# Clustering Adverse Events of COVID-19 Vaccines across the United States

Ahmed Askar[1,2][0000−0002−4537−5418] and Andreas Züfle[1][0000−0001−7001−4123]

[1] Department of Geography and Geoinformation Science, George Mason University,
4400 University Drive, MS 6C3, Fairfax, VA 22030, USA; aaskar@gmu.edu,
azufle@gmu.edu

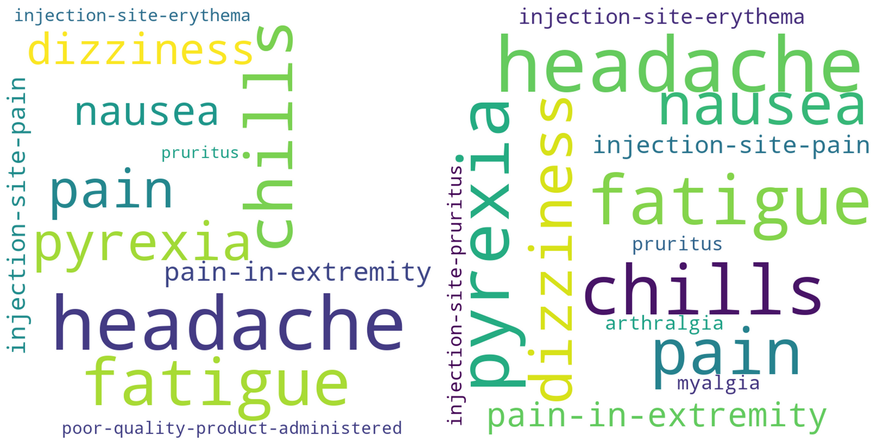
[2] Food and Drug Administration, Silver Spring, MD 20993, USA

**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.

**Keywords:** Spatial Clustering, COVID-19, Vaccines, Adverse Events, Similarity Search, Pharmacovigilance, Health Geography

## 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 [10]. The U.S. Centers for Disease Control and Prevention (CDC) has stated that all U.S. authorized vaccines are safe and efficient [6]. 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, 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) [14]. As cases of severe symptoms gain public visibility in the news [28], these seemingly contradicting facts of general safety and possibly severe side-effects are a source of confusion leading to vaccine hesitancy among the population [31].



(a) California

(b) Florida

Fig. 1: COVID-19 Adverse Effect Clouds per Region.

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 1 shows word clouds of adverse effects for California (Figure 1a) and for Florida (Figure 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 [17] and news topics (such as “Sports”, “Politics”, “Entertainment”) from news articles [29]. 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 2 and formally define an adverse event database in Section 3. Our approach to extract latent topics of adverse events using topic modeling is described in Section 4. 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 5. We explore the global and local spatial autocorrelation of COVID-19 vaccine adverse events in Section 7 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 8 and identify future directions.

## 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” [34]. 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 [12]. Yet, there are adverse effects [14] 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 [11]. 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 [4]. In the context of pharmacovigilance, LDA has been used to find potentially unsafe dietary supplements [35], but without the consideration of the spatial distribution of latent topics among adverse effects. In our prior work in [2] 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 [2] 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 [18,21]. 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 [3,27,25], environment [26,20], and drug quality [7]. Specifically for vaccines, there is evidence that stress may have an amplifying effect on immune response and adverse events [16]. 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.

## 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 1 (Adverse Effect).** *An Adverse Effect is a textual representation of an undesirable experiences associated with the use of a medical product. We let $\mathcal{A} = \{A_1, ..., A_N\}$ 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:

**Definition 2 (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 1 shows exemplary adverse events from the VAERS database. The first line in Table 1 implies that “Dizziness”, “Injection site pruritus”, “Injection site rash”, and “Somnolence” are adverse effects reported in Maryland Moderna vaccine.

| 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 |

Table 1: Sample records of Adverse Event Report Database. Each Line is an Adverse Event.

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

**Definition 3 (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.

## 4 Latent Adverse Event Topic Modeling

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.

