

RESEARCH ARTICLE

Process Mining for Characterizing Adverse Event Transition Patterns in GLP-1 Receptor Agonist Using OpenFDA Data

Adil Priman Hati Hulu ¹, Winda Irawati Zebua ², Fathree Halawa ³ and Jernih Hati Sarumaha ⁴

¹Informatics Engineering, STMIK Methodist Binjai, 20716, Indonesia ²Medical Laboratory Technology, Pharmacy and Health Sciences, Sari Mutiara University of Indonesia, Medan 20123, Indonesia ³Information Systems, Science, Technology and Information, Sari Mutiara University of Indonesia, Medan 20123, Indonesia ⁴Business and Management, Management, Mitra Bangsa University, Pasar Minggu, 12520, Indonesia

*Correspondence adilprimanhatihulu@gmail.com

Received: 12 September 2026; Revised: 20 September 2026; Accepted: 28 September 2026, Published: 30 September 2026

ABSTRACT

The increasing use of glucagon-like peptide-1 receptor agonists (GLP-1 RAs) for type 2 diabetes mellitus and obesity has increased the need for structured post-marketing safety analysis. Pharmacovigilance studies based on spontaneous reporting systems generally emphasize adverse event frequency and disproportionality, while sequential relationships among multiple events within individual reports remain insufficiently represented. This study addresses this gap by applying process mining to adverse event reports obtained from the U.S. Food and Drug Administration Adverse Event Reporting System through the OpenFDA Drug Adverse Event Application Programming Interface for semaglutide, liraglutide, dulaglutide, and tirzepatide during 2023–2025. A total of 188,734 reports were processed through cleaning, normalization, deduplication, event-log construction, Top-25 adverse event selection, Directly-Follows Analysis, Heuristics Miner, and Inductive Miner. The final event log contained 496,212 event records from 188,247 cases, including 48,429 cases with at least two reported events. Nausea → Vomiting was the most frequent transition, occurring 4,459 times, while Incorrect dose administered → Injection site pain occurred 2,267 times for tirzepatide. Heuristics Miner achieved precision values of 0.9671–0.9909 and F-scores of 0.8474–0.8682, whereas Inductive Miner achieved fitness of 1.0000 with lower precision values of 0.5315–0.7481. These results indicate that process mining can complement conventional pharmacovigilance analysis by characterizing recurring reporting-transition structures without implying clinical onset, causality, or incidence.

KEYWORDS

Process Mining; Adverse Event; GLP-1 Receptor Agonists; Pharmacovigilance; FAERS; OpenFDA

1. Introduction

Pharmacovigilance supports the continuous monitoring of medicine safety after medicinal products enter routine clinical use. Post-marketing surveillance is necessary because pre-approval clinical trials cannot represent all patient characteristics, concomitant treatments, patterns of medicine use, and long-term exposure conditions that occur in broader populations. Spontaneous reporting systems therefore provide an additional source of real-world safety information by collecting adverse events and medication errors reported after product authorization. The U.S. Food and Drug Administration Adverse Event Reporting System (FAERS) is one of the main sources of post-marketing pharmacovigilance data. FAERS contains adverse event and medication error reports submitted by manufacturers, healthcare professionals, patients, and other reporters. OpenFDA provides computational access to FAERS records through an application programming interface and includes Harmonized Fields for structured drug identification. These data support large-scale safety-signal exploration and computational analysis of reported adverse events.

FAERS data require careful interpretation. A report does not establish that a medicinal product caused an adverse event because underlying diseases, concomitant medications, and other clinical factors can influence the reported outcome. FAERS also contains duplicate reports, incomplete information, differences in reporting quality, and reporting bias. In addition, spontaneous reporting data cannot be used directly to estimate adverse event incidence because the exposed population and the number of unreported events are unknown.

Glucagon-like peptide-1 receptor agonists (GLP-1 RAs) have become an important focus of post-marketing safety surveillance because their use has expanded in the management of type 2 diabetes mellitus and obesity. Recent pharmacovigilance studies have examined adverse event patterns associated with semaglutide, liraglutide, dulaglutide, tirzepatide, and other agents in this therapeutic group. A study of semaglutide identified differences in adverse event patterns according to the route of administration [1]. A more recent FAERS analysis evaluated neurological adverse events associated with several GLP-1 RAs and identified reported safety signals through disproportionality analysis [2].

Tirzepatide has also received increasing attention in recent post-marketing studies. One study analyzed tirzepatide reports in FAERS and identified adverse event signals through disproportionality analysis [3]. Another study examined tirzepatide safety in patients with type 2 diabetes and identified several reported adverse event patterns using multiple disproportionality algorithms [4]. These studies provide updated information on reported safety signals, but their analytical focus remains on frequency, disproportionality, signal strength, and related pharmacovigilance measures.

Most FAERS-based studies of GLP-1 RAs analyze adverse events as individual observations or drug–event associations. Common approaches include Reporting Odds Ratio, Proportional Reporting Ratio, Empirical Bayesian Geometric Mean, Bayesian Confidence Propagation Neural Network, demographic stratification, system-organ classification, and time-to-onset analysis. These methods are useful for identifying potential safety signals. However, they do not explicitly represent the sequential structure of multiple adverse events recorded within the same safety report.

This limitation creates a specific analytical gap. A single FAERS report can contain several reported adverse events. Frequency analysis can identify which events appear most often, while disproportionality analysis can identify which drug–event combinations are reported more frequently than expected. However, these methods do not directly describe which reported event is followed by another event within the same report. As a result, the transition structure among reported adverse events remains insufficiently represented in current GLP-1 RA pharmacovigilance analysis.

Process mining provides a computational approach for analyzing this type of sequential information. Process mining transforms event data into event logs that contain a case identifier, an activity, and temporal information. The resulting traces can be used to identify recurring transitions and process structures. Process mining generally includes process discovery, conformance checking, and process enhancement. This study focuses on process discovery because the objective is to reconstruct process structures from observed event-log behavior.

The use of process mining in healthcare has increased with the availability of digital clinical and administrative data. A systematic literature review published in 2025 identified applications of process mining in clinical processes, organizational workflows, resource management, and healthcare process evaluation [5]. A tertiary study published in 2025 also reported that process discovery remains one of the most frequently used process mining activities in healthcare [6]. These studies also identified data quality, process variability, noise, missing information, and model interpretation as important challenges in healthcare process mining.

