Spatio-Temporal Clustering of Adverse Events for Post-Market Approved Drugs

Askar, Ahmed and Zufle, Andreas

Food And Drug Administration / Office of Commissioner -- George Mason University / Geography and Geoinformation Science

Abstract

Our goal of this study is to propose a framework to explore spatiotemporal adverse event surveillance model by identifying most frequent item-sets, which co-locate together and is associated with FDA approved drug using machine learning and spatial science. Adverse side effects of a drug may vary over space and time due to different populations, environments, drug quality and standard of public health surveillance. A major challenge in drug development is determining possible pitfalls. Discovering all side effects during the development process is not plausible. Once the drug is approved and available to the public, regulators rely on a combination of surveillance, reporting (by doctors and patients) and data mining to identify any post-market issues with the drug. Our approach can be applied to other drugs, vaccines or medical devices to study if spatially localized adverse effects may justify further investigation.

Introduction

Public health surveillance is the base of effective public health practice , and it has been put in the spotlight due to the coronavirus disease pandemic, which has impacted aspects of public health governance in its response and recovery . Pharmacovigilance is a branch of public health surveillance and is the detection, assessment, understanding, and prevention of adverse effects . We propose an approach to find spatial autocorrelation among adverse effects using data mining and spatial statistics to support pharmacovigilance. 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 . Adverse Events (AEs) are any undesirable experiences associated with the use of a medical product. To monitor and track Adverse Events, the FDA has developed an Adverse Event self-reporting tool called MedWatch . 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 practitioner . There are many existing studies, which focus on finding significant links between drugs and adverse effects however theses studies do not put any consideration on spatiotemporal links. Location and time are an important variable for public health surveillance and incorporating it in pharmacovigilance studies might give new insight. Could some patterns between drugs and adverse effects be explained by spatial distribution? 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 has currently left it unanswered.

Figure 1. Does adverse event repeat across space and time?

Materials and Methods

Data Collection:
As of 12/01/2020, there are 19 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 or by OpenFDA API. FAERS data comes in multiple files with a primary key linking all files.

Frequent (k) Adverse Event Set Mining (TopkFAE):
We leverage a variant of the classic Apriori algorithm, which mines Top-k frequent itemsets in transaction databases. Our 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 prune the tree to avoid evaluating itemsets whose support must be lower than the selected k 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 large, then a large set of rare AEs may create noise that may cause regions to appear too more dissimilar. Our experiments have shown that values of k ∈ {5, 10} provide a good balance. Our study area is Europe, countries available for analysis may change due to data availability to several parameters such as k, drug and time frame of AE.

Similarity Measure between Sets of Adverse Events:
To measure the similarity between the Top-k most frequent adverse effects TopkFAE(DBR_1,Y) and TopkFAE(DBR_2,Y) of two countries R_1 and R_2 in a year Y, we use text similarity measure known as Gestalt Pattern Matching.

$$\text{Gestalt}(S_1,S_2) = \frac{2*LCSS(S_1,S_2)}{|S_1|+|S_2|}$$, where in LCSS(S_1,S_2) is the longest common subsequence

Next, we use the Gestalt Pattern Matching similarity function to define a distance function by subtracting from one.

$$\text{Distance}((\text{TopkFAE}(DBR_1,Y),(\text{TopkFAE}(DBR_2,Y)) = 1-\text{Gestalt}(\text{TopkFAE}(DBR_1,Y),\text{TopkFAE}(DBR_2,Y))$$

Next, we use hierarchical agglomerative clustering algorithm to cluster regions by adverse events using the output from distance function above. The closer the similarity coefficients, the more robust the clusters generated from hierarchical agglomerative clustering.

Measure of Spatial Autocorrelation:
First Law of Geography: Everything is related to everything else. But near things are more related than distant things. Moran I was used to test for spatial dependencies among our spatial cluster.

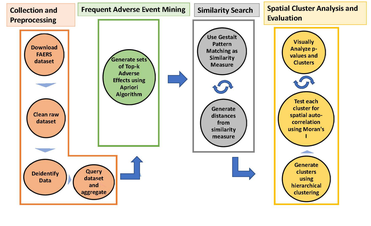

Figure 2. Road map for spatiotemporal association model using Adverse Events Report Submission to FDA FAERS Database

Results and Discussion

For our experimental evaluation we collected AEs data in Europe from the FAERS database for three anticoagulant medications - Apixaban, Rivaroxaban and Dabigatran grouped by four years (2014-2017) and using values for k∈{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 1. This table shows only and all clusters having P-values less than 0.1 before Bonferroni correction. 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"). We conclude that after Bonferroni correction to account for the multiple hypothesis testing problem, none of the clusters of countries having similar adverse events exhibit a significant and interesting spatial auto-correlation.

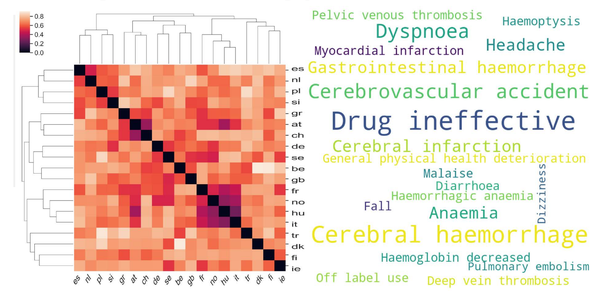

Figure 3. Results for Drug Rivaroxaban,k=5, Year 2014. Distance matrix and hierarchical clustering(left). Adverse effect cloud weighted by frequency across the cluster of countries {at, be, ch, de, gb, gr,se} (right)

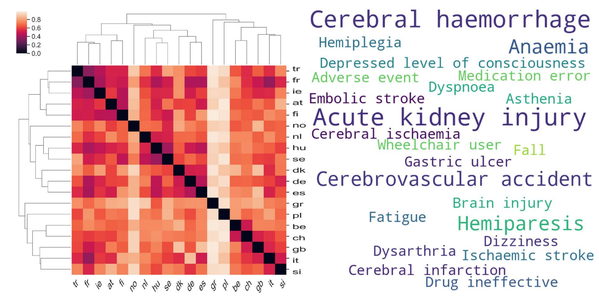

Figure 4. Results for Drug Rivaroxaban,k=10, Year 2016. Distance matrix and hierarchical clustering(left). Adverse effect cloud weighted by frequency across the cluster of countries {de, dk, es} (right)

Table 1. Statistically significant clusters of AEs for Rivaroxaban, Dabigatran and Apixaban for Year 2014-2017 in Europe fork∈{5, 10}

Fig	Spatial Pattern	P-value	P-value Bonferroni correction	Moran's I	List of Countries (two-letter ISO code)	#	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.3	Clustered	0.03818	0.4311	0.43452	at, be, ch, de, gb, gr, se	7	2014	Rivaroxaban	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	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

Conclusion

We conclude that our proposed approach was able to find spatial clusters among adverse events of drugs however they did not have significant spatial auto-correlation. This is possibly due to data aggregation in FAERS, which may contribute to intricate spatial pattern lose in the data. We hope our first approach at mining publicly available adverse event databases (FAERS), improves our understanding of the spatio-temporal change of the adverse effects of a drug.

Limitation

Apriori algorithm is memory exhausting and can benefit from high performance computing environment in order to expand the geographic and temporal coverage in future research

Data aggregation in FAERS is too conservation going by HIPPA regulation, which states that data needs to be aggregated in zip-codes less than 20000 population. Spatial pattern analysis of adverse events and its intricacies is lost due to strict data aggregation and can benefit from data with some spatial information similar to the VAERS database.

References

References available upon request.