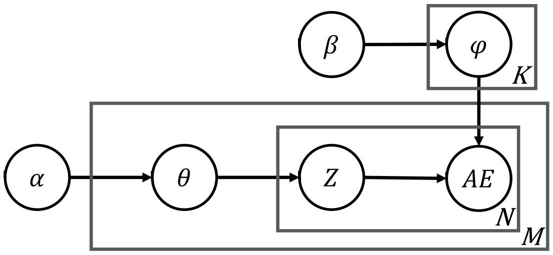


Fig. 2: LDA Topic Modeling of Adverse Events. For each adverse event a topic distribution $\theta$ is estimated and for each topic $i$, an adverse effect distribution $\varphi_i$ is estimated. Given a topic $Z$ generated from $\theta$, observable adverse effects (AEs) are generated from $\varphi_Z$.

A challenge of mining adverse events is the potentially large number of different adverse effects. The FAERS Adverse Event Databases use MedDRA codes [5] 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) [4] – 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. 2. A vector $\alpha$ 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 $\alpha$ which we use to obtain a topic distribution $\theta$ for each of our $M =$ adverse events. Thus, the large plate in Fig. 2 corresponds to a set of $M$ adverse events, each having a topic distribution $\theta$ drawn randomly (and Dirichlet distributed) from $\alpha$.

For each topic, the prior parameter $\beta$ 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 $\beta$ of length $|\mathcal{A}|$ as parameter, where $\mathcal{A}$ is the set of observed adverse effects (c.f. Definition 1). 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 $\theta$ and, given this topic, a number of $N_i$ adverse effects are generated randomly from the adverse effect distribution $\varphi$ – where $N_i$ is assumed to be independent from the chosen topic and uniformly distributed.

In Fig. 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 [23].

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 $\alpha$. 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 [4] 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 $\varphi_z$ 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 $\theta_m$ 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 5 and test these clusters for spatial autocorrelation using Moran's I in Section 7.

## 5 Spatial Clustering of Vaccine Adverse Event Topics

The latent topic modeling of Section 4 provides us with a topic distribution $\theta_i$ 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 4 (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 4, 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 distance function of Definition 4 and employ a hierarchical agglomerative clustering approach [8]. The advantage of such an approach is that we neither have to guess the number of clusters as often needed for partitioning clustering approaches [22] nor have to define a density threshold as required by density-based clustering algorithms [13,32]. 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 3 shows the pair-wise distance (see Definition 3) 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 3 darker colors correspond to a higher pair-wise 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 3 can be found on our project website https://github.com/ahmedaskar64/Spatio-Temporal-AEs-Similarity/tree/main.

## 6 Spatial Autocorrelation

Given the latent topics of vaccine adverse events as described in Section 4 and the clustering approach of Section 5, 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 [24]. Moran's I statistic tests if a variable measured on spatial regions exhibits a significant spatial autocorrelation, either positive (clustered) or