Process discovery algorithms differ in how they represent event-log behavior. Heuristics Miner uses

event frequency and dependency relationships to identify recurring transitions. The algorithm is suitable for event logs that contain infrequent behavior and noise. Inductive Miner constructs structured process models recursively from observed event-log behavior. These differences affect the characteristics of the resulting process models.

Process model evaluation therefore requires several quality dimensions. Fitness measures how well a model can reproduce the behavior recorded in the event log. Precision measures how strongly the model restricts behavior that is not present in the event log. Generalization evaluates whether the model can represent related behavior beyond the observed traces. Simplicity describes the structural complexity of the resulting process model. The combined use of these metrics provides a broader basis for comparing process discovery algorithms.

Despite recent developments in pharmacovigilance analysis and healthcare process mining, their integration for GLP-1 RA adverse event reporting remains limited. Recent GLP-1 RA studies continue to emphasize disproportionality, signal detection, and time-to-onset analysis [2–4]. Recent healthcare process mining studies mainly focus on care pathways, clinical workflows, organizational processes, and healthcare operations [5, 6]. Therefore, a methodological gap remains between drug-safety signal analysis and process-oriented analysis of reported adverse event sequences.

This study addresses this gap by applying process mining to adverse event reports associated with semaglutide, liraglutide, dulaglutide, and tirzepatide. The study uses OpenFDA Harmonized Fields from FAERS for the 2023–2025 period. Each safety report is represented as a case. Each reported adverse event is represented as an activity. The available temporal information is used to establish the event order. The resulting event log is analyzed using Directly-Follows Analysis, Heuristics Miner, and Inductive Miner. The main contributions of this study are threefold. First, the study transforms OpenFDA/FAERS reports into an event log that represents safety reports as cases and reported adverse events as activities. This representation supports the analysis of reporting sequences beyond individual event frequencies and disproportionality measures. Second, the study applies Directly-Follows Analysis to identify recurring transitions among the most frequently reported adverse events and to characterize reporting patterns associated with semaglutide, liraglutide, dulaglutide, and tirzepatide. Third, the study applies Heuristics Miner and Inductive Miner to the same event-log samples and compares the resulting models using fitness, precision, generalization, simplicity, and F-score.

The empirical results provide an initial description of the reporting-transition structures identified through this framework. After preprocessing, the event log contains 496,212 event records representing 188,247 cases. Among the Top 25 reported adverse events, 48,429 cases contain at least two events and contribute to the transition analysis. The most frequent directly-follows transition is Nausea → Vomiting, with 4,459 occurrences. Heuristics Miner produces precision values of 0.9671–0.9909 and F-scores of 0.8474–0.8682. Inductive Miner produces a fitness value of 1.0000 for all four drugs and precision values of 0.5315–0.7481. These results show different model characteristics between the two process discovery algorithms. The event sequences analyzed in this study represent reporting sequences rather than clinical onset sequences. The analysis uses *receivedate* to establish temporal order and uses *reaction_number* as a tie-breaker when several events share the same timestamp. Therefore, the identified transitions describe reporting patterns and should not be interpreted as causal relationships, incidence estimates, or measures of clinical risk.

The remainder of this paper is organized as follows. Section 2 describes the research design, data source, preprocessing, event-log construction, process discovery algorithms, and model evaluation procedures. Section 3 presents the data extraction results, adverse event frequency analysis, Directly-Follows Analysis, process discovery results, model evaluation, and drug-specific transition patterns. Section 4 summarizes the findings and states the interpretation boundaries of the analysis.

2. Related Work

Recent pharmacovigilance studies have examined adverse events associated with glucagon-like peptide-1 receptor agonists (GLP-1 RAs) using spontaneous reporting databases. Semaglutide has been evaluated using the FDA Adverse Event Reporting System (FAERS) to compare adverse event patterns between oral and subcutaneous administration. The analysis identified differences in reported adverse events and time-to-onset patterns between the two administration routes [1]. Gastrointestinal adverse events associated with semaglutide have also been evaluated through FAERS-based disproportionality analysis [7]. Another post-marketing study analyzed semaglutide reports and identified several drug–event signals that met the applied disproportionality criteria [8].

Pharmacovigilance studies have also examined adverse events across several GLP-1 RAs. A recent FAERS analysis evaluated neurological adverse events associated with six GLP-1 RAs using disproportionality analysis [2]. Metabolic and nutritional adverse events have also been analyzed across GLP-1 RAs using spontaneous reporting data [9]. A review of serious adverse events associated with GLP-1 agonists examined evidence across several clinical outcome categories [10]. These studies extend safety assessment beyond individual gastrointestinal events but retain the drug–event association as the main analytical unit.

Tirzepatide has received specific attention in post-marketing pharmacovigilance. A FAERS-based study applied disproportionality methods to identify reported adverse event signals associated with tirzepatide [11]. Another analysis examined both the overall reporting population and sex-specific subgroups and reported frequent events involving incorrect dose administration, injection-site reactions, off-label use, and nausea [12]. More recent post-marketing monitoring also evaluated tirzepatide through FAERS-based signal detection [3]. A separate study applied multiple disproportionality algorithms to characterize adverse event reports associated with tirzepatide in type 2 diabetes [4].

These pharmacovigilance studies primarily analyze event frequency, disproportionality, demographic characteristics, system-organ classifications, and time-to-onset information. These methods provide information about reported drug–event associations. However, they do not explicitly represent the sequential relationships among multiple adverse events recorded within the same safety report. This limitation separates conventional pharmacovigilance signal analysis from event-sequence analysis.

Process mining provides a different analytical perspective because it represents observed activities as ordered traces within event logs. A systematic literature review published in 2025 examined process mining applications across healthcare processes and identified applications involving clinical workflows, organizational processes, and process evaluation [5]. A tertiary study published in 2025 also reviewed healthcare process mining research and identified process discovery as a major analytical activity in this field [6]. These studies indicate that process mining has become an established approach for examining structured event data in healthcare.

Healthcare process mining also presents methodological challenges that affect event-log construction and model interpretation. A review of process mining characteristics in healthcare identified process variability, data characteristics, and domain-specific requirements as important considerations [13]. An updated review of 263 healthcare process mining publications examined project stages, domain-expert involvement, and the use of performance indicators [14]. These findings show that healthcare process mining requires both computational analysis and domain-specific interpretation.

