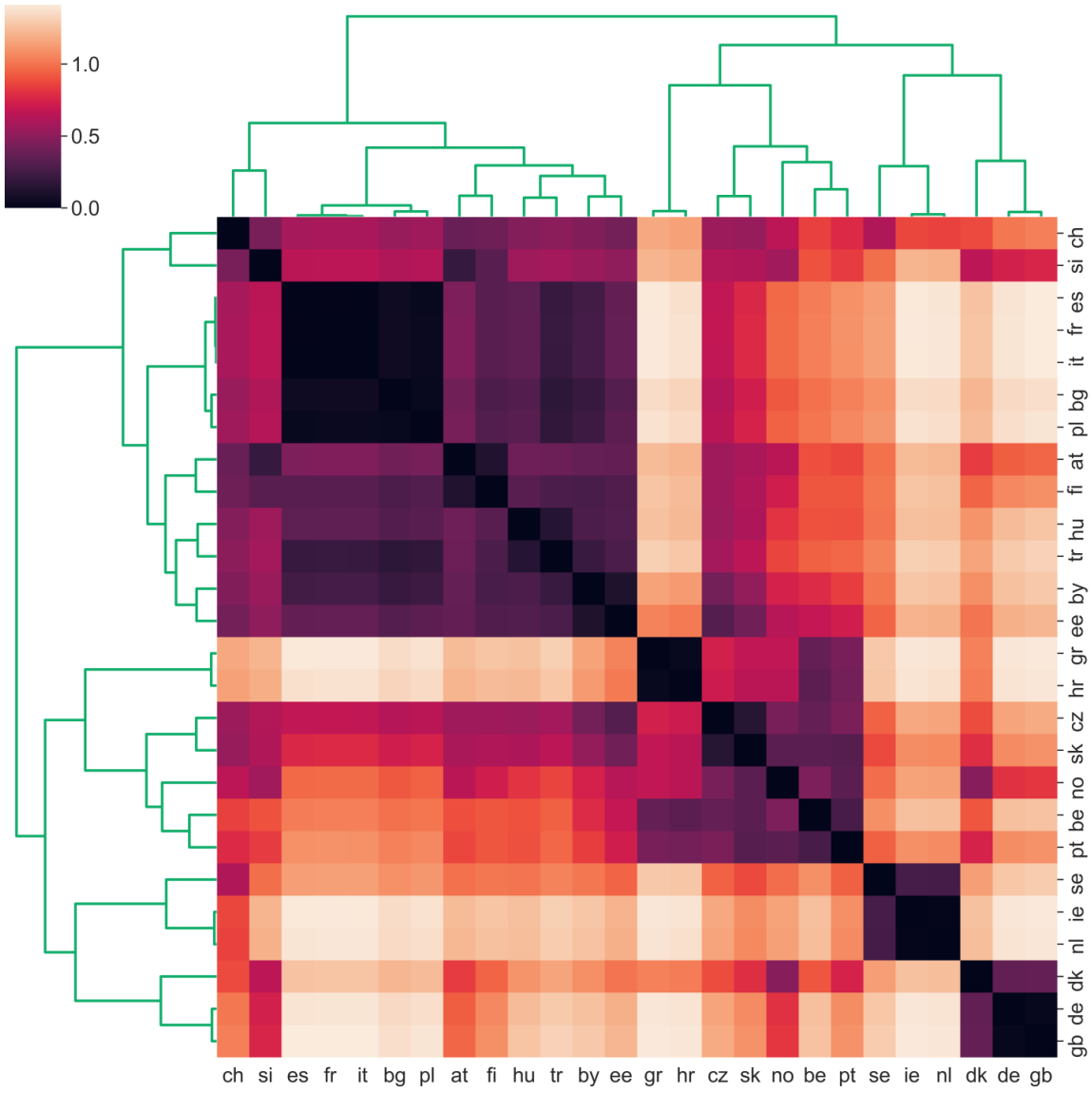


Figure 5.5: Clustering for Drug Rivaroxaban, $k = 5$, Year 2015

sufficiently large adverse event database for a drug, a topic modeling approach can indeed model latent underlying causes, and that the resulting topic distributions of regions may indeed be a proper representation that allows to compare the adverse events across countries in a meaningful way.

5.5.2 Qualitative Analysis of Adverse Event Clusters

In this section, we answer Problem Statement 3 to evaluate whether we can find discriminative clusters of adverse events among regions. Figures 5.5-5.7 shows the similarity matrices between all regions used for the respective drugs (cf. Table 5.3) for $k = 5$ topics for the specified years using the distance measure defined in Section 4.4.3. For drug Rivaroxaban, Figure 5.5 shows one very large cluster of 13 out of 26 regions each having low distance among each other but having a large distance to the remaining regions. Among these remaining regions, we find two well-discriminated clusters of size three and three containing countries {Sweden, Ireland, Netherlands} and {Denmark, United Kingdom, Germany} respectively. For Drug Dabigatran we observe excellent discrimination between adverse events of countries in Figure 5.6. We observe three clusters of sizes six, four, and four which exhibit very high intra-cluster similarity and very low cross-cluster similarity. A similar high quality clustering can be observed in Figure 5.7 for Drug Apixaban which only includes nine countries in the study region showing two clear clusters of sizes four, and three.

Overall, we observe that many countries indeed exhibit very high adverse event similarity and our hierarchical clustering algorithm is able to well discriminate groups of highly similar countries.

5.5.3 Spatial Autocorrelation

Next, we investigate Problem Statement 4 to see if the resulting clusters of regions exhibit significant spatial autocorrelation. For this purpose, our null hypothesis states that there is no spatial autocorrelation between countries within the same cluster of adverse events. that is, adverse events appear randomly without a spatial pattern. An excerpt of results of our experimental evaluation for *selected* years, drugs, and, and values of $k \in \{5, 10\}$ clusters can be found in Table 5.7. For clusters of regions exhibiting similar adverse events (for a given drug, given year, and given value of k), this table provides the corresponding Moran's I measure of spatial autocorrelation and corresponding p-value (probability that I may be coincidental under the null hypothesis) as described in Section 5.4.5.