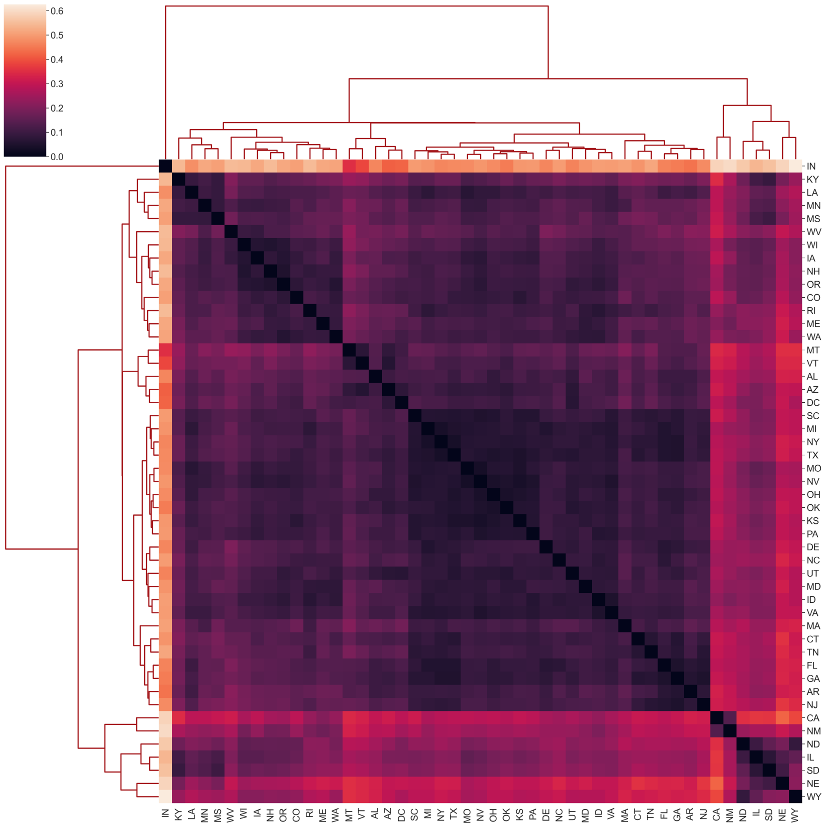


Fig. 3: Pair-wise similarity matrix of latent topics of COVID-19 vaccine adverse events of counties in the United States.

negative (dispersed). To measure the spatial autocorrelation of clusters obtained as described in Section 5, 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 [15], 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 [9]. 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 [19,30].

| 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” |

Table 2: Top-10 most probably adverse effects per topics across all regions and all COVID-19 vaccine brands.

## 7 Experimental Evaluation

For our experimental evaluation we collected data from the VAERS database as described in Section 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.

### 7.1 Qualitative Analysis of Topics

For $K = 10$ latent topics of COVID-19 adverse events Table 2 shows the $\varphi_i$ vectors of our LDA model which correspond to the adverse effect distribution of the $i$'th topic. For each topic in Table 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

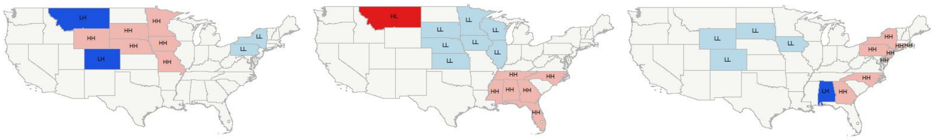
| Pattern | p-value | Moran's Index | z-score | Topic ID |
|---|---|---|---|---|
| Clustered | 0.0006 | 0.2756 | 3.4512 | 1 |
| Random | 0.6214 | -0.0635 | -0.4938 | 2 |
| Clustered | 0.0966 | 0.1216 | 1.6616 | 3 |
| Random | 0.6643 | -0.0464 | -0.4340 | 4 |
| Random | 0.2054 | 0.0920 | 1.2662 | 5 |
| 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 |

Table 3: Moran's I measure of global spatial autocorrelation for each of the $K = 10$ topics of COVID-19 adverse events.

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 (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.

### 7.2 Spatial Anaylsis of COVID-19 Adverse Event Topics

Table 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 $\varphi_{ik}$ 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 [24]. Moran's I is a test statistic to test the hypothesis that a spatial phenomenon appears uniformly at random without any spatial pattern. We observe in Table 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 [33] (for example, using Bonferroni correction [36]), the dispersed pattern of Topic #7 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



(a) Topic #1 (b) Topic #7 (c) Topic #10

Fig. 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.

the Local Indicator of Spatial Autocorrelation (LISA [1]) in Figure 4, showing the spatial location of clusters of regions that exhibit high (or low) probabilities of the corresponding topic. 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.

## 8 Conclusions

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.

## 9 Acknowledgements

This research was prepared or accomplished by Ahmed Askar in his personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of the U.S Food and Drug Administration, the Department of Health and Human Services, or the United States government. This research received no external funding.