Process mining has also been applied to chronic disease data. A systematic review reported growing use of process mining in chronic disease research, particularly for diabetes and cancer pathways [15]. The review also described the combined use of statistical methods and process mining for obtaining information from healthcare data. This direction demonstrates the applicability of process-oriented analysis to longitudinal and pathway-based healthcare data.

Process discovery methods differ in the way they construct models from event logs. A Stable Heuristic Miner was developed to identify common patient pathways from location event logs by applying statistical stability to threshold selection [16]. The method calculates thresholds at statistically stable limits and generates a reference process model from the analyzed event log. Fodina extends heuristic process discovery by addressing limitations of conventional Heuristics Miner and providing additional configuration options for model discovery [17]. These methods illustrate different strategies for handling variability and noise in event-log data. Process model evaluation also requires several quality dimensions. Fitness measures the ability of a model to reproduce behavior recorded in the event log. Precision measures the extent to which a model restricts behavior that is absent from the observed log. Generalization and simplicity provide additional information about model behavior and structural complexity. The X-Processes framework explicitly considers fitness, precision, generalization, and simplicity when searching for process models [18]. These dimensions provide a suitable basis for comparing process discovery models generated from the same event log.

Healthcare event logs can also contain sensitive information. Privacy-preserving process mining has been examined to address privacy requirements while retaining data utility for process analysis [19]. Although public pharmacovigilance data differ from patient-level hospital event logs, privacy-oriented process mining research shows that event-log structure and information governance remain relevant methodological considerations. Process mining has previously been applied to adverse event analysis in healthcare. An event-log-based investigation used process mining to examine adverse events within care processes and demonstrated how process information can support the analysis of process variations associated with adverse events [20]. However, this application focused on care-process data rather than large-scale spontaneous

pharmacovigilance reports. The reviewed literature therefore presents two related but distinct analytical directions. Pharmacovigilance studies of GLP-1 RAs primarily use frequency analysis, disproportionality measures, subgroup analysis, and time-to-onset analysis. Healthcare process mining studies analyze sequential behavior, pathways, process structures, and process-model quality. The reviewed studies provide limited integration of these approaches for examining transitions among multiple adverse events recorded within individual GLP-1 RA safety reports.

This study addresses this methodological gap by transforming OpenFDA/FAERS reports into a process mining event log. Each safety report represents a case, while each reported adverse event represents an activity. Directly-Follows Analysis is used to characterize recurring reported-event transitions. Heuristics Miner and Inductive Miner are then applied to the same event-log samples, and the resulting models are compared using fitness, precision, generalization, simplicity, and F-score. The analysis therefore complements conventional pharmacovigilance methods by examining the structure of reported-event sequences.

3. Methods

3.1. Research Design

This study employed a quantitative process mining approach to characterize reported adverse event transition patterns associated with glucagon-like peptide-1 receptor agonists (GLP-1 RAs). The research data consisted of adverse drug event reports obtained from the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS) through the OpenFDA Drug Adverse Event API. OpenFDA provides FAERS records in a machine-readable JSON format and includes information on safety reports, medicinal products, patients, and reported reactions [21]. The data source therefore supports systematic extraction and transformation of large-scale pharmacovigilance records into an event-log structure suitable for computational analysis. The research methodology consisted of nine main stages: (1) data collection, (2) preprocessing and normalization, (3) event-log construction, (4) selection of the Top 25 adverse events, (5) Directly-Follows Analysis, (6) process discovery using Heuristics Miner and Inductive Miner, (7) process-model quality evaluation, (8) selection of the primary model, and (9) drug-specific pattern analysis. Recent healthcare process mining studies have applied event-log transformation, Directly-Follows Graphs, Heuristics Miner, Inductive Miner, and other process discovery algorithms to reconstruct observed healthcare processes and identify recurring process patterns [22, 23].

Process mining was selected because it reconstructs observed process behavior from event logs containing case identifiers, activities, and temporal information [24, 25]. Recent methodological developments have further demonstrated that event attributes and transition features can enrich process models by representing information associated with individual events and consecutive event relationships [26]. In healthcare, process mining has been applied to clinical pathways, patient trajectories, process variation, and decision-support analysis using real-world event data [13, 14, 27]. These applications support the use of process mining for examining sequential structures that are not directly represented by conventional frequency-based pharmacovigilance analysis. Figure 1 illustrates the research stages, beginning with the collection of adverse drug event reports from OpenFDA/FAERS for the 2023–2025 period covering four GLP-1 receptor agonists: semaglutide, liraglutide, dulaglutide, and tirzepatide. The collected data then undergo preprocessing, including data cleaning, drug name normalization, deduplication, and event ordering based on the report receipt date. The processed data are subsequently transformed into an event log, with the safety report serving as the case, the reported adverse event as the activity, and *receivedate* as the timestamp. The Top 25 adverse events, selected based on frequency, are then used to construct the process mining event log. Next, Directly-Follows Analysis is performed to identify reporting sequences between adverse events, followed by process discovery using the Heuristics Miner and Inductive Miner. The resulting models are evaluated based on fitness, precision, generalization, simplicity, and F-score using the same event log sample to ensure a comparable evaluation. Subsequently, drug-specific analysis is conducted, accompanied by Directly-Follows Graph (DFG) and Petri Net visualizations, followed by interpretation of the identified transition patterns within the context of pharmacovigilance.

