

Comparing FDA-Approved Clinical Drug Trials to the “Wisdom of the Crowd”: A Case Study Using Semaglutide

Parisa Momeni, Gabriel Laverghetta, Lingyao Li, Jay Ligatti

University of South Florida
4202 E Fowler Ave,
Tampa, FL 33620 USA
{parisamomeni,glaverghetta,lingyaol,ligatti}@usf.edu
Corresponding author: Parisa Momeni

Abstract

Understanding how real-world patients experience and report adverse drug reactions (ADRs) is essential for post-market pharmacovigilance. In this work, we compare ADR distributions for semaglutide-based medications across two complementary data sources: FDA clinical trial reports and patient-reported social media posts. To enable structured comparison, we apply a biomedical natural language processing pipeline to extract, normalize, and align ADR mentions from social media data. We then introduce a multidimensional comparison framework based on origami plots. Additionally, we propose a novel metric, the Geometric Overlap Score (GOS), to quantify overlap between FDA and social media ADR profiles. This framework enables assessment of both overall similarity across ADR distributions and localized differences in specific symptom categories.

Our results reveal both shared patterns and notable divergences. Common gastrointestinal symptoms such as nausea and vomiting appear consistently across sources. However, social media places greater emphasis on abdominal pain and headache, whereas FDA reports more strongly emphasize diarrhea, constipation, and decreased appetite. At the brand level, Wegovy exhibits the highest GOS, while Rybelsus shows the weakest alignment. Coarse-level analysis yields a modest GOS of 0.365, suggesting limited population-wide agreement between sources.

These findings highlight the value of structured entity extraction, geometric metrics, and patient-centered data for surfacing clinically relevant but underreported experiences in pharmacovigilance.

1 Introduction

The concept of the “wisdom of the crowd” holds that a large and diverse group of individuals can produce better results than a single expert [1]. The study of crowd wisdom has a long history. During a 1906 county fair in Plymouth, England, 800 people individually estimated the weight of an ox; Francis Galton noted that the median guess of 1207 pounds was less than 1% off from the correct weight of 1198 pounds [2]. With the surge in the availability of decentralized information during the Information Age, the notion of crowd wisdom has gained wide attention.

Traditional knowledge systems rely on centralized information structures, where authoritative domain experts or institutions generate and validate content—such as United

States Food and Drug Administration (FDA) trials, medical journals, and regulatory databases [3]. In contrast, online resources such as Wikipedia, Reddit, Stack Overflow, and X are crowdsourced in nature, with a much lower barrier to entry for contributors [4, 5]. These two paradigms reflect fundamentally different approaches to information generation, quality control, and validation.

Prior work has explored this tension across domains such as governance, encyclopedia accuracy, and clinical data curation [6, 7]. This centralized–distributed dichotomy appears across domains—for instance, in software security, GitHub serves as a decentralized collaboration platform (anyone can create a GitHub account and post there), while the CVE database reflects a more centralized software vulnerability reporting model [8]. Centralized systems often offer rigor, uniformity, and expert oversight, while distributed systems offer scale, speed, and adaptability. However, direct comparisons between the two remain limited in healthcare contexts, where real-world data from social platforms has only recently been studied alongside regulatory gold standards.

Within the pharmaceutical sciences, the practice of pharmacovigilance involves the detection and prevention of adverse drug reactions (ADRs) [9]. Clinical trials such as randomized control trials are the standard tool for assessing drug safety, but they suffer from limitations such as difficulty identifying rare ADRs and underrepresentation of certain populations [10, 11]. To address these limitations, prior work has developed biological models to predict drug ADRs based on their molecular properties [12, 13].

In addition, the proliferation of social media platforms has sparked interest in the use of social media data for pharmacovigilance, due to its real-time nature and the vast quantity of available data [14, 15]. The growing influence of user-generated health discourse on platforms like X (formerly Twitter) has fundamentally altered the landscape of pharmacovigilance. Despite the increasing reliance on decentralized sources for real-time ADR monitoring, there remains a limited understanding of how these crowd-sourced observations align with structured clinical evidence used by regulatory bodies like the FDA.

Many studies have utilized social media data to identify or estimate the prevalence of ADRs [16–19]. Prior work has often focused on identifying novel ADRs from social media (i.e., ADRs that were not discovered during clinical tri-

als) [20, 21]. In addition, prior work has compared ADRs identified in electronic health records to a compiled social media database [22] and compared ADRs reported by the FDA Adverse Event Reporting System (FAERS) to ADRs reported by social media users [21, 23, 24].

Modern natural language processing (NLP) and large language model (LLM) tools have made this sort of analysis much less time-consuming and allowed a much larger scale of analysis. However, to our knowledge, few studies have compared ADRs reported by social media users to ADRs reported in FDA-approved clinical trials. Although clinical trials suffer from drawbacks in identifying rare ADRs [25, 26], they remain one of the primary tools used to gain FDA approval for new drugs [27–29].

To address these research gaps and explore the “wisdom of the crowd” as it pertains to the field of pharmacovigilance, this study compares ADRs reported from FDA-approved clinical trials to ADRs identified in social media posts. We use semaglutide as the focus of our case study. Semaglutide was originally developed as a medication used to treat diabetes but has gained widespread notoriety in off-label obesity treatment [30]. Discussions surrounding semaglutide are highly prevalent on social media platforms [31], making the drug an appropriate focal point for our comparison of centralized and decentralized ADR reporting. Our pipeline builds on previous work analyzing semaglutide discussions across user subpopulations [32].

We investigate the following research questions:

- **RQ1:** Can adverse drug reactions be reliably extracted, clinically canonicalized, and semantically clustered from informal, user-generated social media text using domain-adapted NLP and biomedical concept normalization techniques?
- **RQ2:** Which methods are appropriate for measuring and visualizing similarity between datasets?
- **RQ3:** How does the set of ADRs identified during FDA-approved clinical drug trials for semaglutide differ from the set of ADRs identified from semaglutide-related social media posts?

Motivation and Scope of Each Research Question

RQ1 addresses the full methodological pipeline required to transform noisy, linguistically diverse social media data into clinically interpretable insights. It is motivated by three core challenges:

1. Extracting symptom mentions from informal text using domain-adapted models like BioBERT,
2. Canonicalizing those mentions into standardized biomedical concepts via UMLS and semantic normalization, and
3. Structuring the resulting data through semantic embeddings and clustering to identify latent symptom groupings (e.g., gastrointestinal vs. psychological effects).

Together, these components enable scalable, interpretable ADR extraction from real-world discourse, bridging informal patient language with formal pharmacovigilance frameworks.

RQ2 is a foundational methodological question that addresses a core challenge in cross-dataset pharmacovigilance: how to meaningfully compare ADR distributions between structured clinical trials and noisy, user-generated social media data. Existing approaches—such as frequency counts and statistical tests like chi-square—are limited in scope. They are scale-sensitive, one-dimensional, and fail to preserve the distributional shape of adverse event profiles. These limitations obscure important structural differences between datasets. In response, we justify the need for a more expressive and intuitive methodology. Specifically, we build on the established Intersection-over-Union (IoU) metric [33] to define the Geometric Overlap Score (GOS)—a directional, interpretable metric that enables multidimensional geometric comparison between ADR distributions from clinical and social media sources. GOS allows similarity to be calculated among an arbitrary number of datasets, where each dataset contains a given number of features. We visualize dataset similarity using the preexisting origami plot framework [34]. Our methodological contribution lies in reformulating IoU so it can be computed directly from the normalized ADR distributions used in our origami plots. This approach quantifies overlap between our dataset in a visually intuitive and scale-invariant manner.

RQ3 is the central applied question of our study, aiming to assess the real-world alignment between ADRs documented in FDA-approved clinical trials and those reported in decentralized social media platforms. While prior research has compared social media ADRs with aggregated databases like FAERS, direct comparisons to FDA clinical trial documents remain rare. This gap is crucial because FDA trials represent the regulatory gold standard for drug safety evaluation. By examining where and how crowd-sourced ADRs converge with or diverge from these clinical benchmarks, RQ3 offers critical insight into the reliability and limitations of crowd-based pharmacovigilance.

Our methodology—combining NLP-based extraction with intuitive, multidimensional visualizations—enables granular comparisons across brands, revealing not just whether the datasets differ, but how they differ and why that matters for public health surveillance.