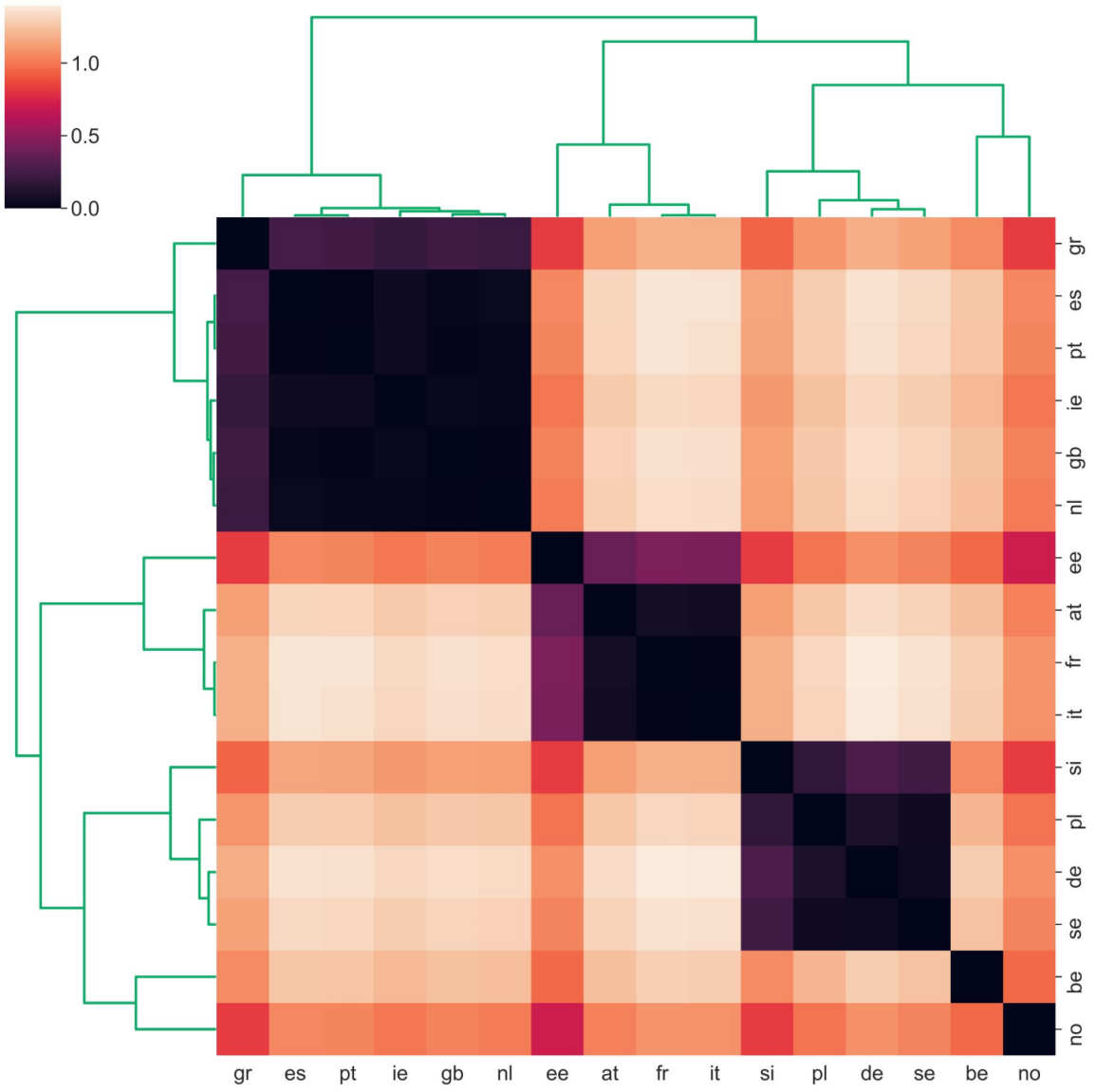


Figure 5.6: Clustering for Drug Dabigatran, $k = 5$, Year 2014

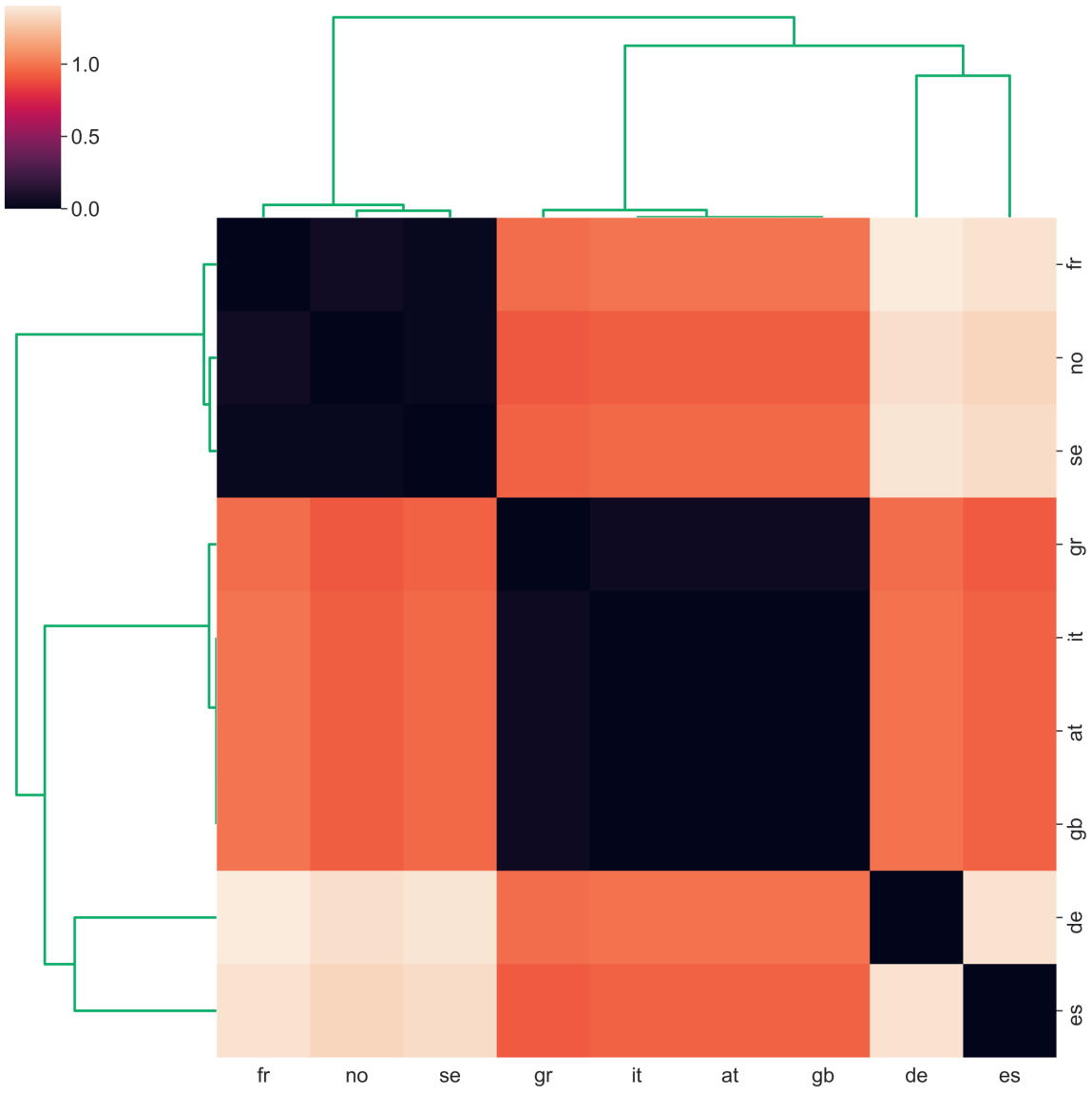


Figure 5.7: Clustering for Drug Apixaban, $k = 5$, Year 2017

First and foremost, we note that we only show clusters having a p-value of less than 0.05 in Table 5.7, thus omitting a large number of non-interesting clusters. Since we used each (non-root) node of the cluster dendrogram for each drug, we obtained $n-2$ clusters per drug, where n is the number of countries used in our study for each drug. Since we used 26, 16, and 9 countries for Rivaroxaban, Dabigatran, and Apixaban, respectively (see Table 5.3), this yields $(26-2)+(16-2)+(9-2) = 45$ clusters for each of the four Years 2014-2017 and for each $k \in \{5, 10\}$, yielding a total of $45 * 4 * 2 = 360$ hypothesis tests. Table 5.7 shows that only 19 (5.3%) of these hypothesis yield a significant spatial autocorrelation. Using a two-tailed hypothesis test at a level of signifiance (or false positive rate), we expect 5% of false positives assuming that the null hypothesis holds. Getting 19 out of 360 Bernoulli trials has a p-value of > 0.43. Thus, we can *not* reject the null hypothesis that adverse events appear randomly in space. This negative result is also confirmed using Bonferroni correction [115] to account for testing 360 hypothesis. After applying Bonferroni correction, none of the spatial autocorrelations of Table 5.7 remain significant.

We conclude that our experimental evaluation does not confirm spatial autocorrelation among clusters of countries exhibiting similar adverse events. Yet, we note that Table 5.7 shows more positive than negative cases of spatial autocorrelation, indicating that there may indeed be a trend. More research is needed to investigate spatial patterns among adverse events. The authors suspect that the coarse spatial granularity of the data availble at the FDA FAERS database on country level may hide interesting spatial patterns.

5.6 Conclusion

We proposed a first approach to measure the similarity of reported adverse events between spatial regions based on the latent topics of adverse events. Based on this similarity, we proposed a clustering approach to group countries having similar adverse events and evaluated the degree of spatial autocorrelation among regions in the same group. Our experimental has shown that we can indeed find clusters of countries that exhibit similar adverse events.

However, we were not able to confirm spatial autocorrelation between these regions. We note that more research in this field is needed.

One limitation of our approach is the aggregation at country level, which may have interesting local spatial patterns. Applying our solutions to smaller spatial regions may find such patterns. We also note that a different measure of spatial proximity may yield stronger autocorrelation by considering not only topological distance but also including political and socioeconomic similarities. To summarize, we did show that some countries exhibit similar topics of adverse events, but an deeper investigation of patterns and their causality is needed. We hope our approach at mining publicly available adverse event databases improves our understanding of the spatio-temporal change of the adverse effects of a drug.

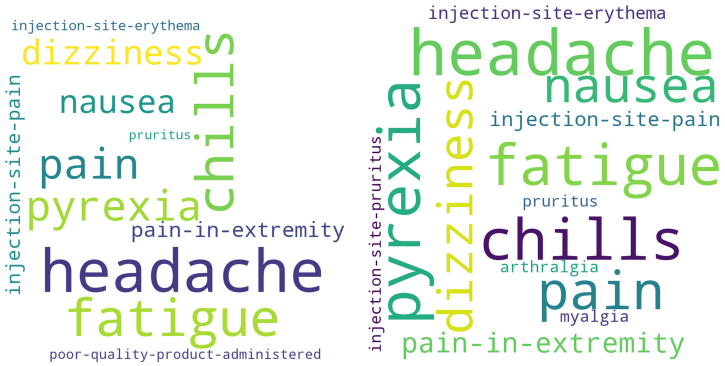
Chapter 6: Clustering of Adverse Events for Post-Market Approved Drugs using Frequent Itemset

Abstract

We study the similarity of adverse effects of COVID-19 vaccines across different states in the United States. We use data of 300,000 COVID-19 vaccine adverse event reports obtained from the Vaccine Adverse Event Reporting System (VAERS). We extract latent topics from the reported adverse events using a topic modeling approach based on Latent Dirichlet allocation (LDA). This approach allows us to represent each U.S state as a low-dimensional distribution over topics. Using Moran's index of spatial autocorrelation we show that some of the topics of adverse events exhibit significant spatial autocorrelation, indicating that there exist spatial clusters of nearby states that exhibit similar adverse events. Using Anselin's local indicator of spatial association we discover and report these clusters. Our results show that adverse events of COVID-19 vaccines vary across states which justifies further research to understand the underlying causality to better understand adverse effects and to reduce vaccine hesitancy.

6.1 Introduction

By June 12th, 2021, more than 2.3 billion doses of various brands of COVID-19 vaccines had been administered world-wide with more than 300 million doses administered in the United States [116]. The U.S. Centers for Disease Control and Prevention (CDC) has stated that all U.S. authorized vaccines are safe and efficient [117]. While generally safe, the COVID-19 vaccines have adverse effects, including common side effects such as injection site pain and fever, but also including rare adverse effects that can be more severe. In the United States alone, by June 1st, 2021, a total of 297,410 of adverse events have been reported, collected,



(a) California (b) Florida

Figure 6.1: COVID-19 Adverse Effect Clouds per Region.

and made publicly available by the CDC and the U.S. Food and Drug Administration in a database called the Vaccine Adverse Event Reporting System (VAERS) [118]. As cases of severe symptoms gain public visibility in the news [119], these seemingly contradicting facts of general safety and possibly severe side-effects are a source of confusion leading to vaccine hesitancy among the population [120].