3.2. Data Source and Collection

The data were obtained from the OpenFDA Drug Adverse Event API, which is sourced from the FDA Adverse Event Reporting System (FAERS). OpenFDA indicates that a single adverse event report may contain multiple

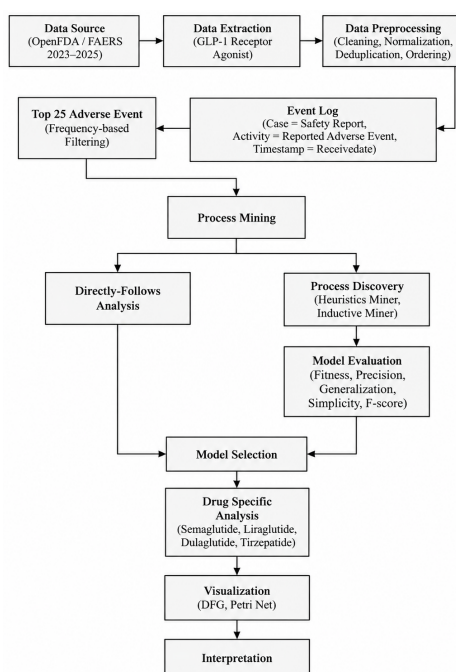


Figure 1. Research Design

medicinal products and multiple reported reactions. Therefore, the presence of a particular drug and a reported reaction within the same report cannot, based solely on the report structure, be automatically interpreted as evidence of a causal transition [21]. Data were collected from January 1, 2023, to December 31, 2025. Four drugs were analyzed: semaglutide, liraglutide, dulaglutide, and tirzepatide. Data retrieval was performed using the OpenFDA harmonized field `patient.drug.openfda.generic_name`, with the search separated by drug and year. The `receivedate` field was used to represent the report receipt date and establish the temporal information for each report. OpenFDA provides `receivedate` as one of the fields that can be used to search and filter adverse event reports [21, 28]. Because the number of retrieved records was relatively large, the data extraction process was performed incrementally using the OpenFDA pagination mechanism. Data for each combination of drug and year were stored in JSONL format before undergoing further preprocessing. The extraction process resulted in 188,734 adverse event reports, with the distribution across the four analyzed drugs presented in Table 1. By drug, the number of reports retrieved was 43,827

Table 1. Adverse Event Report Period 2023–2025

Year	Number of Reports
2023	41,861
2024	56,483
2025	90,390
Total	188,734

reports containing semaglutide, 6,085 reports containing liraglutide, 22,670 reports containing dulaglutide, and 116,152 reports containing tirzepatide across the entire study period. Since a single report may contain more than one target drug, the drug-specific report counts should not be interpreted as unique case counts or summed to represent the total number of reports.

3.3. Data Preprocessing and Normalization

The raw OpenFDA data have a hierarchical structure, as a single report may contain patient information, multiple medicinal products, and multiple reported reactions. Therefore, the data need to be transformed into a tabular structure suitable for process mining. Figure 2 illustrates the preprocessing stage, in which each report is identified using the `safetyreportid` attribute as the report identifier. This attribute serves

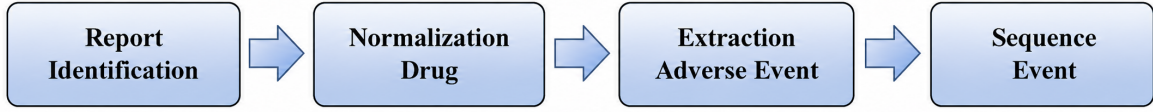


Figure 2. Data Preprocessing and Normalization

as the basis for defining a case in the event log; therefore, reports sharing the same identifier are treated as a single case unit. At this stage, the completeness of the key attributes required for the analysis is also examined, including the report identifier, drug information, adverse event, and report receipt date. The use of `safetyreportid` enables each adverse event to be traced back to its original report while preserving the relational structure between reports and events.

Drug names are then normalized to ensure that each target drug is represented consistently. This step is necessary because the OpenFDA harmonized fields may contain name variations that substantively refer to the same drug. In the study dataset, one of the identified variations was “Oral Semaglutide,” which was normalized to “Semaglutide.” Meanwhile, Liraglutide, Dulaglutide, and Tirzepatide were retained as their respective categories. This normalization was performed to ensure that drug grouping, frequency calculations, and process mining analysis were not affected by variations in drug naming.

Adverse events were extracted from the `patient.reaction.reactionmeddrapt` field. Each reaction term contained in a report was treated as a separate activity in the event log. Accordingly, when a single report contained multiple adverse events, each adverse event was represented as a separate event while retaining the same `safetyreportid`. Following extraction, duplicate events were identified and removed based on the combination of `safetyreportid`, drug, and `reaction_number`. This step produced a more structured representation of adverse events for subsequent analysis.

The extracted events were then ordered to establish the sequence of activities within each case. Ordering was performed primarily based on `receivedate`, representing the date on which the report was received, while `reaction_number` was used as a tie-breaker when multiple adverse events had the same receipt date. This ordering was necessary because process mining requires a defined sequence of activities to construct traces and analyze directly-follows transition between events. It should be emphasized that `receivedate` represents the date on which a report was received by the FAERS/OpenFDA system, rather than the date on which the adverse event occurred or began. Therefore, the event sequence constructed in this study represents a reported-event sequence, rather than a clinical onset sequence, and is not used to infer causal transition between adverse events. The use of `reaction_number` as a tie-breaker serves only to establish a deterministic ordering when multiple events share the same timestamp.

3.4. Heuristics Miner

The Heuristics Miner was used to discover process structures based on the frequency and dependency transition between activities. This heuristic approach is useful for handling real-world event logs that may contain noise and infrequent behavior. In this study, the Heuristics Miner generated a Petri Net for each GLP-1 RA. The resulting models were then evaluated using five process model quality metrics. The use of multiple quality dimensions is necessary because a model with high fitness does not necessarily achieve high precision. A highly permissive model may represent the observed event log well while also allowing behavioral patterns that are not actually observed in the log [5]. The following formula for heuristics miners is presented in [16].

$$a \Rightarrow_w b = \frac{|a >_w b| - |b >_w a|}{|a >_w b| + |b >_w a| + 1} \quad (1)$$

where the term $a \Rightarrow_w b$ denotes the dependency measure between activities a and b . The term $|a >_w b|$ represents the number of times activity a is directly followed by activity b , while $|b >_w a|$ represents the number of times activity b is directly followed by activity a . The constant $+1$ in the denominator prevents division by zero and stabilizes the dependency calculation.

3.5. Inductive Miner

Inductive Miner was used as a comparative process discovery algorithm. The algorithm constructs a process model recursively based on the structure of the event log and produces a structured process model. Inductive

Miner was selected as the comparison algorithm because it is a widely used process discovery approach and is capable of generating structured process models from event logs. In this study, the resulting Inductive Miner models were evaluated using the same quality metrics and event log samples as the Heuristics Miner models, enabling a comparable assessment of their ability to represent the observed reported adverse-event sequences [25]. The comparison between Heuristics Miner and Inductive Miner was performed using the same event-log samples, ensuring that differences in model evaluation were not attributable to differences in the input data. This approach provides a consistent basis for comparing the performance and quality of the process models generated by the two algorithms.