Contributions

This paper makes the following contributions:

1. **A robust NLP pipeline for ADR extraction and normalization from noisy social media text.** Our pipeline integrates BioBERT NER, UMLS semantic normalization, negation detection, and contextual clustering to map informal user-reported symptom descriptions to standardized clinical ADR categories.
2. **The first systematic comparison of semaglutide ADRs from FDA-approved clinical trials with those extracted from 859,751 real-world X posts.** Unlike prior work that compares social media to FAERS or focuses on discovering novel ADRs, we directly align social media discourse with the ADR profiles used to obtain FDA approval.

3. **A novel multidimensional framework for comparing ADR distributions between FDA clinical trial data and large-scale social media data.** We combine origami plots with geometric similarity metrics to quantify distributional overlap across multiple ADR dimensions, addressing limitations of traditional one-dimensional statistical tests.
4. **A multilevel analysis (coarse- and brand-level) of ADR prevalence across social and clinical sources.** We show how ADR overlap varies substantially between Ozempic, Wegovy, and Rybelsus at different granularity levels, revealing population-level alignments and divergences not captured by single-brand or aggregate analyses.
5. **Empirical evidence showing where X ADR reports align strongly with FDA findings, and where they diverge.** Our geometric similarity metrics highlight consistent agreement on high-prevalence gastrointestinal ADRs and notable divergence in ADRs that are disproportionately emphasized in social media discourse.
6. **A justification for using interpretable, multidimensional visualization and overlap metrics as a complement to classical tests such as chi-square.** We demonstrate that chi-square is limited in interpreting high-dimensional ADR distributions—particularly when sample sizes differ or when expected counts are sparse—while our method provides explainable, visual, and scale-invariant comparisons.
7. **A direct geometric relationship between similarity metrics and normalized ADR vectors in the origami framework.** We establish a direct geometric relationship between similarity metrics and the normalized ADR vectors in our origami framework. Specifically, we derive a closed-form expression for the area of overlap between origami charts using the Geometric Overlap Score (GOS)—a novel metric tailored to measure similarity among datasets with an arbitrary number of features. GOS enables interpretable, scale-invariant comparisons between ADR distributions from social media (X) and FDA clinical data, computed directly from the geometric structure of the plots. To our knowledge, this is the first study to formally compute area-based overlap between two origami charts, addressing a methodological gap with a reproducible and interpretable approach for cross-source pharmacovigilance analysis.

2 Data and Methods

Figure 1 provides an overview of this study. We compile two datasets, one consisting of X posts pertaining to semaglutide and another consisting of FDA-approved semaglutide drug trial reports. We apply a modular pipeline to extract ADR mentions from the X posts. We compare the prevalences of the ADRs reported in both datasets using traditional and geometric similarity tests.

2.1 Datasets

FDA-Approved Clinical Drug Trial Reports We query the Drugs@FDA [35] database to obtain clinical trial re-

ports pertaining to “Ozempic”, “Wegovy”, and “Rybelsus” (the three primary semaglutide brands). For each of the three brands, we obtain its approval label. These labels contain various prescribing information, including ADRs reported in clinical trials. We manually inspect the labels and record the ADRs and their associated prevalences in the clinical trials.

Social Media Dataset Our social media dataset consists of X posts compiled via Brandwatch, a social media analytics platform that collects posts via X’s historical API [36]. We use the query terms “Ozempic”, “Wegovy”, and “Rybelsus”, along with “semaglutide” (to collect posts in which the user did not specify the particular drug brand). Our dataset contains the text of each post, along with user information and metadata such as the posting date. We limit our dataset to original posts, replies, reposts (retweets), and quotes that were created between July 1, 2021 (the date when semaglutide gained FDA approval for weight management [37]) and April 30, 2024. The final X dataset consists of 859,751 posts.

To provide for ethical compliance, all collected data is anonymous and contains no personally identifiable information. In addition, we adhere to the rate limits for the X API and avoid collecting deleted posts. Information about accessing an anonymized version of the X dataset is available in the Data Availability subsection. In addition, all of our code is publicly available [38].

2.2 ADR Extraction from Social Media

To identify ADRs from social media discourse, we develop a modular and interpretable NLP pipeline designed to process noisy, user-generated content. This is particularly important given that social media posts often employ informal, non-standard, and diverse language to describe ADRs [39]. Our approach integrates biomedical named entity recognition (NER), domain-aware normalization, negation detection, and embedding-based semantic generalization. This framework is tailored to address the lexical variability and informality inherent in platforms such as X.

Biomedical Named Entity Recognition

We employ a fine-tuned BioBERT model (Ishan0612/biobert-ner-disease-ncbi) via the Hugging Face Transformers API to identify biomedical entities from social media text. BioBERT, pretrained on PubMed and PMC articles, has demonstrated superior performance in biomedical NER tasks [40]. Posts are pre-processed using standard techniques including lowercasing, lemmatization, and tokenization to ensure consistent input to the model.

Concept Normalization and Synonym Mapping Following NER, we standardize entity mentions using a three-stage normalization process:

1. **Rule-based Synonym Mapping:** Informal or slang phrases (e.g., “queasy”, “light headed”) are normalized to canonical terms (e.g., “nausea”, “dizziness”) via a handcrafted dictionary inspired by MedDRA and UMLS synonymy.

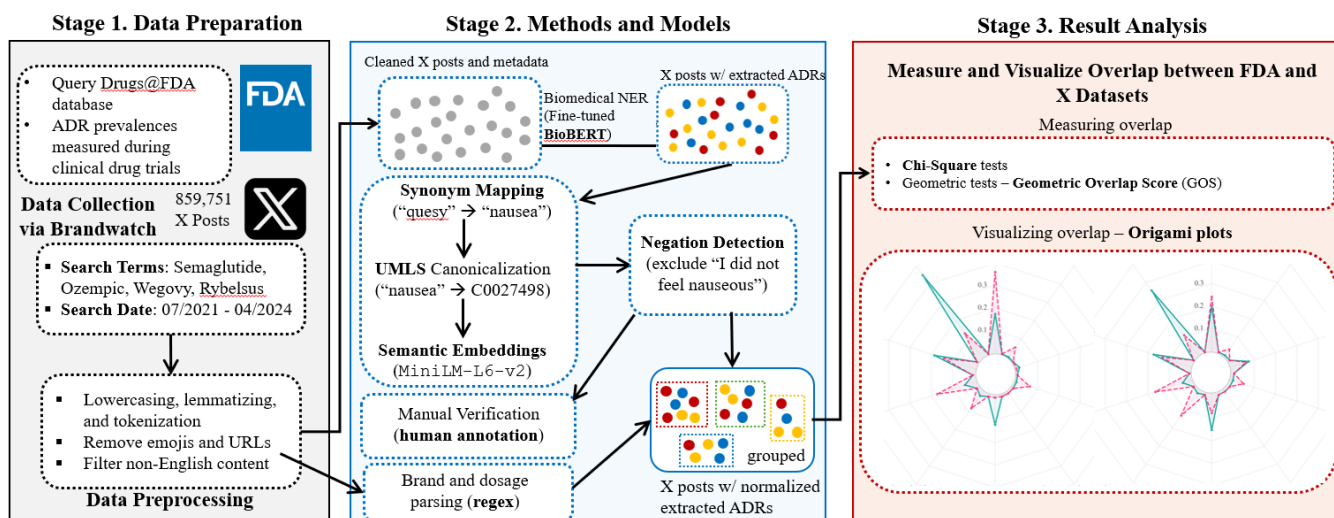


Figure 1: Overview of our research framework.

2. **UMLS Canonicalization:** Using the `scispaCy` library [41] with its built-in UMLS linker [42], we map extracted entities to their corresponding Concept Unique Identifiers (CUIs), filtering for clinically relevant semantic types (e.g., T184: Sign/Symptom, T046: Pathological Function). We utilize the `en_core_sci_sm` model (`scispaCy` version 0.5.1), a lightweight pipeline trained on biomedical literature and optimized for entity recognition and UMLS concept linking.
3. **Semantic Embedding Matching:** Terms not found in UMLS or the synonym map are projected into a latent space using the `all-MiniLM-L6-v2` model [43]. We then compute cosine similarity with pre-encoded ADR vocabularies to identify semantically close matches above a threshold.

This layered normalization ensures robustness across lexical variations, spelling errors, and non-standard phrasing common in user-reported ADRs [44], and directly supports RQ3, which seeks to compare ADRs from FDA clinical trial data with those identified from social media. By reducing linguistic noise and enhancing conceptual alignment, semantic clustering improves the fidelity of this comparison.