Towards a better understanding of COVID-19 vaccine adverse events we propose a similarity measure to quantify the similarity of sets of adverse events. To illustrate the challenge tackled in this work, Figure 6.1 shows word clouds of adverse effects for California (Figure 6.1a) and for Florida (Figure 6.1b). These word clouds show the font size of the most frequent adverse effects proportional to their relative frequency observed in that state. We observe that common side effects such as headache, pyrexia (fever), and chills appear with similar relative frequency in both states and we also observe that some adverse effects appear more frequently in one region than another. For example, it pyrexia and dizziess are more frequently observed in Florida. Our goal is to measure the (dis-)similarity of the adverse effects observed in different regions. This similarity allows to understand how reported adverse events vary over space, over time, across different vaccine brands, and across different populations. We use our proposed similarity measure to study if we can

observe statistically significant clusters of regions exhibiting similar adverse effects using VAERS data for the United States. While our work does not answer the question whether vaccines are safe, we hope that public health researchers and health officials may find our similarity measure useful to better understand adverse events, their variations over space, and the underlying causal factors.

Summarizing our approach, we use a bag-of-words model to describe a set of adverse events, such as reported in a spatial region. We leverage Latent Dirichlet Allocation (LDA) to extract latent topics of adverse effects for each region. LDA has been successfully used to extract domains and research topics from scientific research papers [121] and news topics (such as "Sports", "Politics", "Entertainment") from news articles [122]. To extract latent topics of adverse events, we treat the adverse events reported in a spatial region as documents and individual adverse effects as words. We qualitatively evaluate the modeled topics and show that they are able to represent, for example, adverse events related to "pyrexia/fever" and adverse effects related to "vertigo/dizziness". Then, we describe states of the U.S. by their adverse event topic distribution to evaluate whether topics of vaccine adverse effects vary across the United States. We quantitatively evaluate if this variation exhibits any significant spatial autocorrelation, that is, if spatially close states exhibit similar topics of adverse events.

For this purpose, we first survey existing work in Section 6.2 and formally define an adverse event database in Section 6.3. Our approach to extract latent topics of adverse events using topic modeling is described in Section 6.4.1. Using these topics as a low-dimensional embedding of adverse events in a spatial region, our approach to quantify spatial autocorrelation and to find spatial clusters of states that exhibit significantly similar (or dissimilar) topics of adverse effects is described in Section 6.4.2. We explore the global and local spatial autocorrelation of COVID-19 vaccine adverse events in Section 6.5 to discover significant spatial autocorrelation, showing that some topics of adverse events indeed vary in different parts of the United States. Finally, we conclude in Section 6.6 and identify future directions.

6.2 Related Work

Adverse Effects of Vaccines

Vaccines are, without any doubt, a paramount weapon to fight deadly diseases evident by the fact that "In 1900, for every 1,000 babies born in the United States, 100 would die before their first birthday, often due to infectious diseases" [54]. Furthermore, vaccines not only protect those receiving the vaccines but also vulnerable groups around them, such as new born babies, who may not be able to receive a vaccine [55]. Yet, there are adverse effects [118] including the 300,000 adverse events reported for the COVID-19 vaccines by June 1st, 2021. Understanding and mitigating these adverse events will not only improve the well-being of those receiving the vaccines, but will also decrease fear of vaccines that leads to high vaccine hesitancy as observed during the COVID-19 pandemic [56]. To the best of our knowledge, this is the first study investigating the similarity of adverse effects of COVID-19 vaccines to understand their spatial autocorrelation. We hope that our proposed techniques will find adaption by epidemiologists to improve our understanding of the ecology of past, present, and future infectious diseases.

Topic Modeling of Adverse Events

Topic modeling is an unsupervised learning technique to discover underlying themes of a collection of documents. Latent Dirichlet Allocation (LDA) is one of the more common topic modeling techniques in the literature [23]. In the context of pharmacovigilance, LDA has been used to find potentially unsafe dietary supplements [26], but without the consideration of the spatial distribution of latent topics among adverse effects. In our prior work in [50] we performed a spatio-temporal study on the adverse events of blood thinning drugs and their spatial auto-correlation. This study mainly limited by data availability, having adverse events reported by country only. For this reason, our prior study in [50] used European countries, but most countries had to be removed due to having too few reported adverse events. The wide availability of VAERS COVID-19 vaccine data at United States state level

enables us to directly explore the latent adverse event features for spatial auto-correlation.

Pharmacovigilance

The field of pharmacovigilance aims at understanding the occurrence of adverse effects of drugs [8, 9]. Existing work has shown that adverse effects of a single drug or multiple combination of drugs may vary over space and time due to racial and ethnic disparities [10, 11, 12], environment [13, 14], and drug quality [15]. Specifically for vaccines, there is evidence that stress may have an amplifying effect on immune response and adverse events [49]. However, such aspects of understanding the interactions between drugs and other external factors are out of scope of this work. In this work, we investigate the effect of location on adverse effects of the COVID-19 vaccines. While location may be a proxy of other factors (such as stress), this work does not provide or imply any causality between location and adverse events. Yet, we hope that an understanding of the spatial distribution and autocorrleation of adverse events may help experts discover such causalities.

6.3 Problem Definition

This section formally defines adverse events, adverse effects, and the problem of spatio-temporal clustering of adverse events. First, we provide a definition of adverse effects and events.

Definition 11 (Adverse Effect)**.** *An Adverse Effect is a textual representation of an undesirable experiences associated with the use of a medical product. We let* $\mathcal{A} = \{A_1, ..., A_N\}$ *denote the set of all adverse events and* N *denotes the number of all (possible) adverse effects.*

Data such as collected in the VAERS database is a collection of records each associated with a set of adverse effects, a specific pharmaceutical drug, a location, and time. We call such as record an Adverse Event (AE), formally defined as follows:

Table 6.1: Sample records of Adverse Event Report Database. Each Line is an Adverse Event.

Adverse Event ID	Drug	Location	Set of Adverse Effects
1139067	Moderna	MD	Dizziness, Injection site pruritus, Injection site rash, Somnolence
1004857	Moderna	PA	Nausea, Palpitations, Presyncope, Pyrexia, Tremor
1115746	Moderna	NY	Chills,Headache,Nausea,Pain,Pain in extremity
1148711	Moderna	CA	Axillary pain, Fatigue, Headache, Nausea, Pain in extremity
1240185	Pfizer	IN	Fatigue,Headache,Pain,Pyrexia
1120846	Pfizer	UT	Nausea,Pain in extremity, Sleep disorder, Tinnitus, Vertigo
1104541	Pfizer	GA	Injection site reaction, Rash pruritic
1138693	Pfizer	WI	Eye pruritus, Lip swelling, Nasal pruritus, Swelling face, Urticaria
1200860	Janssen	TX	Headache
1114482	Janssen	MI	Chills, Hyperhidrosis, Pyrexia
1244933	Janssen	IL	Heart rate, Heart rate increased, Pain, Poor quality sleep, Pyrexia
1202067	Janssen	RI	Chills, Injection site erythema, Menstruation irregular, Pyrexia

Definition 12 (Vaccine Adverse Event Database)**.** *Let $\mathcal{A}$ denote a set of adverse effects, let $\mathcal{S}$ denote a set of spatial regions, and let $\mathcal{D}$ denote a set of vaccine brands. An Adverse Event Report Database $\mathcal{DB}$ is a collection of adverse event reports (s, A, d), where $s \in \mathcal{S}$ is a spatial region, $A \subseteq \mathcal{A}$ is a set of adverse effects, and $d \in \mathcal{D}$ is the brand for which the adverse effects are reported. We let $M := |\mathcal{DB}|$ denote the number of adverse event reports in $\mathcal{DB}$*

We note that a single adverse event may report multiple adverse effects. As an example, Table 6.1 shows exemplary adverse events from the VAERS database. The first line in Table 6.1 implies that "Dizziness", "Injection site pruritus", "Injection site rash", and "Somnolence" are adverse effects reported in Maryland Moderna vaccine.

Our goal is to find clusters of locations that exhibit similar adverse events. Towards this goal, we group adverse events by region.

Definition 13 (Spatial Adverse Events)**.** *Let $\mathcal{DB}$ be an adverse event report database and let $s' \in \mathcal{S}$ be a spatial region. We define*

$$\mathcal{DB}_{s'} := \{(s, A, d) \in \mathcal{DB} | s = s'\}$$

as the set of all adverse events reported in regions'.

In the next section, we describe how we obtain latent topics of adverse events to represent each region as a low dimensional topic distribution.

6.4 Methodology

This section presents our Latent Dirichlet Allocation (LDA) based approach to extract latent topics from adverse events. All our code to access the data and to run the topic modeling can be found at https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Similarity/tree/main.