3.6. Model Evaluation

The process models were evaluated using five metrics: fitness, precision, generalization, simplicity, and F-score. Fitness measures the ability of a model to represent the behavior observed in the event log, whereas precision measures the extent to which the model restricts behavior that is not observed in the log. Generalization assesses the model's ability to represent patterns that have not explicitly appeared in the observed data, while simplicity reflects the complexity of the resulting process model.

4. Results and Discussion

4.1. Results Data Extraction and Event Log Construction

The research data were obtained from the OpenFDA Drug Adverse Event API, which is sourced from the FDA Adverse Event Reporting System (FAERS), covering the 2023–2025 period. Data extraction was performed for four glucagon-like peptide-1 receptor agonists (GLP-1 RAs): semaglutide, liraglutide, dulaglutide, and tirzepatide.

Table 2. Number of Adverse Event Reports Extracted by Drug and Year

Drug	2023	2024	2025	Total
Semaglutide	9,504	14,310	20,013	43,827
Liraglutide	2,043	1,839	2,203	6,085
Dulaglutide	9,512	7,286	5,872	22,670
Tirzepatide	20,802	33,048	62,302	116,152
Total	41,861	56,483	90,390	188,734

Based on Table 2, a total of 188,734 adverse event reports were successfully extracted during the 2023–2025 study period. Tirzepatide accounted for the largest number of reports, with 116,152 reports, followed by semaglutide with 43,827 reports, dulaglutide with 22,670 reports, and liraglutide with 6,085 reports. Temporally, the number of reports increased from 41,861 reports in 2023 to 56,483 reports in 2024 and 90,390 reports in 2025. The number of tirzepatide reports was substantially higher than that of the other drugs and therefore requires careful interpretation. A high number of reports does not necessarily indicate that the drug carries a higher risk, as FAERS is based on a spontaneous reporting system and is influenced by factors such as the extent of drug use, attention toward the product, reporting patterns, and characteristics of the user population. Therefore, in this study, report counts are primarily used to identify reporting patterns, rather than to estimate adverse event incidence or relative risk.

After cleaning, normalization, deduplication, and event log construction, the extracted data comprised 188,734 raw reports from OpenFDA/FAERS during the 2023–2025 study period. The resulting final event log contained 496,212 event records representing 188,247 unique cases. The number of event records was greater than the number of cases because a single report may contain more than one reported adverse event, resulting in multiple event records within a single case.

Figure 3 presents the distribution of final event records by drug and reporting year during the 2023–2025 study period. Tirzepatide had the highest number of event records, with 251,105 events, followed by semaglutide with 157,197 events, dulaglutide with 63,429 events, and liraglutide with 24,481 events. The final event log comprised 6,742 unique reported adverse event terms, which were used as activity labels. This figure represents the number of distinct adverse event types in the event log, whereas the 496,212 event records represent the total occurrences of reported adverse events and serve as the observational units in the event

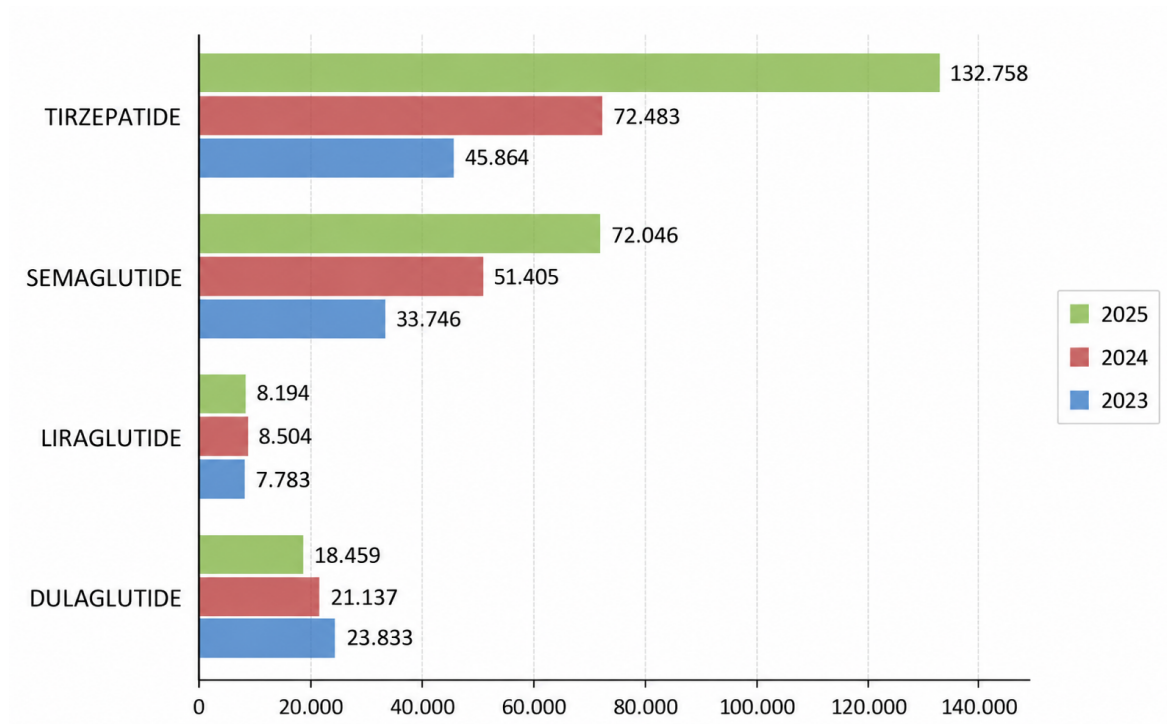


Figure 3. Distribution Final Event Records based on Drug

log. Therefore, the frequency analysis in the subsequent stage was based on the number of event records, rather than the number of distinct adverse event types.

Table 3. Result Event Log

Parameter	Result	Description
Raw reports	188,734	Number of reports extracted from OpenFDA/FAERS
Unique safety report ID	184,969	Number of unique reports IDs in the extracted data
Unique cases	188,247	Number of cases represented in the event log
Final event records	496,212	Number of all event records after preprocessing
Unique adverse events	6,742	Number of unique adverse event types/labels
Duplicate events removed	50,340	Number of duplicate events removed
Invalid timestamp	0	Events lacking a valid timestamp
Multi-target reports	3,252	Reports with more than one target drug
Period	2023–2025	

Table 3 shows that a single report may contain more than one adverse event, resulting in a higher number of event records than reports. This structure provides the basis for constructing traces, which are required for subsequent process mining analysis.