Negation Detection and Filtering To distinguish affirmed ADRs from negated or hypothetical mentions, we implement the `NegEx` algorithm [45] using the `negspacy` extension, configured to target UMLS-relevant entity types. This enables reliable exclusion of false positives such as “I didn’t feel nauseous” or “others reported headaches”, ensuring precision in ADR counting and downstream analysis.

Validity Filtering and ADR Classification To isolate legitimate ADRs, we define a conservative filtering schema:

- Mentions of core diseases (e.g., *diabetes*, *hypertension*) are excluded to focus on secondary adverse effects.
- Terms normalized via trusted sources (synonym dictionary or UMLS) are accepted.

- Embedding-based matches are validated against a list of generic non-medical terms (e.g., “life”, “problem”) to avoid overgeneralization.

Final outputs are returned as structured dictionaries containing both positive and negated ADRs per post, along with provenance metadata.

Pipeline Contribution and Novelty Our system introduces several advancements over prior work:

- **Multilevel Normalization:** Combines rule-based, ontology-driven, and embedding-based techniques for high-coverage normalization.
- **Negation Sensitivity:** Correctly filters negated ADRs, addressing a known limitation in large language models [24].
- **Dynamic Vocabulary Expansion:** Learns previously unseen ADRs via UMLS type inference and semantic clustering, without manual supervision.

Manual Verification To assess the quality of our ADR extraction pipeline, we evaluate a manually annotated subset of our dataset. We randomly select posts from our dataset and the extracted ADRs for those posts. For each post, we classify the pipeline’s results into one of four categories:

- True positive (TP): The pipeline correctly flags the post as mentioning an ADR and identifies the correct ADR(s).
- False positive (FP): The pipeline incorrectly flags the post as mentioning an ADR or identifies an incorrect ADR.
- True negative (TN): The pipeline correctly flags the post as not mentioning any ADRs.
- False negative (FN): The pipeline incorrectly flags the post as not mentioning any ADRs.

We perform two manual verification analyses. To measure the pipeline’s ability to correctly classify posts as containing ADRs, we manually annotate 1000 random posts from the dataset. In addition, to assess the quality of the ADR

extraction, we manually annotate 1000 random posts that our pipeline flagged as containing ADRs. These analyses are performed without reposts, as we exclude reposts from the ADR extraction pipeline to avoid skewing the prevalence results with duplicate ADR mentions. For both analyses, we measure precision ($TP/(TP + FP)$) and accuracy ($(TP + TN) / \text{total number of posts}$). Two of the authors independently perform the manual verification; intercoder agreements between the authors are calculated using Krippendorff’s Alpha [46]. Our manual verification results are shown in Table 1.

2.3 Brand & dosage parsing

Since ADR prevalence rates vary among semaglutide brands and dosages and off-label dosage (e.g., over-dosing and under-dosing) is common among users [47, 48], we identify brand names within posts through regular expression (regex) matches. We treat Wegovy, Ozempic, and Rybelsus as the canonical brands and match them with a case-insensitive regex over a curated alias list (common misspellings, abbreviations, and hashtags like `weg`, `ozzy`, `#ozempic`). When multiple brands appear, we designate the first valid match in text order as the primary brand. If no recognizable brand appears, we assign “No Brand Given”.

For each detected primary brand occurrence, we search for dose mentions within ± 60 characters. The dose extraction regex handles cases without spaces (e.g., `1mg`), with multiple or Unicode whitespace characters (e.g., `tabs`, `non-breaking spaces`, `thin spaces`), European decimal notation (e.g., `1,0 mg`), and optional approximation symbols (`~`, `≈`) and excludes units such as `mg/dL` or `mg/L`.

We then aggregate doses by brand (union across all windows and brand-specific subsets near the primary brand). Detected doses are matched to brand-specific canonical sets using a tolerance of ± 0.15 mg (i.e., Ozempic: 0.5 mg, 1 mg; Rybelsus: 7 mg, 14 mg; Wegovy: 2.4 mg). Posts that contain a numeric dose with “mg” but fell outside the FDA-reported ranges for that brand were specifically categorized as Off-label. Mentions that reference a brand but did not include any dosage are categorized as “No Dosage Given”.

For comparability, we report three prevalence measures: All (%) – the share of the entire corpus, Within-brand (%) – the share within that brand, and Brand All (%) – the brand’s total share of the corpus, repeated across its rows. The overall distribution of brand and dosage mentions at the post level is summarized in Table 2.

Per-authors aggregation Since individual authors may create multiple posts, we derive author-level measures to avoid inflating counts due to prolific posters. Specifically, we collapse the per-post annotations into an author-level table, identifying unique authors and summarizing their brand and dosage mentions. To ensure balanced population-level analysis, we track whether each author ever mentioned a specific brand, reported a dosage, or referenced an off-label amount. This aggregation mitigates distortion from repetitive posting behavior and provides a clearer picture of author-level discourse. Together, this annotation and aggregation framework enables structured comparisons between social media data

and clinical trial findings, while supporting flexible stratification by brand and usage patterns. The resulting brand and dose prevalence statistics are presented in Table 3, and the corresponding user-level engagement metrics appear in Table 4.

2.4 Measuring and Visualizing Dataset Overlap

With the ADRs, brand names, and dosages extracted, we seek to quantify and visualize the similarity of these ADRs to those reported in FDA-approved semaglutide trials. We report a chi-square test [49] at the coarse level (all semaglutide-related ADRs) as a baseline comparison, since chi-square is the most commonly used method in ADR similarity analysis.

The chi-square test requires observed and expected frequencies for each ADR category. For X , observed counts O_i represent the actual number of times ADR i appears in the collected posts. However, the FDA reports ADR prevalence as percentages rather than raw counts. To enable statistical comparison, we first normalize each FDA prevalence value to form a probability distribution and then convert into pseudo-counts by multiplying each proportion by the total number of X ADR mentions. Formally, the FDA pseudo-count for ADR i is computed as:

$$O_i^{FDA} = p_i^{FDA} \times N_X, \quad (1)$$

where p_i^{FDA} is the normalized FDA prevalence and N_X is the total number of ADR extractions from X . Expected counts E_i are then computed by the chi-square procedure under the null hypothesis that the two distributions are identical.

Prior work has highlighted that one-dimensional tests such as chi-square cannot fully represent multidimensional shape or distributional overlap. For example, [50] show that Pearson’s chi-square statistic is non-invariant under scaling of count matrices, and [51] discuss the curse of dimensionality in goodness-of-fit testing in arbitrary dimensions. Additionally, the chi-square test assumes independence and a sufficient sample size per category, assumptions that are not always satisfied in real-world social media data.

Graphical tools and statistics are crucial for understanding complex phenomena, including ADRs [52, 53]. To visualize the similarity between our datasets, we plot the ADR prevalences graphically using radar charts, a common tool for visualizing multivariate data [54, 55]. We specifically use origami plots, an improvement over standard radar charts whose area is invariant to axis order [34]. After constructing origami plots for the ADR prevalence distributions in the FDA and X datasets, we introduce and apply geometric analysis using the Geometric Overlap Score (GOS)—a closed-form metric designed to quantify the overall shape similarity and directional overlap among datasets. The origami plot representations and geometric tests complement traditional chi-square statistics by clarifying the cause of the differences between the datasets (i.e., the ADR prevalences that differ).

Constructing Origami Plots We define origami plots as a general geometric representation for comparing multivariate feature distributions across a set of D datasets. To construct

Table 1: Manual verification results

	TP	FP	TN	FN	Precision	Accuracy	Intercoder Agreement
1000 random posts, excluding reposts	3	0	993	4	1	0.996	0.933
1000 random posts flagged as containing ADRs, excluding reposts	854	152	0	0	0.854	0.854	0.852

Table 2: Brand prevalence (sorted by Brand All %) with per-dose details for the X dataset. “All %” indicates the share of the entire corpus, and “Within %” indicates the share within that brand.

Brand	Brand All (%)	Dose details (Count / All % / Within %)			
Ozempic	688,082 (80.02)%	Dose	Count	All %	Within %
		0.5 mg	133	0.02%	0.02%
		1 mg	471	0.05%	0.07%
		Off-label	1,014	0.12%	0.15%
		No Dosage Given	686,464	79.84%	99.76%
Wegovy	56,225 (6.54)%	Dose	Count	All %	Within %
		2.4 mg	462	0.05%	0.82%
		Off-label	159	0.02%	0.28%
		No Dosage Given	55,604	6.47%	98.89%
Rybelsus	3,679 (0.43)%	Dose	Count	All %	Within %
		7 mg	34	0.00%	0.92%
		14 mg	116	0.01%	3.15%
		Off-label	80	0.01%	2.17%
		No Dosage Given	3,449	0.41%	93.75%
No Brand Given	111,886 (13.01)%	Dose	Count	All %	Within %
		No Dosage Given	111,886	13.01%	100.00%

Table 3: Brand and dose-level prevalence, grouped by authors, for the X dataset. “All %” indicates the share of the entire corpus, and “Within %” indicates the share within that brand. Note that the totals for each brand sum to more than 100% because multiple posts may originate from the same author (e.g., an author may mention multiple brands across separate posts).