6.4.1 Latent Adverse Event Topic Modeling

A challenge of mining adverse events is the potentially large number of different adverse effects. The FAERS Adverse Event Databases use MedDRA codes [86] and terminology to standardize adverse effects such as using "pyrexia" instead of "heightened temperature" of "fever". Yet, the number of possible adverse effects is too large and the resulting feature space of using bag-of-words semantics to represent adverse effects is too high dimensional. To address this issue, we acknowledge that adverse effects are symptoms of unknown (latent) underlying causes. While one way of identifying causes is involving a medical expert, we propose a data-driven approach to identify underlying topics among adverse events using topic modeling that we interpret as causes. For that, we employ Latent Dirichlet Allocation (LDA) [23] – a generative probabilistic model which assumes that each adverse event is a

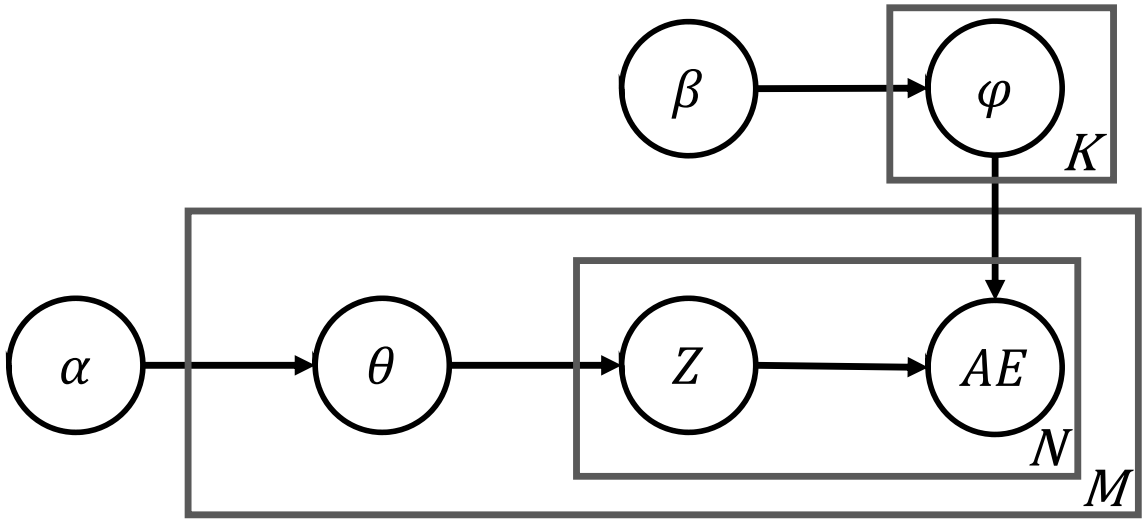


Figure 6.2: LDA Topic Modeling of Adverse Events. For each adverse event a topic distribution θ is estimated and for each topic i, an adverse effect distribution φ_i is estimated. Given a topic Z generated from θ, observable adverse effects (AEs) are generated from φ_Z.

mixture of underlying (latent) topics, and each topic has a (latent) distribution of more and less likely adverse effects.

A graphical representation of our LDA model using plate notation is shown in Fig. 6.2. A vector α of length K is used to parameterize the *a priori* distribution of topics. The parameter K corresponds to the number of latent topics used to model adverse events. When an adverse event is created, we assume that its topics are chosen following a *Dirichlet distribution* having parameter α which we use to obtain a topic distribution θ for each of our $M =$ adverse events. Thus, the large plate in Fig. 6.2 corresponds to a set of M adverse events, each having a topic distribution θ drawn randomly (and Dirichlet distributed) from α.

For each topic, the prior parameter β is used to generate the distribution of adverse effects within a topic. Thus, we assume that a topic generates adverse effects following a Dirichlet distribution having a vector β of length $|\mathcal{A}|$ as parameter, where $\mathcal{A}$ is the set of observed adverse effects (c.f. Definition 11). For each of our K topics, a resulting vector

$\varphi_i, 1 \leq i \leq K$ stores the adverse effect distribution of topic K.

To generate the adverse effects of an adverse event, a topic is chosen randomly from the topic distribution θ and, given this topic, a number of N_i adverse effects are generated randomly from the adverse effect distribution φ – where N_i is assumed to be independent from the chosen topic and uniformly distributed. In Fig. 6.2, the node AE denotes the (observable) set of all $N = \sum_i N_i$ adverse effects, and Z is a function that maps each word to the topic that generated it. The reason for choosing a Dirichlet distribution rather than a more straightforward uniform or multinomial distribution for the topic and word priors is inspired by research showing that the distribution of words in text can be better approximated using a Dirichlet distribution [111].

To infer the topics of our adverse event database $\mathcal{DB}$, we employ a generative process. Given the observed adverse effects, LDA optimizes the latent variables to maximize the likelihood of matching the observed adverse events and corresponding adverse effects. This generative process works as follows. Adverse events are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all N adverse effects. LDA assumes the following generative process for database $\mathcal{DB}$ consisting of M adverse events, each having a number of N_i adverse effects.

- For each adverse event choose a topic distribution $\theta_m \sim Dir(\alpha), 1 \leq m \leq M$, where $Dir(\alpha)$ is a Dirichlet distribution with prior α. In our experiments, we initially assume each topic to have uniform prior probabilities, having $\alpha_i = \alpha_j$ for $1 \leq i, j \leq K$. This apriori distribution is adapted using Bayesian inference [23] to maximize the likelihood of generating the observed keywords.

- For each topic, choose an adverse effect distribution $\varphi_i \sim Dir(\beta)$, where $1 \leq i \leq K$. For our experiments, we assume each adverse effect to have the same prior probability N^{-1}.

- For each adverse effect ae in adverse event j:

    1. Choose a topic $z \sim Multinomial(\theta_j)$ from the topic distribution of j, and

2. Choose a word $w \sim Multinomial(\varphi_z)$ from the adverse effect φ_z of topic z.

Here, $Multinomial(x)$ corresponds to a multinomial distribution drawing from a stochastic vector x.

To describe each adverse event in a latent topic space, we use the adverse event specific topic distributions θ_m which describe each adverse event m as a set of K latent features corresponding to the weight of the respective latent topic. While this topic modeling does not provide us with any semantic of the underlying topics, we know that adverse events having similar latent features also exhibit similar adverse effects. Based on the similarity of latent topics we propose a hierarchical agglomerative clustering approach to find regions that exhibit similar adverse events in Section 6.4.2 and test these clusters for spatial autocorrelation using Moran's I in Section 6.5.

6.4.2 Spatial Clustering of Vaccine Adverse Event Topics

The latent topic modeling of Section 6.4.1 provides us with a topic distribution θ_i for each adverse event report $d \in \mathcal{DB}$. To describe the topic distribution of a region, we use the average topic distribution of all adverse events reported in the region. To measure similarity between the topics of adverse events of two regions, we use Euclidean distance between these resulting average topic distributions. Formally,

Definition 14 (Region-Wise Adverse Event Distance)**.** *Let $\mathcal{DB}$ be an adverse event database, let $\mathcal{DB}_{s_1}, \mathcal{DB}_{s_2} \subseteq \mathcal{DB}$, let K be a positive integer and let $\theta(ae)$ denote the latent topic distribution of an adverse event $ae \in \mathcal{DB}$ using the LDA model described in Section 6.4.1, then:*

$$dist(\mathcal{DB}_{s_1}, \mathcal{DB}_{s_2}) := \left\| \frac{\sum_{\mathcal{DB}_{s_1}} \theta(ae)}{|\mathcal{DB}_{s_1}|} - \frac{\sum_{\mathcal{DB}_{s_2}} \theta(ae)}{|\mathcal{DB}_{s_2}|} \right\|_2,$$

where $\|.\|_2$ denotes the Euclidean norm.

To find clusters among regions having similar topics of adverse events we leverage the

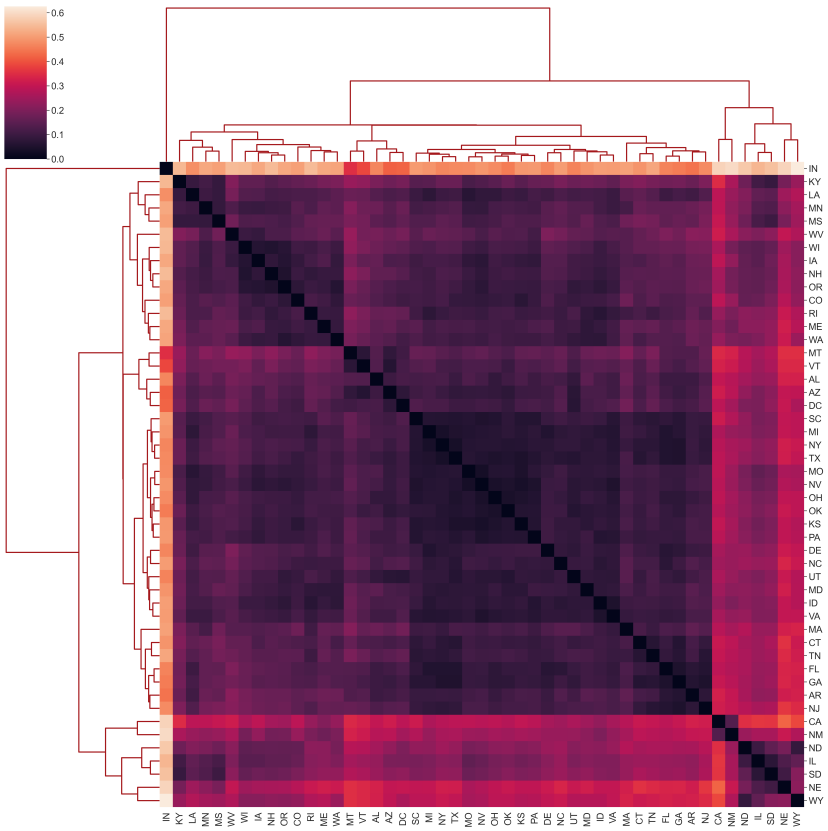


Figure 6.3: Pair-wise similarity matrix of latent topics of COVID-19 vaccine adverse events of counties in the United States.

distance function of Definition 14 and employ a hierarchical agglomerative clustering approach [99]. The advantage of such an approach is that we neither have to guess the number of clusters as often needed for partitioning clustering approaches [100] nor have to define a density threshold as required by density-based clustering algorithms [123, 101]. To merge clusters, we employ complete linkage, which defines the distance between two clusters of regions as the maximum pair-wise distance of regions among the clusters.