4.2. Selection Top 25 Adverse Event

After the final event log was constructed, a frequency analysis was performed for each reported adverse event term. The final event log contained 496,212 event records. Based on the frequency of occurrence of each event

term, the 25 most frequently reported adverse events (Top 25) were selected using the following formula:

$$\text{Percentage}_i = \frac{\text{Frequency}_i}{496,212} \times 100\% \quad (2)$$

The percentage was used to indicate the proportion of occurrences of each reported adverse event relative to the total number of event records in the final event log. The results of the selection are presented in Table 4.

Table 4. Top 25 Reported Adverse Event

Item	Reported Adverse Event	Frequency	Persentase
1	Incorrect dose administered	27,896	5.62%
2	Nausea	21,336	4.30%
3	Injection site pain	15,293	3.08%
4	Off label use	13,338	2.69%
5	Diarrhoea	12,897	2.60%
6	Vomiting	12,319	2.48%
7	Extra dose administered	9,088	1.83%
8	Constipation	8,765	1.77%
9	Drug ineffective	6,591	1.33%
10	Blood glucose increased	6,363	1.28%
11	Decreased appetite	6,360	1.28%
12	Fatigue	5,938	1.20%
13	Injection site haemorrhage	5,378	1.08%
14	Abdominal pain upper	5,371	1.08%
15	Impaired gastric emptying	5,239	1.06%
16	Product dose omission issue	5,222	1.05%
17	Weight decreased	4,967	1.00%
18	Accidental underdose	4,941	1.00%
19	Headache	4,762	0.96%
20	Injection site erythema	4,695	0.95%
21	Abdominal pain	4,622	0.93%
22	Weight increased	4,314	0.87%
23	Dizziness	4,314	0.87%
24	Inappropriate schedule of product administration	3,834	0.77%
25	Dyspepsia	3,644	0.73%

Based on Table 4, Incorrect dose administered ranked first with the highest frequency, accounting for 27,896 event records (5.62%), followed by Nausea with 21,336 event records (4.30%) and Injection site pain with 15,293 event records (3.08%). In addition to adverse events representing clinical symptoms, such as nausea, vomiting, diarrhoea, constipation, and abdominal pain, other events were related to product use or administration, including incorrect dose administered, extra dose administered, product dose omission issue, and inappropriate schedule of product administration. Collectively, the 25 selected adverse events accounted for 207,487 event records.

$$\frac{207,487}{496,212} \times 100\% = 41.81\% \quad (3)$$

Therefore, the Top 25 reported adverse events accounted for 41.81% of all event records in the final event log.

4.3. Top 25 Event Log Characteristics

After filtering based on the Top 25 reported adverse events, a total of 207,487 event records were obtained from 120,235 cases, comprising 25 activities. Of these, 48,429 cases (40.28%) contained at least two events and were therefore eligible for Directly-Follows transition analysis, whereas 71,806 cases (59.72%) contained only a single event. Furthermore, the 48,429 multi-event cases comprised 135,681 event records and generated 87,252 transition records, which were used to identify transition patterns among the reported adverse events.

Table 5. Top 25 Event Log Characteristics

Parameter	Result
Top 25 Event records	207,487
Cases	120,235
Activities	25
Cases ≥ 2 event	48,429 (40.28%)
Cases 1 event	71,806 (59.72%)
Event records on multi-event cases	135,681
Transition records	87,252

Based on Table 5, the Top 25 event log had an average trace length of 1.73 events, with a median of 1 event. This indicates that most cases contained only a single event included in the Top 25 set. However, 48,429 cases contained two or more events and therefore provided information on transition patterns among reported adverse events.

4.4. Directly Follows Analysis Results

Directly Follows analysis was conducted on 48,429 multi-event cases data which resulted in 87,252 transition records from the Top 25 activities. This analysis was used to identify the transition patterns of reported adverse events that appeared sequentially in the trace, the results are shown in Table 6.

Table 6. Result of Directly Follows Analysis

Item	Transition ($\rightarrow$) Reported Adverse Event	Frequency
1	Nausea $\rightarrow$ Vomiting	4,459
2	Incorrect dose administered $\rightarrow$ Injection site pain	2,267
3	Vomiting $\rightarrow$ Diarrhoea	1,924
4	Vomiting $\rightarrow$ Nausea	1,721
5	Impaired gastric emptying $\rightarrow$ Nausea	1,494
6	Nausea $\rightarrow$ Diarrhoea	1,411
7	Nausea $\rightarrow$ Constipation	1,245
8	Diarrhoea $\rightarrow$ Nausea	1,222
9	Incorrect dose administered $\rightarrow$ Injection site haemorrhage	1,101
10	Diarrhoea $\rightarrow$ Vomiting	1,070

Based on Table 6, the analysis showed that the most frequent transition was Nausea $\rightarrow$ Vomiting, occurring 4,459 times, followed by Incorrect dose administered $\rightarrow$ Injection site pain with 2,267 occurrences, and Vomiting $\rightarrow$ Diarrhoea with 1,924 occurrences. Other dominant transition patterns also involved gastrointestinal events, including Vomiting $\rightarrow$ Nausea, Nausea $\rightarrow$ Diarrhoea, and Diarrhoea $\rightarrow$ Nausea. Overall, the transition patterns were predominantly associated with two major groups: gastrointestinal transitions and drug administration-related transitions. These findings represent reported adverse-event sequences within the event log, rather than causal transition or clinical event sequences. The event ordering was established based on the receivedate timestamp available in the FAERS data.

4.5. Results Process Discovery and Model Evaluation

In addition to Directly-Follows Analysis, process discovery was performed to obtain process models representing the relational structure among reported adverse events. Two algorithms were employed in this study: Heuristics Miner (HM) and Inductive Miner (IM). The Heuristics Miner was used to identify transition between activities based on their frequency and dependency, whereas the Inductive Miner was employed as a comparative algorithm to construct a process tree based on the structure of the event log. Both algorithms were applied to the same event-log sample for each drug, ensuring that the resulting evaluation metrics could be compared consistently.