Brand	Per-brand-authors	Dose details (Count / All % / Within %)			
Ozempic	383,668 (87.89)%	Dose	Count	All %	Within %
		0.5 mg	117	0.03%	0.03%
		1 mg	461	0.11%	0.12%
		Off-label	971	0.22%	0.25%
		No Dosage Given	383,015	87.74%	99.83%
Wegovy	31,320 (7.17)%	Dose	Count	All %	Within %
		2.4 mg	395	0.09%	1.26%
		Off-label	132	0.03%	0.42%
		No Dosage Given	31,060	7.11%	99.17%
Rybelsus	2,755 (0.63)%	Dose	Count	All %	Within %
		7 mg	33	0.01%	1.20%
		14 mg	113	0.01%	4.10%
		Off-label	74	0.02%	2.69%
		No Dosage Given	2,607	0.60%	94.61%
No Brand Given	61,027 (13.98)%	Dose	Count	All %	Within %
		No Dosage Given	61,027	13.98%	100.00%

an origami plot, we fix a common set of N features (variables) that define the axes of the plot. For each dataset indexed by $i \in \{1, \dots, D\}$ and each feature $j \in \{1, \dots, N\}$, we associate a nonnegative feature height $h_j^i \in [0, 1]$, representing the normalized relative occurrence of feature j in dataset i .

Each origami plot is constructed by mapping the feature

heights to polar coordinates. Specifically, each feature j is assigned a uniformly spaced angle

$$\theta_j = \frac{2\pi(j-1)}{N}, \quad (2)$$

and the corresponding point for dataset i is given by

$$(h_j^i \cos \theta_j, h_j^i \sin \theta_j). \quad (3)$$

Table 4: Author-level engagement statistics and brand mentions for X data.

Metric	Value
Total unique authors	436,551
Authors mentioning at least 1 brand	402,126 (92.11%)
Authors mentioning more than 1 brand	15,221 (3.49%)
Authors mentioning exactly 1 brand & 1 FDA-approved dose (no off-label)	728 (0.17%)
Authors mentioning any off-label dose	1,155 (0.26%)
Authors mentioning both brand & (any) dose	2,145 (0.49%)
Authors mentioning brand only (no dose)	399,981 (91.62%)
Authors never mentioning any brand	34,425 (7.89%)
Authors mentioning both FDA-approved brand and no brand in different posts	26,602 (6.09%)
Most-mentioned brand by unique authors	Ozempic (383,668 authors, 87.89%)

This mapping embeds the origami profile for each dataset in the plane.

Figure 2 shows an origami plot for $N = 10$ features constructed for a dataset indexed by i . To compute the area of the origami, each wedge can be decomposed into two congruent isosceles triangles of the form $\triangle Oh_j^i B_j$ and its symmetric partner $\triangle Oh_j^i B_{j-1}$. Since the N axes are equally spaced, the central angle at O between consecutive axes is π/N , and the distance from O to each auxiliary point B_j is fixed at r . The area of each triangle depends on the spike length h_j^i , the constant radius r of the inner ring, and the sine of the central angle. Specifically, the area of each triangle is

$$s_j^i = \frac{1}{2} h_j^i r \sin\left(\frac{\pi}{N}\right). \quad (4)$$

Each axis contributes two such triangles, one on either side of the tip h_j^i . Summing over all axes leads to a compact formula for the total shaded area:

$$S^i = \sum_{j=1}^N 2s_j^i = \sin\left(\frac{\pi}{N}\right) r \sum_{j=1}^N h_j^i. \quad (5)$$

Geometric Overlap Score (GOS) *Geometric Overlap Score* is a generalized metric that quantifies the overlap among a set of D datasets, where each dataset has a set of N features. To motivate this metric, we begin with the widely used *Intersection over Union (IoU)*—also known as the *Jaccard Similarity Coefficient* [33]—which quantifies the degree of overlap between two sets in relation to their combined coverage. The general form of IoU is given by:

$$\text{IoU}(S^1, S^2, \dots, S^D) = \frac{|S^1 \cap S^2 \cap \dots \cap S^D|}{|S^1 \cup S^2 \cup \dots \cup S^D|}, \quad (6)$$

where $|S^1 \cap S^2 \cap \dots \cap S^D|$ represents the area of geometric intersection among the D datasets and $|S^1 \cup S^2 \cup \dots \cup S^D|$ represents their total combined area.

Based on Figure 2, we observe that the base of every triangle in the origami plot is the same. Thus, the height of each of the N triangles is directly proportional to the value of the dataset feature for that triangle, h_j^i . When multiple origami plots are plotted within the same plane using the same center point, they share the same interior area, implying that

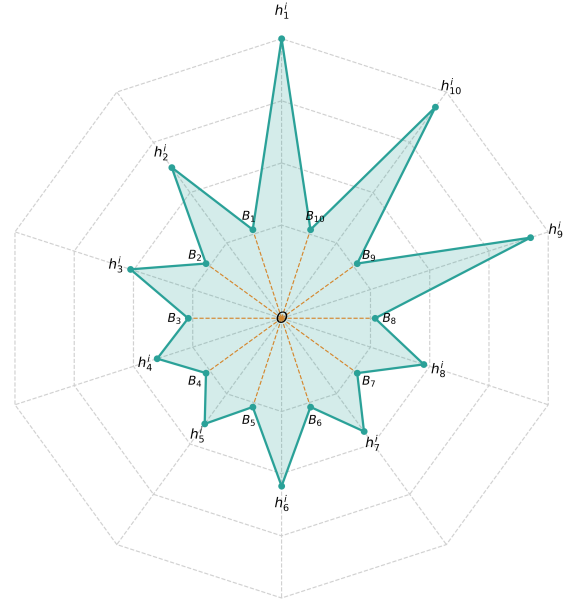


Figure 2: General representation of an origami plot. Each feature height h_j^i forms a spike, with auxiliary points B_j lying on an inner ring of radius r . The complete polygon is constructed by joining consecutive wedges $\triangle Oh_j^i B_j$ and $\triangle Oh_j^i B_{j-1}$. The shaded area is $\sin\left(\frac{\pi}{N}\right) r \sum_{j=1}^N h_j^i$.

the overlap among the origami plots is entirely determined by the triangle heights. We define GOS for origami plots as follows:

$$\text{GOS}(D, N, H) = \frac{\sum_{j=1}^N (h_j^1 \cap h_j^2 \cap \dots \cap h_j^D)}{\sum_{j=1}^N (h_j^1 \cup h_j^2 \cup \dots \cup h_j^D)} \quad (7)$$

In set-theoretic terms, the intersection $h_j^1 \cap h_j^2 \cap \dots \cap h_j^D$ corresponds to the minimum value of feature j across every dataset, capturing the shared area among the datasets at each axis. Similarly, the union $h_j^1 \cup h_j^2 \cup \dots \cup h_j^D$ is taken as the element-wise maximum, reflecting the total coverage across all datasets.

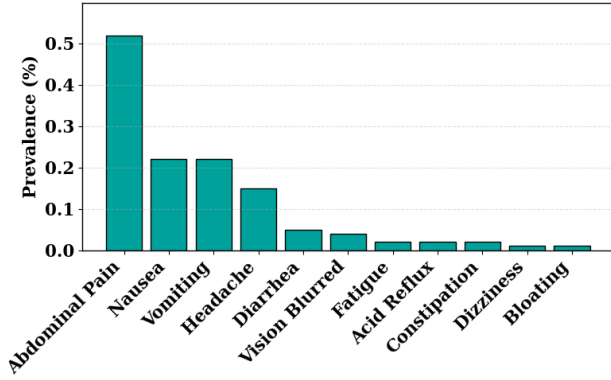


Figure 3: Top ADRs ranked by prevalence in X posts, reflecting the most frequently discussed ADRs in social media discourse.