Figure 6.3 shows the pair-wise distance (see Definition 13) for each pair of states for the 49 states of the United States excluding Alaska, Puerto Rico, and Hawaii using

$K = 10$ adverse event topics. In Figure 6.3 darker colors correspond to a higher pairwise similarity. We observe a large group of mutually similar states having smaller nested clusters of similar states thus explaining our choice for hierarchical clustering. We also observe that is not trivial to delineate clusters due to noise, which explains our choice of complete link clustering to maximize delineation and avoid having clusters "grow together". A high resolution version of Figure 6.3 can be found on our project website https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Similarity/tree/main.

6.4.3 Spatial Autocorrelation

Given the latent topics of vaccine adverse events as described in Section 6.4.1 and the clustering approach of Section 6.4.2, we next investigate if the observed adverse event topics exhibit significant spatial autocorrelation. In other words, can we reject the null hypothesis that topics are independent of location by observing that spatially close regions exhibit similar topics?

For this purpose, we retain all clusters (of all sizes) corresponding to all nodes in the dendrogram excluding clusters of size one and excluding the root of the dendrogram that contains all regions. Given any such cluster of regions that exhibit similar topics of adverse events, we employ Moran's I measure of spatial autocorrelation [112]. Moran's I statistic tests if a variable measured on spatial regions exhibits a significant spatial autocorrelation, either positive (clustered) or negative (dispersed). To measure the spatial autocorrelation of clusters obtained as described in Section 6.4.2, we use one-hot encoding (or dummy-coding) to encode each individual cluster membership into a binary variable. Thus, for a cluster C, the cluster membership variable of a region r is set to 1 if $r \in C$ and 0 otherwise. Moran's I requires an adjacency metric on regions to assess the similarity between polygonal regions. For this purpose, we employ the *Queen Contiguity* model [124], that is, two regions are considered adjacent if they share boundary. We directly report Moran's I test statistic whose range is in $[-1, -1]$, ranging from strongly dispersed (close to -1) to strongly clustered (close to 1). We also report the p-value of the null-hypothesis that the

Table 6.2: Top-10 most probably adverse effects per topics across all regions and all COVID-19 vaccine brands.

Topic	(Probabilities in %) Adverse Effects
1	(4.5)"headache", (3.6)"pyrexia", (3.6)"fatigue", (3.3)"pain", (3.1)"chills", (3.0)"nausea", (2.3)"pain-in-extremity", (1.7)"dizziness", (1.7)"injection-site-erythema", (1.7)"arthralgia"
2	(4.1)"headache", (2.8)"dizziness", (2.6)"pyrexia", (2.6)"pain-in-extremity", (2.5)"fatigue", (2.5)"chills", (2.4)"nausea", (2.4)"pain", (2.1)"injection-site-pain", (1.6)"dyspnoea"
3	(6.9)"headache", (4.1)"pyrexia", (3.8)"fatigue", (3.7)"chills", (3.0)"pain", (2.9)"dizziness", (2.8)"nausea", (1.9)"pain-in-extremity", (1.8)"injection-site-erythema", (1.8)"injection-site-pain"
4	(8.7)"chills", (8.3)"pyrexia", (7.2)"headache", (7.2)"pain", (6.4)"fatigue", (3.9)"nausea", (3.2)"pain-in-extremity", (2.6)"injection-site-pain", (2.2)"myalgia", (2.1)"dizziness"
5	(4.5)"pyrexia", (4.1)"headache", (4.0)"chills", (3.4)"pain", (3.1)"fatigue", (2.5)"nausea", (2.5)"dizziness", (2.1)"injection-site-pain", (2.1)"arthralgia", (2.1)"pain-in-extremity"
6	(3.8)"dizziness", (3.3)"headache", (2.4)"chills", (2.3)"nausea", (2.2)"fatigue", (2.2)"pain", (2.1)"pain-in-extremity", (1.5)"dyspnoea", (1.5)"injection-site-erythema", (1.5)"pyrexia"
7	(6.5)"headache", (5.5)"pyrexia", (5.1)"chills", (4.8)"pain", (4.7)"fatigue", (3.2)"nausea", (2.6)"injection-site-pain", (2.4)"dizziness", (2.0)"injection-site-erythema", (1.7)"pain-in-extremity"
8	(5.7)"headache", (4.4)"fatigue", (4.0)"chills", (3.8)"pain", (3.2)"pyrexia", (3.0)"pain-in-extremity", (2.7)"nausea", (2.1)"injection-site-pain", (1.8)"injection-site-erythema", (1.8)"dizziness"
9	(4.0)"headache", (3.9)"fatigue", (3.6)"pain", (3.2)"chills", (2.9)"nausea", (2.8)"pyrexia", (2.5)"dizziness", (1.9)"pain-in-extremity", (1.9)"injection-site-pain", (1.6)"pruritus"
10	(3.8)"pyrexia", (3.3)"fatigue", (2.9)"headache", (2.8)"pain", (2.6)"chills", (2.4)"dizziness", (2.1)"nausea", (1.9)"pruritus", (1.9)"rash", (1.9)"injection-site-erythema"

regions are distributed randomly without any spatial pattern by transforming Moran's I values to z-values and employing a two-tailed z-test [113]. The resulting p-values indicate whether a cluster of regions having similar topics of adverse events are significantly spatially clustered or dispersed. We used the geopandas library for handling spatial attributes and Pysal library for Moran's I test of spatial autocorrelation [80, 103].

6.5 Experimental Evaluation

For our experimental evaluation we collected data from the VAERS database as described in Section 6.1 grouped by U.S. states and grouped by the three brands of vaccines authorized by 06/14/2021: Janssen, Moderna, and Pfizer. The experiments are conducted on a PC with Intel(R) Xeon(R) CPU $E3$-1240 v6 @3.70GHz and 32GB RAM. Windows 10 Enterprise

64-bit is the operating system, and all the algorithms are implemented by Python 3.7. All code, including code to obtain data from the VAERS API, is available at https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Similarity/tree/main

6.5.1 Qualitative Analysis of Topics

For $K = 10$ latent topics of COVID-19 adverse events Table 6.2 shows the φ_i vectors of our LDA model which correspond to the adverse effect distribution of the i'th topic. For each topic in Table 6.2 we show the Top-10 highest probability adverse effects. First, we observe that the resulting ten topics are hard to discriminate, as they all contain common adverse effects such as "headache", "pyrexia" (fever). Yet, we do observe different distributions of these adverse effects. We observe that Topic #4 has high probabilities for common symptoms and consequently low probabilities for rare symptoms. Topic #6 seems to corresponds to light symptoms with a low probability of fever, but higher probability of "dizziness". However, we note that our team does not include a medical expert, thus we refrain from a deeper analysis of these topics and conclude that our LDA approach has been able to find topics that differ in distribution of adverse effects. We note that due to truncation to only showing the Top-10 most probable adverse effects, we do not show uncommon and rare adverse effects which may define a topic (thus having most of it's probability mass focused within this single topic). The interested reader may find the full list of adverse effect per topic probabilities on our project website at https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Similarity/tree/main, also including the per-topic adverse effect distributions for $K = 3$ and $K = 20$ topics.

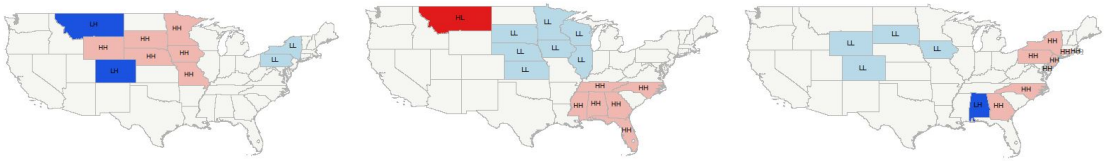
6.5.2 Spatial Anaylsis of COVID-19 Adverse Event Topics

Table 6.3 shows the degree of spatial autocorrelation of each of the $K = 10$ topics of adverse events. For this purpose, we associated each U.S. state i with it's corresponding φ_{ik} probability of topic $k \in \{1, ..., 10\}$. With each states having it's corresponding probability for topic k, we use Moran's I measure of spatial autocorrelation [112]. Moran's I is a test

Table 6.3: Moran's I measure of global spatial autocorrelation for each of the $K = 10$ topics of COVID-19 adverse events.

Pattern	p-value	Moran's Index	z-score	Topic ID
Clustered	0.0006	0.2756	3.4512	1
Random	0.6214	-0.0635	-0.4938	2
Clustered	0.0966	0.1216	1.6616	3
Random	0.6643	-0.0464	-0.4340	4
Random	0.2054	0.0920	1.2662	5
Random	0.6867	0.0109	0.4033	6
Dispersed	0.0754	-0.1785	-1.7782	7
Clustered	0.0071	0.2149	2.6938	8
Random	0.1988	0.0875	1.2850	9
Clustered	0.0002	0.3163	3.7895	10

statistic to test the hypothesis that a spatial phenomenon appears uniformly at random without any spatial pattern. We observe in Table 6.3 that out of the ten topics, six topics show no spatial autocorrelation (unable to reject the null hypothesis of a random pattern), one topic shows negative spatial autocorrelation (implying a significant dispersed pattern), and three topics exhibit a positive spatial autocorrelation (spatially clustered patterns). First, we note testing ten hypothesis, and at the high p-value of 0.0754 we'd expect one such pattern by chance under the null hypothesis. Accounting for the multiple hypothesis testing problem [106] (for example, using Bonferroni correction [105]), the dispersed pattern of Topic #8 is no significant. However, for the clustered patterns of Topics #1 and #8, and #10 we observe highly significant p-value of 0.0006, 0.0071, and 0.0002, respectively, showing that these three topics of COVID-19 adverse events do exhibit significant spatial autocorrelation. This results shows that some latent topics among the adverse effects of the COVID-19 vaccines indeed depend on location. For a deeper study, we show the Local Indicator of Spatial Autocorrelation (LISA) [107] in Figure 6.4, showing the spatial location of clusters of regions that exhibit high (or low) probabilities of the corresponding topic.