Table 7. Results of Process Discovery and Model Evaluation

Drug	Algorithm	Fitness	Precision	Generalization	Simplicity	F-score
Dulaglutide	HM	0.7776	0.9827	0.6901	0.4405	0.8682
Dulaglutide	IM	1.0000	0.6478	0.9160	0.5543	0.7862
Liraglutide	HM	0.7497	0.9743	0.4864	0.3511	0.8474
Liraglutide	IM	1.0000	0.5315	0.8858	0.4146	0.6941
Semaglutide	HM	0.7865	0.9671	0.6114	0.3344	0.8675
Semaglutide	IM	1.0000	0.6876	0.9174	0.4959	0.8149
Tirzepatide	HM	0.7711	0.9909	0.6191	0.3635	0.8673
Tirzepatide	IM	1.0000	0.7481	0.9334	0.4775	0.8559

Table 7 shows that the Inductive Miner achieved a fitness value of 1.0000 across all four drugs, whereas the Heuristics Miner produced fitness values ranging from 0.7497 to 0.7865. In contrast, the Heuristics Miner achieved higher precision values, ranging from 0.9671 to 0.9909, compared with the Inductive Miner, which ranged from 0.5315 to 0.7481. Differences were also observed in the F-score results. The Heuristics Miner produced F-score values ranging from 0.8474 to 0.8682, while the Inductive Miner produced values ranging from 0.6941 to 0.8559. On average, the Heuristics Miner obtained a fitness of 0.7712, precision of 0.9787, generalization of 0.6017, simplicity of 0.3724, and F-score of 0.8626. Meanwhile, the Inductive Miner obtained a fitness of 1.0000, precision of 0.6537, generalization of 0.9131, simplicity of 0.4856, and F-score of 0.7878. Based on these results, the Heuristics Miner was used as the primary model for analyzing reported adverse event transition patterns because it consistently achieved higher precision and F-score across the four drugs. The Inductive Miner was retained as a comparative model to examine differences in the characteristics of the resulting process models.

4.6. Drug Specific Analysis

The drug-specific analysis was conducted to identify the characteristics of adverse event transition patterns for each GLP-1 receptor agonist, namely semaglutide, liraglutide, dulaglutide, and tirzepatide. The analysis used directly-follows transitions from cases containing at least two Top 25 adverse events. The results showed that semaglutide, liraglutide, and dulaglutide exhibited transition patterns predominantly involving nausea, vomiting, diarrhoea, and other gastrointestinal events. For tirzepatide, in addition to gastrointestinal patterns, relatively frequent transitions were observed between incorrect dose administered and various injection-site reactions, including injection site pain, injection site haemorrhage, and injection site erythema. The drug specific transition patterns are presented in Table 8.

Table 8. Transition Pattern Results Based on Drug

Drug	Dominant Transition Patterns
Semaglutide	Nausea → Vomiting; Vomiting → Diarrhoea; Impaired gastric emptying → Nausea
Liraglutide	Nausea → Vomiting; Impaired gastric emptying → Nausea; Vomiting → Diarrhoea
Dulaglutide	Nausea → Vomiting; Impaired gastric emptying → Nausea; Extra dose administered → Inappropriate schedule
Tirzepatide	Incorrect dose administered → Injection site pain; Nausea → Vomiting; Incorrect dose administered → Injection site haemorrhage

5. Conclusion

Based on the results of the process mining analysis of adverse event reports for GLP-1 receptor agonist therapies during the 2023–2025 period, the event-log-based approach was able to characterize reported adverse-event sequences and identify directly-follows transition patterns. Following cleaning, normalization, extraction, temporal ordering, and deduplication, the resulting dataset comprised 496,212 event records from 188,247 unique cases. Frequency-based filtering identified 25 reported adverse events as activities for process mining.

Among the 120,235 cases containing at least one Top-25 activity, 48,429 cases contained at least two adverse events and were therefore eligible for directly-follows transition analysis. The analysis revealed recurring transition patterns, particularly Nausea → Vomiting, Vomiting → Diarrhoea, and transitions involving Incorrect dose administered and various injection-site events. The process discovery results demonstrated different characteristics between Heuristics Miner and Inductive Miner. Using the same event-log samples, Inductive Miner achieved a fitness of 1.0000 across all four drugs, whereas Heuristics Miner achieved fitness values ranging from 0.7497 to 0.7865. In contrast, Heuristics Miner achieved higher precision values (0.9671–0.9909) than Inductive Miner (0.5315–0.7481), and also produced higher F-score values across the four drugs. These findings indicate that the two algorithms provide different representations of the observed event-log behavior, emphasizing the importance of evaluating process models using multiple quality dimensions rather than a single metric. The drug-specific analysis further indicated that transition patterns varied across the four GLP-1 receptor agonists. Semaglutide, liraglutide, and dulaglutide predominantly exhibited transition patterns involving gastrointestinal events, whereas tirzepatide showed prominent transitions involving Incorrect dose administered and injection-site events.

Author Contributions

Conceptualization: A.P.H.H., W.I.Z., F.H., and J.H.S.; Methodology: A.P.H.H., W.I.Z., and J.H.S.; Software: A.P.H.H.; Validation: W.I.Z., F.H., and J.H.S.; Formal analysis: A.P.H.H. and W.I.Z.; Investigation: A.P.H.H. and F.H.; Resources: W.I.Z. and J.H.S.; Data curation: A.P.H.H. and F.H.; Writing—original draft preparation: A.P.H.H.; Writing—review and editing: W.I.Z., F.H., and J.H.S.; Visualization: A.P.H.H.; Supervision: J.H.S.; Project administration: A.P.H.H. and J.H.S.; Funding acquisition: Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The lung cancer dataset used in this study was obtained from Kaggle. The dataset contains 1,000 patient records with 25 attributes related to lung cancer risk factors and symptoms.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

ChatGPT, developed by OpenAI, was used to assist with language editing and manuscript formatting. The authors reviewed and verified all generated content and remain fully responsible for the accuracy, interpretation, and integrity of the manuscript.