We define GOS using min/max as follows:

$$\text{GOS}(D, N, H) = \frac{\sum_{j=1}^N \min(h_j^1, h_j^2, \dots, h_j^D)}{\sum_{j=1}^N \max(h_j^1, h_j^2, \dots, h_j^D)}. \quad (8)$$

This formulation provides a compact and interpretable measure of geometric similarity across arbitrary numbers of datasets, where each dataset has a given number of features. The use of min/max to define wedge overlap and union is consistent with prior radar chart area-based similarity metrics [56]. In the above equation, the application of min and max reflects standard set-theoretic definitions of intersection and union.

3 Results and Analysis

3.1 ADR Prevalence Across X and FDA Sources

Our dataset comprises ADR terms extracted from 859,751 posts on X for semaglutide-based drugs (Ozempic, Wegovy, and Rybelsus). As part of the preprocessing steps, we exclude reposts (retweets) to avoid redundant or duplicated ADR mentions that could skew prevalence and frequency distributions. The filtered dataset consists of 411,747 original posts.

Addressing RQ1, we identify a total of 13,056 ADR mentions from the filtered dataset via our ADR extraction pipeline, of which 12,745 (97.62%) are affirmed (non-negated), and 311 (2.38%) are explicitly negated through the implementation of a NegEx-based detection system.

The ADRs are then aggregated to compute mention counts and prevalence percentages. Figure 3 illustrates the distribution of the top most frequently reported ADRs in terms of prevalence across unique posts. The most common symptoms include abdominal pain, nausea, and vomiting, with abdominal pain being reported in over 50% of all posts that contain an ADR.

The ADR prevalence percentages in Table 5 are calculated by averaging reported frequencies from FDA trial data across available doses for each brand. For instance, if an ADR like nausea was reported in 15.8% of patients on Ozempic 0.5 mg and 20.3% on Ozempic 1 mg, its average prevalence is computed as $(15.8 + 20.3)/2 = 18.05\%$. This

approach provides a normalized comparison across brands with varying dosage.

To ensure balanced cross-platform comparison, we extract semaglutide-related ADRs from both X and FDA-approved clinical trial reports. Using the FDA database, we parse prescribing information and FDA-label PDFs for Wegovy, Ozempic, and Rybelsus. For each brand and dosage, its ADRs are normalized and merged. This yields a consolidated FDA ADR frequency vector spanning all semaglutide use cases. The same procedure is applied to social media data using NLP pipelines described in Section 2.2.

To unify comparison, we identify the top ten ADRs with highest combined frequency across platforms. These are used for axes for visualization in origami plots, enabling comparison of distributional structure across platforms. These ten most common ADRs serve as the axes in our origami plots, enabling visual comparison of FDA vs. X distributions for each brand. Unlike Figure 3, which ranks ADRs by prevalence in X alone, Figure 4 orders ADRs by their combined prominence across X and FDA sources to facilitate direct comparison.

3.2 Distributional Comparison of ADR Profiles

ADR patterns can be compared between datasets at multiple levels of granularity. In this study, we compare ADR distributions from X and FDA clinical trial data at two primary levels: the intermediate level, which considers brand-specific differences (e.g., Ozempic, Wegovy, Rybelsus), and the coarse level, which aggregates all semaglutide-related mentions regardless of brand or dosage. Fine-grained comparisons—such as those incorporating specific dosage variations—are feasible using our dataset and could be explored in the future.

Our chi-square test provides statistical grounding to determine whether overall distributions differ between social media and clinical trial data. The test reveals a statistically significant difference ($\chi^2 = 2638.61$, $\text{df} = 9$, $P < 0.001$), indicating a divergence in frequency patterns between the two sources. Although this result demonstrates strong statistical divergence, the chi-square statistic reflects only a global measure of difference and does not preserve ADR-specific distributional behavior, as discussed in Section 2.4. Moreover, as shown in Table 2, the number of ADR mentions differs substantially across brands. This variation violates the assumptions required for valid application of the chi-square test at the intermediate level. Applying the test in this context would conflate brand-specific reporting imbalances with actual distributional differences. Therefore, we restrict the use of chi-square to the coarse level and instead adopt geometric comparison methods for brand-level analysis.

This section focuses on addressing RQ2 and RQ3 (RQ1 is addressed in Section 3.1). To address RQ2, we use origami plots, which visualize ADR prevalence across multiple dimensions in a single, interpretable shape. Unlike classic statistical tests, origami charts preserve distributional structure, enable geometric comparison, and highlight nuanced differences between datasets. We quantify similarity using GOS, measuring the extent of distributional overlap between social media and FDA clinical trial data. For our analysis, we

Table 5: Top 10 FDA-reported ADRs for each semaglutide brand, ranked by average prevalence across standard clinical trial doses.

Rank	Ozempic	Wegovy	Rybelsus
1	Nausea (18.05%)	Nausea (44.0%)	Decreased appetite (75.0%)
2	Diarrhea (8.65%)	Diarrhea (30.0%)	Vomiting (70.0%)
3	Vomiting (7.1%)	Vomiting (24.0%)	Constipation (55.0%)
4	Abdominal Pain (6.5%)	Constipation (24.0%)	Diarrhea (50.0 %)
5	Constipation (4.05%)	Abdominal Pain (20.0%)	Nausea (15.5 %)
6	Dyspepsia (3.1%)	Headache (14.0%)	Abdominal Pain (10.5%)
7	Hypoglycemia (2.7%)	Fatigue (11.0%)	Abdominal Distension (2.5%)
8	Eructation (1.9%)	Dyspepsia (9.0%)	GERD (Reflux Disease) (2.0%)
9	GERD (Reflux Disease) (1.7%)	Dizziness (8.0%)	Gastritis (2.0%)
10	Flatulence (0.95%)	Abdominal Distension (7.0%)	Dyspepsia (1.8%)

apply Equation (8) with $D = 2$ datasets corresponding to X and FDA sources, $N = 10$ ADR features, and H denoting the normalized ADR prevalence vectors. The resulting GOS values are shown in Figure 4. We intentionally unshade the inner polygons within our origami plots (e.g., in Figure 4), de-emphasizing the shared geometric scaffolding across all datasets. This choice foregrounds the data-dependent outer spikes h_j^i —the true source of variation between the X and FDA ADR profiles—and aligns the visualization with the mathematical formulation used in our comparisons.

At the intermediate level of granularity, we analyze ADR patterns by aggregating all posts by brand, regardless of dosage. The comparisons illustrated in Figure 4 a–c provide an intuitive and interpretable visualization of the alignment and divergence between distributed (X) and centralized (FDA) ADR reporting.

Across all three brands we observe that vomiting appears consistently in both datasets, confirming strong agreement on core gastrointestinal effects. However, subtle but meaningful differences also emerge. Notably, abdominal pain appears more prominently in X discourse, whereas other ADRs such as diarrhea and constipation are more emphasized in the FDA dataset. This heightened emphasis may reflect user expectations surrounding weight loss and could be partially shaped by placebo effects or framing biases inherent in social media conversations [57]. The differences may also be influenced by user hesitancy to reveal their experiences with certain ADRs within public spaces. Users may also employ euphemisms rather than medical terminology to describe their ADRs [58].

The origami plots support RQ3 by showing that while the two data sources agree on common side effects, X surfaces additional patterns. These patient-centered effects likely reflect experiences shared on social media that may be underrepresented in formal clinical reports. These patterns justify the need for multidimensional and intuitive comparison tools, such as the geometric metrics and origami plots used in this work.

Figure 4 summarizes the Geometric Overlap Score results for the each semaglutide brand. As shown in Figure 4, Wegovy (GOS = 0.453) exhibits the strongest alignment between X-reported ADRs and FDA-documented ADRs, reflected in the highest GOS among the three brands. In con-

trast, Rybelsus shows noticeably weaker alignment (GOS = 0.208), suggesting that the ADRs most frequently mentioned on X for this medication diverge more substantially from those documented by the FDA.

While brand-level comparisons reveal nuanced ADR profiles, an aggregated perspective offers insight into broader population trends. We now examine the coarsest level of granularity, where all brands and dosages are pooled. The coarse-level aggregation allows us to assess the population-wide similarity between user-reported ADRs on X and FDA clinical trial findings, irrespective of brand or dosage. This analysis provides a population-wide view of how social media discourse aligns with FDA-documented ADRs.

This representation emphasizes not only the shared core of commonly acknowledged ADRs but also the nuanced distortions that reveal perception gaps between social media discourse and clinical documentation. As shown in Figure 4d, both datasets exhibit moderate alignment. Nausea and vomiting appear similarly in both datasets. Abdominal pain and headache are more prominent in X. In contrast, diarrhea, constipation, and decreased appetite are reported less frequently in X than anticipated from FDA documentation. These discrepancies likely reflect the patient-centered focus of social media discourse, particularly regarding weight-loss experiences, appetite suppression, and real-world symptom narratives.