(a) Topic #1 (b) Topic #8 (c) Topic #10

Figure 6.4: Local Indicator of Spatial Autocorrelation (LISA). Light red areas correspond to high-high clusters. Light blue areas are low-low clusters. Dark red and dark blue areas corresponds to high-low and low-high outliers.

Using LISA, a cluster is defined as a region having a high (low) value that is surrounded by regions that also have high (low) values. Interestingly, we observe that different parts of the United States exhibit high (low) values in these three significant latent topics. We also observe high-low (low-high) outliers, i.e., regions having high (low) topic probabilities that are surrounded by regions having low (high) topic probabilities. These significant clusters that adverse effects indeed vary locally. The underlying causality warrants further study to understand why certain regions of the United States exhibit different topics of adverse events.

6.6 Conclusion

In this work, we tackled the problem of measuring (dis-)similarity between adverse events of COVID-19 vaccines observed in different regions. Our measure leverages a topic modeling approach using LDA to map each adverse event from a (textual) set of adverse effects to a latent topic distribution. Using a database of 300,000 adverse event reports of COVID-19 vaccines in the United States, investigate the underlying topics exhibit any spatial autocorrelation to understand if different places exhibit different adverse events. Our results show that some of the latent topics of COVID-19 adverse events show significant positive spatial autocorrelation. Our local analysis of spatial autocorrelation show that certain topics

of adverse events have increased (or decreased) likelihood in different parts of the United States.

We hope that teams of medical experts may find this result to investigate the underlying causality. Reasons could be due to vaccine quality issues, storage and cooling issues, or simply due to different brands of vaccines. Our own future work will include looking at the correlation between adverse event topics and different vaccine brands to understand topics and possibly the clusters that we have observed. We will also look into temporal changes of topics to gain an understanding how adverse events may change over time and due to climate.

Finally, we note that all of our implementations, experiments, and results are available at our project website:
https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Similarity/tree/main, where we also include additional experiments which we could not fit into this paper.

Chapter 7: Final Conclusion

This dissertation is an ensemble of published and publishable research papers that improved upon existing pharmacovigilance methods with machine learning algorithms enhanced with spatial science techniques. In this dissertation, we investigate the effect of location on adverse effects of blood thinners drugs in chapter 4, 5 and COVID-19 vaccines in chapter 6.

In chapter 4 and 5, we explore spatial temporal clusters of a three FDA approved drugs for post market AEs patterns and trends. We didn't used concomitant drugs of FAERS dataset and used AEs reports associated to a single drug for Dabigatran, Rivaroxaban and Apixaban. We also explored spatial temporal clusters of three COVID-19 vaccines for AEs patterns and trends. Our goal was to investigate if we can identify spatial clusters of regions that exhibit similar adverse effects using two data mining approaches i.e Frequent Item-set mining and Topic mining using latent dirichlet allocation.

We proposed a first approach to measure the similarity of reported adverse events between spatial regions based on the latent topics of adverse events. Based on this similarity, we proposed a clustering approach to group countries having similar adverse events and evaluated the degree of spatial autocorrelation among regions in the same group. Our experimental has shown that we can indeed find clusters of countries that exhibit similar adverse events. However, we were not able to confirm spatial autocorrelation between these regions. We note that more research in this field is needed.

One limitation of our approach is the aggregation at country level, which may have interesting local spatial patterns. Applying our solutions to smaller spatial regions may find such patterns. We also note that a different measure of spatial proximity may yield stronger autocorrelation by considering not only topological distance but also including political and socioeconomic similarities. To summarize, we did show that some countries exhibit similar topics of adverse events, but an deeper investigation of patterns and their

causality is needed. We hope our approach at mining publicly available adverse event databases improves our understanding of the spatio-temporal change of the adverse effects of a drug.

There are many directions of future work to refine our spatiotemporal topic modeling approach. A first direction is to consider different region comparison due to health disparities, different environmental interaction to drugs and public health surveillance standards. In this work, we chose to mine for topics in europe due to data availablity. A second direction is to use different spatial weights between countries to highlight the influence some countries have on each other such as sharing public standards such as the EU or barriers including language or inaccessible terrain between them such as water or mountains. between countries and their openness to each other and finally, a local measure of spatial auto-correlation such as Anselin's Local Indicator of Spatial Association [107] may be used.

In chapter 6, we explore spatial temporal clusters of three Covid19 vaccines Janssen, Moderna, and Pfizer for AEs patterns and trends. Vaccines are, without any doubt, a paramount weapon to fight deadly diseases evident by the fact that "In 1900, for every 1,000 babies born in the United States, 100 would die before their first birthday, often due to infectious diseases" [54]. Furthermore, vaccines not only protect those receiving the vaccines but also vulnerable groups around them, such as new born babies, who may not be able to receive a vaccine [55]. Understanding and mitigating these adverse events will not only improve the well-being of those receiving the vaccines, but will also decrease fear of vaccines that leads to high vaccine hesitancy as observed during the COVID-19 pandemic [56]. We investigate if we can identify spatial clusters of regions that exhibit similar adverse effects using latent dirichlet allocation.

In this work, we tackled the problem of measuring (dis-)similarity between adverse events of COVID-19 vaccines observed in different regions. Our measure leverages a topic modeling approach using LDA to map each adverse event from a (textual) set of adverse effects to a latent topic distribution. Using a database of 300,000 adverse event reports of COVID-19 vaccines in the United States, investigate the underlying topics exhibit any

spatial autocorrelation to understand if different places exhibit different adverse events. Our results show that some of the latent topics of COVID-19 adverse events show significant positive spatial autocorrelation. Our local analysis of spatial autocorrelation show that certain topics of adverse events have increased (or decreased) likelihood in different parts of the United States.

We hope that teams of medical experts may find this result to investigate the underlying causality. Reasons could be due to vaccine quality issues, storage and cooling issues, or simply due to different brands of vaccines. Our own future work will include looking at the correlation between adverse event topics and different vaccine brands to understand topics and possibly the clusters that we have observed. We will also look into temporal changes of topics to gain an understanding how adverse events may change over time and due to climate.

Bibliography

[1] [Online]. Available: https://sunnah.com/ibnmajah:925

[2] S. Lupkin, "Nearly 1 in 3 recent fda drug approvals followed by major safety action," *Scientific American*, 2017.

[3] A. M. Hochberg, S. J. Reisinger, R. K. Pearson, D. J. O'Hara, and K. Hall, "Using data mining to predict safety actions from fda adverse event reporting system data," *Drug information journal: DIJ/Drug Information Association*, vol. 41, no. 5, pp. 633–643, 2007.

[4] R. Rodriguez-Monguio, M. J. Otero, and J. Rovira, "Assessing the economic impact of adverse drug effects," *Pharmacoeconomics*, vol. 21, no. 9, pp. 623–650, 2003.

[5] M. S. Donaldson, J. M. Corrigan, L. T. Kohn *et al.*, "To err is human: building a safer health system," 2000.

[6] J. C. Kohler, E. Pavignani, M. Michael, N. Ovtcharenko, M. Murru, and P. S. Hill, "An examination of pharmaceutical systems in severely disrupted countries," *BMC International Health and Human Rights*, vol. 12, no. 1, pp. 1–11, 2012.

[7] U. Food, D. Administration *et al.*, "Fda adverse event reporting system (faers) public dashboard," *US Food and Drug Administration*, 2018.

[8] G. Jeetu and G. Anusha, "Pharmacovigilance: a worldwide master key for drug safety monitoring," *Journal of Young Pharmacists*, vol. 2, no. 3, pp. 315–320, 2010.

[9] L. Leyens, M. Reumann, N. Malats, and A. Brand, "Use of big data for drug development and for public and personal health and care," *Genetic epidemiology*, vol. 41, no. 1, pp. 51–60, 2017.

[10] A. Baehr, J. C. Peña, and D. J. Hu, "Racial and ethnic disparities in adverse drug events: a systematic review of the literature," *Journal of racial and ethnic health disparities*, vol. 2, no. 4, pp. 527–536, 2015.

[11] C. Piccardi, J. Detollenaere, P. V. Bussche, and S. Willems, "Social disparities in patient safety in primary care: a systematic review," *International Journal for Equity in Health*, vol. 17, no. 1, p. 114, 2018.

[12] J. S. Okoroh, E. F. Uribe, and S. Weingart, "Racial and ethnic disparities in patient safety," *Journal of patient safety*, vol. 13, no. 3, pp. 153–161, 2017.

[13] F. G. F. Pereira, M. B. C. d. Ataíde, R. L. Silva, E. D. R. Néri, G. C. N. Carvalho, and J. Á. Caetano, "Environmental variables and errors in the preparation and administration of medicines," *Revista brasileira de enfermagem*, vol. 71, no. 3, pp. 1046–1054, 2018.