## References

1. Anselin, L.: Local indicators of spatial association—lisa. Geographical analysis **27**(2), 93–115 (1995)
2. Askar, A., Züfle, A.: Spatio-Temporal Clustering of Adverse Events of Post-Market Approved Drugs using Latent Dirichlet Allocation. In: Proceedings of the 17th International Symposium on Spatial and Temporal Databases (To Appear) (2021)
3. Baehr, A., Peña, J.C., Hu, D.J.: Racial and ethnic disparities in adverse drug events: a systematic review of the literature. Journal of racial and ethnic health disparities **2**(4), 527–536 (2015)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
5. Brown, E.G., Wood, L., Wood, S.: The medical dictionary for regulatory activities (meddra). Drug safety **20**(2), 109–117 (1999)
6. Centers for Disease Control and Prevention: Different COVID-19 Vaccines. (https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html)
7. Chircu, A., Sultanow, E., Saraswat, S.P.: Healthcare rfid in germany: an integrated pharmaceutical supply chain perspective. Journal of Applied Business Research (JABR) **30**(3), 737–752 (2014)
8. Day, W.H., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification **1**(1), 7–24 (1984)
9. Dixon, W.J., Massey Jr, F.J.: Introduction to statistical analysis. (1951)
10. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. (https://coronavirus.jhu.edu/map.html). The Lancet infectious diseases **20**(5), 533–534 (2020)
11. Dror, A.A., Eisenbach, N., Taiber, S., Morozov, N.G., Mizrachi, M., Zigron, A., Srouji, S., Sela, E.: Vaccine hesitancy: the next challenge in the fight against covid-19. European journal of epidemiology **35**(8), 775–779 (2020)
12. Dushoff, J., Plotkin, J.B., Viboud, C., Simonsen, L., Miller, M., Loeb, M., David, J.: Vaccinating to protect a vulnerable subpopulation. PLoS Med **4**(5), e174 (2007)
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. pp. 226–231 (1996)
14. FDA, CDC: Vaccine adverse event reporting system. vaers.hhs.gov (2021)
15. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons (2003)
16. Glaser, R., KIECOLT-GLASER, J.K., Malarkey, W.B., Sheridan, J.F.: The Influence of Psychological Stress on the Immune Response to Vaccines. Annals of the New York Academy of Sciences **840**(1), 649–655 (1998)

17. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl 1), 5228–5235 (2004)
18. Jeetu, G., Anusha, G.: Pharmacovigilance: a worldwide master key for drug safety monitoring. Journal of Young Pharmacists **2**(3), 315–320 (2010)
19. Jordahl, K.: Geopandas: Python tools for geographic data. URL: https://github.com/geopandas/geopandas (2014)
20. Kang, J.H., Kim, C.W., Lee, S.Y.: Nurse-perceived patient adverse events and nursing practice environment. Journal of Preventive Medicine and Public Health **47**(5), 273 (2014)
21. Leyens, L., Reumann, M., Malats, N., Brand, A.: Use of big data for drug development and for public and personal health and care. Genetic epidemiology **41**(1), 51–60 (2017)
22. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern recognition **36**(2), 451–461 (2003)
23. Madsen, R.E., Kauchak, D., Elkan, C.: Modeling word burstiness using the dirichlet distribution. In: Proceedings of the 22nd international conference on Machine learning. pp. 545–552 (2005)
24. Moran, P.A.: Notes on continuous stochastic phenomena. Biometrika **37**(1/2), 17–23 (1950)
25. Okoroh, J.S., Uribe, E.F., Weingart, S.: Racial and ethnic disparities in patient safety. Journal of patient safety **13**(3), 153–161 (2017)
26. Pereira, F.G.F., Ataíde, M.B.C.d., Silva, R.L., Néri, E.D.R., Carvalho, G.C.N., Caetano, J.Á.: Environmental variables and errors in the preparation and administration of medicines. Revista brasileira de enfermagem **71**(3), 1046–1054 (2018)
27. Piccardi, C., Detollenaere, J., Bussche, P.V., Willems, S.: Social disparities in patient safety in primary care: a systematic review. International Journal for Equity in Health **17**(1), 114 (2018)
28. 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/)
29. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: 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. pp. 248–256 (2009)
30. Rey, S.J., Anselin, L.: Pysal: A python library of spatial analytical methods. In: Handbook of applied spatial analysis, pp. 175–193. Springer (2010)
31. Sallam, M.: Covid-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. Vaccines **9**(2), 160 (2021)
32. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS) **42**(3), 1–21 (2017)
33. Shaffer, J.P.: Multiple hypothesis testing. Annual review of psychology **46**(1), 561–584 (1995)
34. Stratton, K., Ford, A., Rusch, E., Clayton, E., to Review Adverse Effects of Vaccines, C., et al.: Adverse effects of vaccines: Evidence and causality (2011)
35. Wang, Y., Gunashekar, D.R., Adam, T.J., Zhang, R.: Mining adverse events of dietary supplements from product labels by topic modeling. Studies in health technology and informatics **245**, 614 (2017)
36. Weisstein, E.W.: Bonferroni correction. https://mathworld. wolfram. com/ (2004)