A comparison of Figures 4d (coarse-level origami plot) and 4a (Ozempic) reveals that the aggregated ADR distribution is most visually similar to the Ozempic-specific chart, supporting the notion that Ozempic drives the aggregate shape. This pattern makes sense given that Ozempic dominates the X dataset, accounting for approximately 80% of all brand mentions (Table 2), and therefore exerts the strongest influence on the coarse-level distribution. Wegovy displays similar structural characteristics, though with some variation in dimensions such as headache and decreased appetite. In contrast, Rybelsus diverges more sharply, suggesting a distinct real-world ADR profile that becomes obscured when aggregated.

Quantitatively, the geometric similarity metrics for the coarse-level comparison yield an GOS of 0.365, indicating moderate alignment between X and FDA data. The polygonal similarity between the coarse-level and Ozempic-



Figure 4: Origami plots comparing ADR distributions between social media (X) and FDA data. Subfigures (a–c) show brand-level comparisons for Ozempic, Wegovy, and Rybelsus, respectively. Subfigure (d) displays the coarse-level comparison, aggregating ADRs across all semaglutide brands and dosages. GOS results are shown for all four plots.

specific shapes visually reinforces these metrics. Together, these findings highlight the value of origami charts in preserving ADR-specific structure while enabling interpretable geometric comparisons across levels of granularity. Moreover, the higher alignment observed at the coarse level—largely driven by brand imbalance—underscores the need for intermediate-level analysis to avoid masking meaningful brand-specific variation in real-world discourse.

4 Discussion

Principal Results Our findings demonstrate that X and FDA report overlapping but not identical ADR profiles for

semaglutide-based medications. Using origami plots to encode normalized ADR frequencies, we observe both convergence and divergence in side effect emphasis across sources. For example, FDA reports consistently emphasized clinically severe ADRs such as pancreatitis and kidney injury, while X users disproportionately focused on gastrointestinal effects and weight-related outcomes—particularly for Ozempic and Wegovy. This divergence reflects differences in audience, reporting incentives, and perceived relevance.

The origami visualizations allow us to observe and detect these contrasts at both the aggregate and brand-specific levels. Particularly notable is the case of Rybelsus, where the shape of the X ADR polygon revealed a distinct tilt toward

nausea and appetite suppression—effects that are medically important but may be underreported in clinical channels. By integrating GOS into the origami framework, we provide a unified, closed-form method for quantifying geometric similarity between ADR profiles. This interpretable metric enables meaningful comparisons between clinical trial data and social media discourse across multiple dimensions.

The use of GOS for our case study provides a normalized, scale-invariant measure of similarity that reflects the degree of geometric alignment between real-world (X) and clinical (FDA) ADR distributions. Although we focus on origami plots as our geometric construct, we believe that GOS can convey area of overlap for any reasonable method of graphing the relative occurrences of ADRs in datasets, including, for example, bar charts and other visualizations in which each feature’s area, for a given dataset, is proportionate to its relative frequency (i.e., scalar height).

Practical Implications This work introduces a scalable framework for monitoring ADR signals across formal (FDA) and informal (social media) pharmacovigilance sources. The origami chart framework is compact, interpretable, and dimensionally normalized, enabling meaningful comparisons even when source volumes and ADR distributions differ. As an analytics tool, it supports both visual inspection and algorithmic computation of similarity, offering value for early signal detection, brand surveillance, and post-marketing safety evaluation.

An important strength of our pipeline is its explicit handling of negation during ADR extraction, which improves accuracy over prior approaches that rely solely on LLMs such as GPT. Earlier studies have reported false positives when users mention drugs and side effects in a negated context (e.g., “I didn’t feel nauseous”), which LLMs may incorrectly interpret as positive ADR mentions [59, 60]. By incorporating negation detection, our system minimizes such errors, ensuring cleaner ADR frequency profiles from social media sources. This improvement enhances the reliability of our similarity metrics and strengthens the overall integrity of the origami-based comparison framework.

Our work also fills an important gap in the existing literature on origami charts. While prior studies have used origami-style visualizations, they did not define a method for quantifying similarity—such as area overlap—between charts. As a result, comparisons remained descriptive and lacked formal metrics. In contrast, our approach introduces a closed-form geometric relationship between normalized ADR vectors and similarity metrics—the Geometric Overlap Score (GOS), enabling reproducible, interpretable, and quantitative comparisons between sources. To our knowledge, this is the first use of origami chart overlap as a measurable similarity construct in pharmacovigilance research.

While our empirical evaluation focuses on $N = 10$ ADR features and $D = 2$ datasets (X and FDA), GOS is not limited to this configuration. The formulation naturally generalizes to any number of features and datasets. This generality highlights the method’s potential as a broadly applicable geometric similarity measure.

Limitations While this study offers valuable insights into ADR variations between social media and clinical trial reports, some limitations must be acknowledged. First our dataset is limited to English-language posts on X, excluding non-English speaking users and introducing a potential source of bias.

Second, the demographics of X users may not align with the demographics of patients recruited for clinical drug trials and the general population.

Third, as shown in Table 2, users rarely report their drug dosage, and off-label dosages are common [47, 48], which may decrease the reliability of user-reported dosages.

Fourth, it is important to note that FDA reports are submitted with the explicit intent of documenting adverse drug reactions, while social media posts are informal and not necessarily focused on ADRs. This difference introduces asymmetries in data quality and structure. Still, our framework leverages real-world data at scale to complement traditional pharmacovigilance. By enabling interpretable, quantitative comparisons between sources, it supports near real-time monitoring of emerging ADR signals and aligns with current regulatory interest in real-world evidence for post-marketing surveillance.

Future Work Although this study focuses on semaglutide-related discourse on X, our approach is broadly applicable. The proposed pipeline is drug-agnostic and can be adapted to extract, normalize, and ADRs for a wide range of medications, including those related to both mental and physical health. Furthermore, while we use X as a primary data source, the same methodology can be extended to other social media platforms such as Reddit, Facebook health communities, or patient forums, which may offer richer context and longer-form narratives for ADR analysis.

5 Conclusions

This study bridges the gap between centralized pharmacovigilance mechanisms (FDA clinical trials) and distributed, real-world health discourse on platforms like X. Through a novel combination of domain-adapted NLP techniques and geometric visualization tools, we demonstrate that social media-derived ADR profiles can meaningfully align with FDA data—especially for common, high-prevalence side effects such as nausea, vomiting, and diarrhea. However, discrepancies in ADR prevalence—particularly in appetite-related symptoms—underscore the value of distributed systems in surfacing patient-centered narratives and side effects that may be underemphasized in regulatory reporting. The origami plot framework, combined with the Geometric Overlap Score metric, offers a scale-invariant, interpretable method for visualizing ADR overlap and could serve as a robust supplement to existing pharmacovigilance protocols.

As pharmacovigilance increasingly embraces real-world data sources, our findings advocate for hybrid models that combine the rigor of clinical trials with the agility and patient-centricity of crowd-based platforms. Future extensions of this work could explore cross-platform compar-

isons, longitudinal ADR evolution, and the integration of causal inference models to separate signal from noise in social discourse. Ultimately, this approach may reduce the need for exhaustive pre-market trials, enable faster detection of emerging side effects, and promote a more inclusive, participatory model of drug safety monitoring.

While the case study focuses on semaglutide, the development of the GOS metric forms our core contribution, offering a generalizable tool for structured visual comparison across any domain involving high-dimensional distributional data. It can be extended to other drugs, platforms, languages, or even beyond pharmacovigilance—for example, in comparing user-reported experiences with clinical outcomes in mental health, vaccine reactions, or chronic disease monitoring.

5.1 Data Availability

An anonymized version of the X dataset analyzed during this study, which has the “Author”, “Avatar”, “Entity Info”, “Url”, “Original Url”, “Full Name”, “Twitter Retweet of”, “Twitter Reply to”, “Thread Author”, and “Weblog Title” columns removed, may be downloaded using the following link: https://www.dropbox.com/scl/fi/gksibelrijwnr1wg6xja2/dataset_anonymized.csv?rlkey=ks6dj60ysan9lwh8wdqnygma4&st=294q0sib&dl=0.