[14] J.-H. Kang, C.-W. Kim, and S.-Y. Lee, "Nurse-perceived patient adverse events and nursing practice environment," *Journal of Preventive Medicine and Public Health*, vol. 47, no. 5, p. 273, 2014.

[15] A. Chircu, E. Sultanow, and S. P. Saraswat, "Healthcare rfid in germany: an integrated pharmaceutical supply chain perspective," *Journal of Applied Business Research (JABR)*, vol. 30, no. 3, pp. 737–752, 2014.

[16] G. Scripcaru, C. Mateus, and C. Nunes, "A decade of adverse drug events in portuguese hospitals: space-time clustering and spatial variation in temporal trends," *BMC pharmacology and toxicology*, vol. 18, no. 1, p. 34, 2017.

[17] S. Shekhar, Z. Jiang, R. Y. Ali, E. Eftelioglu, X. Tang, V. Gunturi, and X. Zhou, "Spatiotemporal data mining: a computational perspective," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2306–2338, 2015.

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[19] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser, "Association rules and data mining in hospital infection control and public health surveillance," *Journal of the American medical informatics association*, vol. 5, no. 4, pp. 373–381, 1998.

[20] M. J. Berry and G. S. Linoff, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2004.

[21] V. Kumar, "Data mining algorithms," *Tutorial at IPAM*, 2002.

[22] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European conference on information retrieval*. Springer, 2011, pp. 338–349.

[25] X. Teng, J. Yang, J.-S. Kim, G. Trajcevski, A. Züfle, and M. A. Nascimento, "Fine-grained diversification of proximity constrained queries on road networks," in *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, 2019, pp. 51–60.

[26] Y. Wang, D. R. Gunashekar, T. J. Adam, and R. Zhang, "Mining adverse events of dietary supplements from product labels by topic modeling," *Studies in health technology and informatics*, vol. 245, p. 614, 2017.

[27] F. Khan and D. Singh, "Association rule mining in the field of agriculture: a survey," *International Journal of Scientific and Research Publications*, vol. 329, 2014.

[28] K. Kreimeyer, D. Menschik, S. Winiecki, W. Paul, F. Barash, E. J. Woo, M. Alimchandani, D. Arya, C. Zinderman, R. Forshee *et al.*, "Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems," *Drug Safety*, vol. 40, no. 7, pp. 571–582, 2017.

[29] T. Botsis, T. Buttolph, M. D. Nguyen, S. Winiecki, E. J. Woo, and R. Ball, "Vaccine adverse event text mining system for extracting features from vaccine safety reports," *Journal of the American Medical Informatics Association*, vol. 19, no. 6, pp. 1011–1018, 2012.

[30] R. Ball and T. Botsis, "Can network analysis improve pattern recognition among adverse events following immunization reported to vaers?" *Clinical Pharmacology & Therapeutics*, vol. 90, no. 2, pp. 271–278, 2011.

[31] S. Gao, J. Rao, Y. Kang, Y. Liang, and J. Kruse, "Mapping county-level mobility pattern changes in the united states in response to covid-19," *SIGSpatial Special*, vol. 12, no. 1, pp. 16–26, 2020.

[32] J. Elarde, J.-S. Kim, H. Kavak, A. Züfle, and T. Anderson, "Change of human mobility during covid-19: A united states case study," *PloS one*, vol. 16, no. 11, p. e0259031, 2021.

[33] R. Hinch, W. J. Probert, A. Nurtay, M. Kendall, C. Wymant, M. Hall, K. Lythgoe, A. Bulas Cruz, L. Zhao, A. Stewart *et al.*, "Openabm-covid19—an agent-based model for non-pharmaceutical interventions against covid-19 including contact tracing," *PLoS computational biology*, vol. 17, no. 7, p. e1009146, 2021.

[34] J. Pesavento, A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson, and A. Züfle, "Data-driven mobility models for covid-19 simulation," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, 2020, pp. 29–38.

[35] A. Züfle, C. Wenk, D. Pfoser, A. Crooks, J.-S. Kim, H. Kavak, U. Manzoor, and H. Jin, "Urban life: a model of people and places," *Computational and Mathematical Organization Theory*, pp. 1–32, 2021.

[36] J.-S. Kim, H. Kavak, C. O. Rouly, H. Jin, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle, "Location-based social simulation for prescriptive analytics of disease spread," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 53–61, 2020.

[37] T. Anderson, J. Yu, and A. Züfle, "The 1st acm sigspatial international workshop on modeling and understanding the spread of covid-19," *SIGSPATIAL Special*, vol. 12, no. 3, pp. 35–40, 2021.

[38] T. Anderson, J. Yu, A. Roess, H. Kavak, J.-S. Kim, and A. Züfle, "Proceedings of the 2nd acm sigspatial international workshop on spatial computing for epidemiology (spatialepi 2021)," 2021.

[39] A. Züfle and T. Anderson, "Introduction to this special issue: Modeling and understanding the spread of covid-19: (part i)," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 1–2, 2020.

[40] ——, "Introduction to this special issue: Modeling and understanding the spread of covid-19: (part ii)," *SIGSPATIAL Special*, vol. 12, no. 2, pp. 1–2, 2020.

[41] S. Islam, D. Gandhi, J. Elarde, T. Anderson, A. Roess, T. F. Leslie, H. Kavak, and A. Züfle, "Spatiotemporal prediction of foot traffic," in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, 2021, pp. 1–8.

[42] A. Sikder and A. Züfle, "Augmenting geostatistics with matrix factorization: a case study for house price estimation," *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 288, 2020.

[43] ——, "Emotion predictions in geo-textual data using spatial statistics and recommendation systems," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, 2019, pp. 1–4.

[44] N. Hubig, P. Fengler, A. Züfle, R. Yang, and S. Günnemann, "Detection and prediction of natural hazards using large-scale environmental data," in *International Symposium on Spatial and Temporal Databases*. Springer, 2017, pp. 300–316.

[45] U. Qazi, M. Imran, and F. Ofli, "Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.

[46] M. Mokbel, S. Abbar, and R. Stanojevic, "Contact tracing: Beyond the apps," *SIGSPATIAL Special*, vol. 12, no. 2, pp. 15–24, 2020.

[47] L. Han, R. Ball, C. A. Pamer, R. B. Altman, and S. Proestel, "Development of an automated assessment tool for medwatch reports in the fda adverse event reporting system," *Journal of the American Medical Informatics Association*, vol. 24, no. 5, pp. 913–920, 2017.

[48] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[49] R. Glaser, J. K. KIECOLT-GLASER, W. B. Malarkey, and J. F. Sheridan, "The Influence of Psychological Stress on the Immune Response to Vaccines," *Annals of the New York Academy of Sciences*, vol. 840, no. 1, pp. 649–655, 1998.

[50] A. Askar and A. Zuefle, "Clustering of adverse events of post-market approved drugs," in *17th International Symposium on Spatial and Temporal Databases*, 2021, pp. 106–115.

[51] A. Askar and A. Züfle, "Clustering adverse events of covid-19 vaccines across the united states," in *International Conference on Similarity Search and Applications*. Springer, 2021, pp. 307–320.

[52] P. A. Noseworthy, X. Yao, N. S. Abraham, L. R. Sangaralingham, R. D. McBane, and N. D. Shah, "Direct comparison of dabigatran, rivaroxaban, and apixaban for effectiveness and safety in nonvalvular atrial fibrillation," *Chest*, vol. 150, no. 6, pp. 1302–1312, 2016.

[53] O.-C. W. Rutherford, C. Jonasson, W. Ghanima, F. Söderdahl, and S. Halvorsen, "Comparison of dabigatran, rivaroxaban, and apixaban for effectiveness and safety in atrial fibrillation: a nationwide cohort study," *European Heart Journal-Cardiovascular Pharmacotherapy*, vol. 6, no. 2, pp. 75–85, 2020.

[54] K. Stratton, A. Ford, E. Rusch, E. Clayton, C. to Review Adverse Effects of Vaccines *et al.*, "Adverse effects of vaccines: Evidence and causality," 2011.

[55] J. Dushoff, J. B. Plotkin, C. Viboud, L. Simonsen, M. Miller, M. Loeb, and J. David, "Vaccinating to protect a vulnerable subpopulation," *PLoS Med*, vol. 4, no. 5, p. e174, 2007.

[56] A. A. Dror, N. Eisenbach, S. Taiber, N. G. Morozov, M. Mizrachi, A. Zigron, S. Srouji, and E. Sela, "Vaccine hesitancy: the next challenge in the fight against covid-19," *European journal of epidemiology*, vol. 35, no. 8, pp. 775–779, 2020.

[57] L. Woods-Burnham, J. Johnson, S. Hooker, F. Bedell, T. Dorff, and R. Kittles, "The role of diverse populations in us clinical trials," *Med*, vol. 2, no. 1, pp. 21–24, Jan. 2021, funding Information: L.W.B. is supported by NIH grant number 1T32CA186895. Publisher Copyright: © 2020 Elsevier Inc.

[58] A. C. Mak, M. J. White, W. L. Eckalbar, Z. A. Szpiech, S. S. Oh, M. Pino-Yanes, D. Hu, P. Goddard, S. Huntsman, J. Galanter *et al.*, "Whole-genome sequencing of pharmacogenetic drug response in racially diverse children with asthma," *American journal of respiratory and critical care medicine*, vol. 197, no. 12, pp. 1552–1564, 2018.