References

- [1] Niu K, Fan M, Gao W, Chen C, Dai G. Adverse events in different administration routes of semaglutide: a pharmacovigilance study based on the FDA adverse event reporting system. *Frontiers in Pharmacology*. 2024;15:1414268.
- [2] Chen H, Liu S, Gao S, Shi H, Yan Y, Xu Y, et al. Pharmacovigilance Analysis of Neurological Adverse Events Associated with GLP-1 Receptor Agonists Based on the FDA Adverse Event Reporting System. *Scientific Reports*. 2025;15(1):18063. Available from: <https://doi.org/10.1038/s41598-025-01206-9>.
- [3] Chen H, Ding Y, Shan Y. Post-marketing safety monitoring of tirzepatide: a pharmacovigilance study based on the FAERS database. *Expert Opinion on Drug Safety*. 2025;24(8):959-67.

- [4] Zhang Z, Li J, Zheng J, Liang Y, Ma L, Su L. Tirzepatide Safety in Type 2 Diabetes: A Disproportionality Analysis of Adverse Events Using the FDA FAERS Database. *Endocrine Connections*. 2025;14(7):e250205. Available from: <https://doi.org/10.1530/EC-25-0205>.
- [5] Aversano L, Iammarino M, Madau A, Pirlo G, Semeraro G. Process mining applications in healthcare: a systematic literature review. *PeerJ Computer Science*. 2025;11:e2613.
- [6] Santos A, Leal GCL, Balancieri R. Process Mining in Healthcare: A Tertiary Study. *BMC Medical Informatics and Decision Making*. 2025;25:306. Available from: <https://doi.org/10.1186/s12911-025-02967-z>.
- [7] Shu Y, He X, Wu P, Liu Y, Ding Y, Zhang Q. Gastrointestinal Adverse Events Associated with Semaglutide: A Pharmacovigilance Study Based on FDA Adverse Event Reporting System. *Frontiers in Public Health*. 2022;10:996179.
- [8] Du Y, et al. A Real-World Disproportionality Analysis of Semaglutide: Post-Marketing Pharmacovigilance Data. *Journal of Diabetes Investigation*. 2024;15(10):1422-33.
- [9] He L, et al. Pharmacovigilance Study of GLP-1 Receptor Agonists for Metabolic and Nutritional Adverse Events. *Frontiers in Pharmacology*. 2024;15:1416985.
- [10] Tobaiqy M. A review of serious adverse events linked with GLP-1 agonists in type 2 diabetes mellitus and obesity treatment. *Pharmacological Reports*. 2024;76(5):981-90.
- [11] Liu L. A Real-World Data Analysis of Tirzepatide in the FDA Adverse Event Reporting System (FAERS) Database. *Frontiers in Pharmacology*. 2024;15:1397029.
- [12] Ou Y, et al. Analysis of Tirzepatide in the US FDA Adverse Event Reporting System (FAERS): A Focus on Overall Patient Population and Sex-Specific Subgroups. *Frontiers in Pharmacology*. 2024;15:1463657.
- [13] Munoz-Gama J, et al. Process mining for healthcare: Characteristics and challenges. *Journal of Biomedical Informatics*. 2022;127:103994.
- [14] De Roock E, Martin N. Process mining in healthcare - An updated perspective on the state of the art. *Journal of Biomedical Informatics*. 2022;127:103995.
- [15] Chen K, Abtahi F, Carrero JJ, Fernandez-Llatas C, Seoane F. Process Mining and Data Mining Applications in the Domain of Chronic Diseases: A Systematic Review. *Artificial Intelligence in Medicine*. 2023;144:102645.
- [16] Namaki Araghi S, Fontanili F, Lamine E, Okongwu U, Benaben F. Stable heuristic miner: Applying statistical stability to discover the common patient pathways from location event logs. *Intelligent Systems with Applications*. 2022;14:200071.
- [17] vanden Broucke SKLM, De Weerd J. Fodina: A Robust and Flexible Heuristic Process Discovery Technique. *Decision Support Systems*. 2017;100:109-18.
- [18] Fantinato M, Peres SM, Reijers HA. X-Processes: Process model discovery with the best balance among fitness, precision, simplicity, and generalization through a genetic algorithm. *Information Systems*. 2023;119:102247.
- [19] Pika A, Wynn MT, Budiono S, ter Hofstede AHM, van der Aalst WMP, Reijers HA. Privacy-Preserving Process Mining in Healthcare. *International Journal of Environmental Research and Public Health*. 2020;17(5):1612.
- [20] Caron F, Vanthienen J, Vanhaecht K, Van Limbergen E, Deweerdt J, Baesens B. A process mining-based investigation of adverse events in care processes. *Health Information Management Journal*. 2014;43(1):16-25.
- [21] U S Food and Drug Administration. openFDA: Drug Adverse Event API; 2026. Accessed: Sep. 11, 2026. <https://open.fda.gov/apis/drug/event/>.

- [22] Rabbi F, Banik D, Ibne Hossain NU, Sokolov A. Using Process Mining Algorithms for Process Improvement in Healthcare. *Healthcare Analytics*. 2024;5:100305.
- [23] Özdağoğlu G, Damar M, Erenay FS, Damar HT, Himmetoğlu O, Pinto AD. Monitoring Patient Pathways at a Secondary Healthcare Services through Process Mining via Fuzzy Miner. *BMC Medical Informatics and Decision Making*. 2025;25:199.
- [24] van der Aalst W. Process Mining in the Large. In: van der Aalst W, editor. *Process Mining: Data Science in Action*. Berlin, Heidelberg: Springer; 2016. p. 353-85.
- [25] van der Aalst WMP. Foundations of Process Discovery. In: van der Aalst WMP, Carmona J, editors. *Process Mining Handbook*. Cham: Springer International Publishing; 2022. p. 37-75.
- [26] Rifki O, Peng Z, Perrier L, Xie X. Process Mining with Event Attributes and Transition Features for Care Pathway Modelling. *International Journal of Production Research*. 2025;63(10):3684-708.
- [27] Eili MY, Rezaeenour J, Roozbahani MH. Predicting Clinical Pathways of Traumatic Brain Injuries (TBIs) through Process Mining. *npj Digital Medicine*. 2025;8:112.
- [28] Office of the Commissioner. Data Mining at the Center for Biologics Evaluation and Research; 2022. Accessed: Sep. 11, 2026. <https://www.fda.gov/science-research/data-mining/data-mining-center-biologics-evaluation-and-research>.