References

- [1] James Surowiecki. *The Wisdom of Crowds*. Vintage, 2005. ISBN 978-0-385-50386-0.
- [2] Francis Galton. Vox populi. *Nature*, 75(1949):450–451, Mar 1907. ISSN 1476-4687. doi: 10.1038/075450a0. URL <https://doi.org/10.1038/075450a0>.
- [3] William Badke. *Expertise and Authority in an Age of Crowdsourcing*, pages 191–215. Association of College and Research Libraries, 01 2015. ISBN 978-0838987162.
- [4] Jeff Howe. The rise of crowdsourcing, 2006. URL <https://www.wired.com/2006/06/crowds>. Accessed: 2025-07-01.
- [5] Lingyao Li, Songhua Hu, Yinpei Dai, Min Deng, Parisa Momeni, Gabriel Laverghetta, Lizhou Fan, Zihui Ma, Xi Wang, Siyuan Ma, Jay Ligatti, and Libby Hemphill. Toward satisfactory public accessibility: A crowdsourcing approach through online reviews to inclusive urban design. *Computers, Environment and Urban Systems*, 122:102329, 2025. ISSN 0198-9715. doi: <https://doi.org/10.1016/j.compenvurbsys.2025.102329>. URL <https://www.sciencedirect.com/science/article/pii/S0198971525000821>.
- [6] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, Dec 2005. ISSN 1476-4687. doi: 10.1038/438900a. URL <https://doi.org/10.1038/438900a>.
- [7] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19, 2007.
- [8] Parisa Momeni, Gabriel Laverghetta, Kevin Dennis, Bianca Dehann, and Jay Ligatti. Measuring the absolute and relative prevalence of sql concatenations and sql-identifier injections. In Pierangela Samarati and Sabrina De Capitani di Vimercati, editors, *Security and Cryptography*, pages 243–271, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-09598-5.
- [9] European Medicines Agency. Pharmacovigilance: Overview. <https://www.ema.europa.eu/en/human-regulatory-overview/pharmacovigilance-overview>, 2024. Accessed: 2025-07-01.
- [10] L. Härmärk and A. C. van Grootheest. Pharmacovigilance: methods, recent developments and future perspectives. *European Journal of Clinical Pharmacology*, 64(8):743–752, Aug 2008. ISSN 1432-1041. doi: 10.1007/s00228-008-0475-9. URL <https://doi.org/10.1007/s00228-008-0475-9>.
- [11] Asefeh Heiat, Cary P. Gross, and Harlan M. Krumholz. Representation of the elderly, women, and minorities in heart failure clinical trials. *Archives of Internal Medicine*, 162(15), 08 2002. ISSN 0003-9926. doi: 10.1001/archinte.162.15.1682. URL <https://doi.org/10.1001/archinte.162.15.1682>.
- [12] Liang-Chin Huang, Xiaogang Wu, and Jake Y. Chen. Predicting adverse side effects of drugs. *BMC Genomics*, 12(5):S11, Dec 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-S5-S11. URL <https://doi.org/10.1186/1471-2164-12-S5-S11>.
- [13] Yoshihiro Yamanishi, Edouard Pauwels, and Masaaki Kotera. Drug side-effect prediction based on the integration of chemical and biological spaces. *Journal of Chemical Information and Modeling*, 52(12):3284–3292, Dec 2012. ISSN 1549-9596. doi: 10.1021/ci2005548. URL <https://doi.org/10.1021/ci2005548>.
- [14] Abeer Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O’Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212, 2015. ISSN 1532-0464. doi: 10.1016/j.jbi.2015.02.004. URL <https://www.sciencedirect.com/science/article/pii/S1532046415000362>.
- [15] Dimitra Pappa and Lampros K. Stergioulas. Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions. *International Journal of Data Science and Analytics*, 8(2):113–135, Sep 2019. ISSN 2364-4168. doi: 10.1007/s41060-019-00175-3. URL <https://doi.org/10.1007/s41060-019-00175-3>.
- [16] Jérémy Lardon, Redhouane Abdellaoui, Florelle Bellet, Hadyl Asfari, Julien Souvignet, Nathalie Texier, Marie-Christine Jaulent, Marie-Noëlle Beyens, Anita Burgun, and Cédric Bousquet. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17(7):e171, Jul 2015. ISSN 1438-8871. doi: 10.2196/jmir.4304. URL <http://www.jmir.org/2015/7/e171/>.

- [17] Thin Nguyen, Mark E. Larsen, Bridianne O’Dea, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Estimation of the prevalence of adverse drug reactions from social media. *International Journal of Medical Informatics*, 102:130–137, 2017. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2017.03.013. URL <https://www.sciencedirect.com/science/article/pii/S1386505617300746>.
- [18] Alon Bartal, Kathleen Jagodnik, Nava Pliskin, and Abraham Avi Seidmann. Utilizing ai and social media analytics to discover adverse side effects of glp-1 receptor agonists. Research Paper 4790676, Boston University Questrom School of Business, April 10 2024. URL <https://ssrn.com/abstract=4790676>. Available at SSRN.
- [19] Seraphina Fong, Alessandro Carollo, Lambros Lazuras, Ornella Corazza, and Gianluca Esposito. Ozempic (glucagon-like peptide 1 receptor agonist) in social media posts: Unveiling user perspectives through reddit topic modeling. *Emerging Trends in Drugs, Addictions, and Health*, 4:100157, 2024. ISSN 2667-1182. doi: 10.1016/j.etchd.2024.100157. URL <https://www.sciencedirect.com/science/article/pii/S2667118224000163>.
- [20] Michael P. Bremmer and Christian S. Hendershot. Social media as pharmacovigilance: The potential for patient reports to inform clinical research on glucagon-like peptide 1 (glp-1) receptor agonists for substance use disorders. *Journal of Studies on Alcohol and Drugs*, 85(1):5–10, 2024. doi: 10.15288/jsad.23-00318. URL <https://doi.org/10.15288/jsad.23-00318>. PMID: 37917019.
- [21] Jing Zhang, Xiaofen Wang, and Yiting Zhou. Comparative analysis of semaglutide induced adverse reactions: Insights from faers database and social media reviews with a focus on oral vs subcutaneous administration. *Frontiers in Pharmacology*, 15, 2024. ISSN 1663-9812. doi: 10.3389/fphar.2024.1471615. URL <https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1471615>.
- [22] Maxim Topaz, Kenneth Lai, Neil Dhopeswarkar, Diane L. Seger, Roe Sa’adon, Foster Goss, Ronen Rozenblum, and Li Zhou. Clinicians’ reports in electronic health records versus patients’ concerns in social media: A pilot study of adverse drug reactions of aspirin and atorvastatin. *Drug Safety*, 39(3):241–250, Mar 2016. ISSN 1179-1942. doi: 10.1007/s40264-015-0381-x. URL <https://doi.org/10.1007/s40264-015-0381-x>.
- [23] Ying Li, Antonio Jimeno Yepes, and Cao Xiao. Combining social media and fda adverse event reporting system to detect adverse drug reactions. *Drug Safety*, 43(9):893–903, Sep 2020. ISSN 1179-1942. doi: 10.1007/s40264-020-00943-2. URL <https://doi.org/10.1007/s40264-020-00943-2>.
- [24] Zhijie Duan, Kai Wei, Zhaoqian Xue, Jiayan Zhou, Shu Yang, Siyuan Ma, Jin Jin, and Lingyao li. Crowdsourcing-based knowledge graph construction for drug side effects using large language models with an application on semaglutide, 2025. URL <https://arxiv.org/abs/2504.04346>.
- [25] C A Naranjo, U Busto, and E M Sellers. Difficulties in assessing adverse drug reactions in clinical trials. *Prog Neuropsychopharmacol Biol Psychiatry*, 6(4-6): 651–657, 1982. doi: 10.1016/s0278-5846(82)80162-0.
- [26] Sonal Singh and Yoon K Loke. Drug safety assessment in clinical trials: methodological challenges and opportunities. *Trials*, 13:138, 2012. doi: doi.org/10.1186/1745-6215-13-138.
- [27] Peter Corr and David Williams. *Conflict of Interest in Medical Research, Education, and Practice*, chapter E, The Pathway from Idea to Regulatory Approval: Examples for Drug Development. Institute of Medicine (US) Committee on Conflict of Interest in Medical Research, Education, and Practice, 2009. URL <https://www.ncbi.nlm.nih.gov/books/NBK22930/>.
- [28] Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*. Springer, 2015. doi: 10.1007/978-3-319-18539-2.
- [29] U.S Food & Drug Administration. Overview of our role regulating and approving drugs — video series. <https://www.fda.gov/drugs/information-consumers-and-patients-drugs/overview-our-role-regulating-and-approving-drugs-video-series>, 2024. Accessed: 2025-07-02.
- [30] Aurélie Mailhac, Lars Pedersen, Anton Pottegård, Jens Søndergaard, Torben Mogensen, Henrik Toft Sørensen, and Reimar W Thomsen. Semaglutide (ozempic®) use in denmark 2018 through 2023 – user trends and off-label prescribing for weight loss. *Clinical Epidemiology*, 16:307–318, 2024. doi: 10.2147/CLEP.S456170. URL <https://www.tandfonline.com/doi/abs/10.2147/CLEP.S456170>.
- [31] Miguel Angel Alvarez-Mon, Maria Llaveró-Valero, Angel Asunsolo del Barco, Cristina Zaragozá, Miguel A Ortega, Guillermo Lahera, Javier Quintero, and Melchor Alvarez-Mon. Areas of interest and attitudes toward antiobesity drugs: Thematic and quantitative analysis using twitter. *Journal of Medical Internet Research*, 23(10):e24336, Oct 2021. ISSN 1438-8871. doi: 10.2196/24336. URL <https://www.jmir.org/2021/10/e24336>.
- [32] Parisa Momeni, Gabriel Laverghetta, Jay Ligatti, and Lingyao Li. Topic and sentiment trends in semaglutide discussions on x: Subpopulation-based longitudinal analysis. *Online Journal of Public Health Informatics (forthcoming/in press)*, 2026. doi: 10.2196/80660.
- [33] Vijay Verma and Rajesh Kumar Aggarwal. A comparative analysis of similarity measures akin to the jaccard index in collaborative recommendations: empirical and theoretical perspective. *Social Network Analysis and Mining*, 10(1):43, Jun 2020. ISSN 1869-