[59] L. Dean, "Abacavir therapy and hla-b* 57: 01 genotype," *Medical Genetics Summaries [Internet]. Bethesda (MD): National Center for Biotechnology Information (US)*, 2018.

[60] A. Sanchez-Mazas, V. Černỳ, D. Di, S. Buhler, E. Podgorná, E. Chevallier, L. Brunet, S. Weber, B. Kervaire, M. Testi *et al.*, "The hla-b landscape of africa: signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection," *Molecular ecology*, vol. 26, no. 22, pp. 6238–6252, 2017.

[61] L. S. Karliner, A. Auerbach, A. Nápoles, D. Schillinger, D. Nickleach, and E. J. Pérez-Stable, "Language barriers and understanding of hospital discharge instructions," *Medical care*, vol. 50, no. 4, p. 283, 2012.

[62] B. P. Monahan, C. L. Ferguson, E. S. Killeavy, B. K. Lloyd, J. Troy, and L. R. Cantilena, "Torsades de pointes occurring in association with terfenadine use," *Jama*, vol. 264, no. 21, pp. 2788–2790, 1990.

[63] D. Genser, "Food and drug interaction: consequences for the nutrition/health status," *Annals of Nutrition and Metabolism*, vol. 52, no. Suppl. 1, pp. 29–32, 2008.

[64] Y. Lurie, R. Loebstein, D. Kurnik, S. Almog, and H. Halkin, "Warfarin and vitamin k intake in the era of pharmacogenetics," *British journal of clinical pharmacology*, vol. 70, no. 2, pp. 164–170, 2010.

[65] S. Murphy and R. Roberts, ""black box" 101: how the food and drug administration evaluates, communicates, and manages drug benefit/risk," *Journal of allergy and clinical immunology*, vol. 117, no. 1, pp. 34–39, 2006.

[66] J. M. van Tongeren, S. F. Harkes-Idzinga, H. van der Sijs, R. Atiqi, B. J. van den Bemt, L. W. Draijer, D. Hiel, A. Kerremans, B. Kremers, M. de Leeuw *et al.*, "The development of practice recommendations for drug-disease interactions by literature review and expert opinion," *Frontiers in pharmacology*, vol. 11, p. 707, 2020.

[67] M. M. Diesveld, S. de Klerk, P. Cornu, D. Strobach, K. Taxis, and S. D. Borgsteede, "Management of drug-disease interactions: a best practice from the netherlands," *International journal of clinical pharmacy*, vol. 43, no. 6, pp. 1437–1450, 2021.

[68] A. Poleksic and L. Xie, "Database of adverse events associated with drugs and drug combinations," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[69] U.S. Food and Drug Administration, "January 31, 2018: New england compounding center pharmacist sentenced for role in nationwide fungal meningitis outbreak." [Online]. Available: https://www.fda.gov

[70] U. D. of Health, H. Services *et al.*, "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule," *US Department of Health and Human Services, Washington, DC) Available at: https://www. hhs. gov/hipaa/for-professionals/privacy/special-topics/de-identification/index. html. Accessed September*, vol. 26, p. 2018, 2012.

[71] T. T. Shimabukuro, M. Nguyen, D. Martin, and F. DeStefano, "Safety monitoring in the vaccine adverse event reporting system (vaers)," *Vaccine*, vol. 33, no. 36, p. 4398–4405, Aug 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264410X15009822

[72] V. Janmey and P. L. Elkin, "Re-identification risk in hipaa de-identified datasets: The mva attack," in *AMIA Annual Symposium Proceedings*, vol. 2018. American Medical Informatics Association, 2018, p. 1329.

[73] D. McGraw, "Building public trust in uses of health insurance portability and accountability act de-identified data," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 29–34, 2013.

[74] U.S. Food and Drug Administration (FDA), "FDA Adverse Event Reporting System (FAERS) Quarterly Data Extract Files (url:https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html)."

[75] ——, "Open Data API openFDA (url:https://open.fda.gov/)."

[76] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.

[77] S. Raschka, "Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack," *The Journal of Open Source Software*, vol. 3, no. 24, Apr. 2018. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00638

[78] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[79] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[80] K. Jordahl, "Geopandas: Python tools for geographic data," *URL: https://github.com/geopandas/geopandas*, 2014.

[81] S. J. Rey and L. Anselin, "PySAL: A Python Library of Spatial Analytical Methods," *The Review of Regional Studies*, vol. 37, no. 1, pp. 5–27, 2007.

[82] J. Heaton, "Comparing dataset characteristics that favor the apriori, eclat or fp-growth frequent itemset mining algorithms," in *SoutheastCon 2016*. IEEE, 2016, pp. 1–7.

[83] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[84] S. T. Dumais *et al.*, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.

[85] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," *Advances in neural information processing systems*, vol. 17, 2004.

[86] E. G. Brown, L. Wood, and S. Wood, "The medical dictionary for regulatory activities (meddra)," *Drug safety*, vol. 20, no. 2, pp. 109–117, 1999.

[87] J. W. Ratcliff and D. E. Metzener, "Pattern-matching-the gestalt approach," *Dr Dobbs Journal*, vol. 13, no. 7, p. 46, 1988.

[88] D. Z. Sui, "Tobler's first law of geography: A big idea for a small world?" *Annals of the Association of American Geographers*, vol. 94, no. 2, pp. 269–277, 2004.

[89] M. Edelstein, L. M. Lee, A. Herten-Crabb, D. L. Heymann, and D. R. Harper, "Strengthening global public health surveillance through data and benefit sharing," *Emerging infectious diseases*, vol. 24, no. 7, p. 1324, 2018.

[90] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, "How will country-based mitigation measures influence the course of the covid-19 epidemic?" *The Lancet*, vol. 395, no. 10228, pp. 931–934, 2020.

[91] Office of the Commissioner, "FDA warns repackers distributing pharmaceutical ingredients, including opioids, for putting consumers at risk with significant violations of manufacturing quality standards," Mar 2020. [Online]. Available: https://www.fda.gov

[92] Anonymous, "Xarelto," Sep 2018. [Online]. Available: https://www.ema.europa.eu/en/medicines/human/EPAR/xarelto

[93] ——, "Pradaxa," Sep 2018. [Online]. Available: https://www.ema.europa.eu/en/medicines/human/EPAR/pradaxa

[94] ——, "Eliquis," Sep 2018. [Online]. Available: https://www.ema.europa.eu/en/medicines/human/EPAR/eliquis

[95] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-k frequent closed patterns without minimum support," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 2002, pp. 211–218.

[96] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, pp. 51–56.

[97] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[98] P. E. Black, "Ratcliff/obershelp pattern recognition," *Dictionary of algorithms and data structures*, vol. 17, 2004.

[99] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of classification*, vol. 1, no. 1, pp. 7–24, 1984.

[100] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[101] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[102] A. X. Y. Carvalho, P. H. M. Albuquerque, G. R. de Almeida Junior, and R. D. Guimaraes, "Spatial hierarchical clustering," *Revista Brasileira de Biometria*, vol. 27, no. 3, pp. 411–442, 2009.

[103] S. J. Rey and L. Anselin, "Pysal: A python library of spatial analytical methods," in *Handbook of applied spatial analysis*. Springer, 2010, pp. 175–193.

[104] A. Getis, "Reflections on spatial autocorrelation," *Regional Science and Urban Economics*, vol. 37, no. 4, pp. 491–496, 2007.

[105] E. W. Weisstein, "Bonferroni correction," *https://mathworld. wolfram. com/*, 2004.

[106] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[107] L. Anselin, "Local indicators of spatial association—lisa," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[108] H. J. Duggirala, J. M. Tonning, E. Smith, R. A. Bright, K. Bouri *et al.*, "Use of data mining at the food and drug administration," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 428–434, 2016.

[109] D. A. Kessler *et al.*, "Introducing medwatch: a new approach to reporting medication and device adverse effects and product problems," *Jama*, vol. 269, no. 21, pp. 2765–2768, 1993.

[110] P. Mozzicato, "Meddra," *Pharmaceutical Medicine*, vol. 23, no. 2, pp. 65–75, 2009.

[111] R. E. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the dirichlet distribution," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 545–552.

[112] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.

[113] W. J. Dixon and F. J. Massey Jr, *Introduction to statistical analysis.* McGraw-Hill, 1951.

[114] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 490–499.

[115] C. W. Dunnett, "A multiple comparison procedure for comparing several treatments with a control," *Journal of the American Statistical Association*, vol. 50, no. 272, pp. 1096–1121, 1955.

[116] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time. (https://coronavirus.jhu.edu/map.html)," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[117] Centers for Disease Control and Prevention, "Different COVID-19 Vaccines. (https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html)."

[118] FDA and CDC, "Vaccine adverse event reporting system," *vaers.hhs.gov*, 2021.

[119] PolitiFact, The Poynter Institute, "Federal VAERS database is a critical tool for researchers, but a breeding ground for misinformation. (https://www.politifact.com/article/2021/may/03/vaers-governments-vaccine-safety-database-critical/)."

[120] M. Sallam, "Covid-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates," *Vaccines*, vol. 9, no. 2, p. 160, 2021.

[121] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[122] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 248–256.

[123] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, 1996, pp. 226–231.

[124] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

Curriculum Vitae

Ahmed Askar received his Bachelors of Science in Biology from The Ohio State University in December of 2008, and graduated from Wright State University with a Master's of Public Health in May 2013. Ahmed has worked in public health research and spatial science community at the university level, local government and at the Federal Government since 2008. He has pursued a career in Health Geography and Data Science and currently works at United States Food and Drug Administration since 2011.