5469. doi: 10.1007/s13278-020-00660-9. URL <https://doi.org/10.1007/s13278-020-00660-9>.
- [34] Rui Duan, Jiayi Tong, Alex J Sutton, David A Asch, Haitao Chu, Christopher H Schmid, and Yong Chen. Origami plot: a novel multivariate data visualization tool that improves radar chart. *J Clin Epidemiol*, 156: 85–94, February 2023.
- [35] U.S. Food & Drug Administration. Drugsfda: Fda-approved drugs. <https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm>, 2025. Accessed: 2025-07-05.
- [36] Brandwatch. The most intelligent social suite. <https://www.brandwatch.com/>, 2025. Accessed: 2025-11-10.
- [37] Anastassia Amaro, Danny Sugimoto, and Sean Wharton. Efficacy and safety of semaglutide for weight management: evidence from the step program. *Postgraduate Medicine*, 134(sup1):5–17, 2022. doi: 10.1080/00325481.2022.2147326.
- [38] Parisa Momeni. ADR-FDA-X. <https://github.com/ParisaMomeni/ADR-FDA-X>, 2026. Accessed: 2026-03-31.
- [39] Azadeh Nikfarjam, Abeed Sarker, Karen O’Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 03 2015. ISSN 1067-5027. doi: 10.1093/jamia/ocu041. URL <https://doi.org/10.1093/jamia/ocu041>.
- [40] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [41] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispace: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.
- [42] Liz Amos, David Anderson, Stacy Brody, Anna Ripple, and Betsy L Humphreys. Umls users and uses: a current overview. *Journal of the American Medical Informatics Association*, 27(10):1606–1611, 07 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa084. URL <https://doi.org/10.1093/jamia/ocaa084>.
- [43] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [44] Azadeh Nikfarjam, Abeed Sarker, Karen O’connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.
- [45] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.
- [46] Giacomo Marzi, Marco Balzano, and Davide Marchiori. K-alpha calculator–krippendorff’s alpha calculator: A user-friendly tool for computing krippendorff’s alpha inter-rater reliability coefficient. *MethodsX*, 12:102545, 2024. ISSN 2215-0161. doi: <https://doi.org/10.1016/j.mex.2023.102545>. URL <https://www.sciencedirect.com/science/article/pii/S2215016123005411>.
- [47] Hildtraud Knopf, Ingrid-Katharina Wolf, Giselle Sarganas, Wanli Zhuang, Wolfgang Rascher, and Antje Neubert. Off-label medicine use in children and adolescents: results of a population-based study in germany. *BMC Public Health*, 13(1):631, Jul 2013. ISSN 1471-2458. doi: 10.1186/1471-2458-13-631. URL <https://doi.org/10.1186/1471-2458-13-631>.
- [48] Sirak Tesfamariam, Indermeet Singh Anand, Ghide Kaleab, Samson Berhane, Biruck Woldai, Eyasu Habte, and Mulugeta Russom. Self-medication with over the counter drugs, prevalence of risky practice and its associated factors in pharmacy outlets of asmara, eritrea. *BMC Public Health*, 19(1):159, Feb 2019. ISSN 1471-2458. doi: 10.1186/s12889-019-6470-5. URL <https://doi.org/10.1186/s12889-019-6470-5>.
- [49] Yao-Tsung Chen and Meng Chang Chen. Using chi-square statistics to measure similarities for text categorization. *Expert Systems with Applications*, 38(4):3085–3090, 2011. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2010.08.100>. URL <https://www.sciencedirect.com/science/article/pii/S0957417410008961>.
- [50] Vladimir Gurvich and Mariya Naumova. Critical issues with the pearson’s chi-square test. *arXiv preprint arXiv:2505.06318*, 2025.
- [51] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471, 2018.
- [52] Si Li, Chia-Hui Yu, Yichuan Wang, and Yedurag Babu. Exploring adverse drug reactions of diabetes medicine using social media analytics and interactive visualizations. *International Journal of Information Management*, 48:228–237, 2019. ISSN 0268-4012. doi: <https://doi.org/10.1016/j.ijinfomgt.2018.12.007>. URL <https://www.sciencedirect.com/science/article/pii/S0268401217303456>.
- [53] Krzysztof Górniewicz, Zbigniew Palka, and Waldemar Ratajczak. Measuring the similarity of charts in graphical statistics. *Scientific Reports*, 14(1):6893, Mar 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56156-5. URL <https://doi.org/10.1038/s41598-024-56156-5>.
- [54] John M. Chambers. *Graphical Methods for Data Analysis*. Wadsworth International Group, Bel-

mont, Calif, 1983. URL <https://archive.org/details/graphicalmethods00cham.0>.

- [55] Sangjin Jeong, Yu Mi Jo, Sang-Oh Shim, Yeon-Jung Choi, and Chan-Hyun Youn. A novel model for metabolic syndrome risk quantification based on areal similarity degree. *IEEE Transactions on Biomedical Engineering*, 61(3):665–679, 2013.
- [56] Michael M Porter and Pooya Niksiar. Multidimensional mechanics: Performance mapping of natural biological systems using permutated radar charts. *PloS one*, 13(9):e0204309, 2018.
- [57] Verena Hoffmann, Marina Lanz, Jennifer Mackert, Timo Müller, Matthias Tschöp, and Karin Meissner. Effects of placebo interventions on subjective and objective markers of appetite—a randomized controlled trial. *Frontiers in Psychiatry*, Volume 9 - 2018, 2018. ISSN 1664-0640. doi: 10.3389/fpsyt.2018.00706. URL <https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2018.00706>.
- [58] Stéphane Schück, Avesta Roustamal, Anaïs Gedik, Paméla Voillot, Pierre Foulquié, Catherine Penfornis, and Bernard Job. Assessing patient perceptions and experiences of paracetamol in france: Infodemiology study using social media data mining. *J Med Internet Res*, 23(7):e25049, 2021.
- [59] Michael E. Matheny, Jie Yang, Joshua C. Smith, Colin G. Walsh, Mohammed A. Al-Garadi, Sharon E. Davis, Keith A. Marsolo, Daniel Fabbri, Ruth R. Reeves, Kevin B. Johnson, Gerald J. Dal Pan, Robert Ball, and Rishi J. Desai. Enhancing postmarketing surveillance of medical products with large language models. *JAMA Network Open*, 7(8):e2428276–e2428276, 08 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.28276. URL <https://doi.org/10.1001/jamanetworkopen.2024.28276>.
- [60] Md Muntasir Zitu, Dwight Owen, Ashish Manne, Ping Wei, and Lang Li. Large language models for adverse drug events: A clinical perspective. *Journal of Clinical Medicine*, 14(15), 2025. ISSN 2077-0383. doi: 10.3390/jcm14155490. URL <https://www.mdpi.com/2077-0383/14/15/5490>.