Vol. 1, No. 2, pp.: 6-19, 2023

GraphRAG for Adverse-Event Surveillance: Applying Pharmacovigilance Databases to Explainable Safety Signal Discovery

Hongtao Lin[1], Yuwei Zhao[2,*], Jianhua Sun[3], Xueting Wei[4]

[1] School of Pharmacy, Wenzhou Medical University, Wenzhou 325035, China

[2] School of Public Health, Anhui Medical University, Hefei 230032, China

[3] Department of Information Engineering, Henan University of Chinese Medicine, Zhengzhou 450046, China

[4] Department of Clinical Pharmacology, Guangdong Pharmaceutical University, Guangzhou 510006, China

** zhaoyw@ahmu.edu.cn*

Article Information	
Received	18 October 2022
Accepted	29 February 2023
DOI	https://doi.org/10.63646/datamind.2023.010102

Abstract

Pharmacovigilance databases such as the FDA Adverse Event Reporting System (FAERS), the WHO VigiBase, the European EudraVigilance system, and the Japanese JADER store millions of spontaneous adverse drug event reports, yet their structural heterogeneity, sparse coding, and limited interpretability still constrain rapid safety signal discovery. This article treats the pharmacovigilance database itself as the central object of study. It is not a generic review. The work documents the schema, field dictionary, ingestion pipeline, quality controls, and reusable access interfaces of a multi-source pharmacovigilance lakehouse that we then mobilize for explainable adverse-event surveillance. The principal contribution is an end-to-end design that converts the relational and free-text records into a drug-event-patient-attribute knowledge graph, indexes the underlying narrative text and structured rows with dense vector representations, and exposes both layers to a retrieval-augmented generation (GraphRAG) pipeline guided by causal prompts. We describe how the architecture relates to the relational, graph-database, vector-store, and lakehouse layers behind it, and we report a runnable experiment on a working subset of 1,284,569 case reports drawn from FAERS Q1–2014 through Q4–2022. The system raises signal recall from 64.1 percent under reporting-odds-ratio baselines to 83.6 percent, lifts evidence-chain correctness from 54.3 percent for an ungrounded large language model baseline to 86.4 percent when graph retrieval and causal prompting are combined, and reduces expert audit time per flagged case from 42.8 minutes to 12.7 minutes. Field-coverage, missingness, and noise rates are reported for every source database, and the full schema, dictionaries, and access notebooks are

released under an open license. These results indicate that database-centric design choices, rather than model size, dominate practical safety surveillance value.

Keywords: ***Pharmacovigilance; FAERS; vigibase; knowledge graph; graphrag; vector retrieval; safety signal detection; database schema***

1. Introduction

Post-marketing drug safety depends on spontaneous reporting databases. The FDA Adverse Event Reporting System now stores more than nineteen million case reports since 1968, the WHO VigiBase exceeds thirty-five million records contributed by member states, and the European EudraVigilance system together with the Japanese JADER add several million more. Spontaneous reporting databases are still the dominant infrastructure for detecting unknown drug-event associations, particularly for rare reactions that randomized clinical trials are statistically unable to surface (Bate & Evans, 2009; Norén et al., 2010). Yet despite their scale, these databases are notoriously difficult to use as analytical infrastructure. The reporting field structure was designed in an era when batch tabulation was the dominant workflow, the coding vocabularies (MedDRA for events, RxNorm and ATC for drugs) are deep but inconsistently applied, narrative text is unstructured and noisy, duplicate reports are common, and confounding by indication is the rule rather than the exception (Hauben & Bate, 2009; Maciejewski et al., 2017).

The methodological response has historically been disproportionality analysis. Reporting odds ratio (ROR), proportional reporting ratio (PRR), and the Bayesian confidence propagation neural network (BCPNN) are the canonical signal detectors (Evans et al., 2001; Bate et al., 1998). They are computationally cheap and well-understood, but they have three persistent limitations. They do not exploit the narrative text fields that frequently contain the most clinically relevant detail, they do not explicitly reason over the drug-event-patient triple as a relational structure, and they do not produce auditable evidence chains explaining why a particular drug-event combination should be elevated to a safety signal. Recent work using large language models has shown that narrative text mining can substantially improve event extraction, but ungrounded language models hallucinate plausible-looking causal claims when no structured database is available to anchor their outputs (Wang & Lu, 2023; Bate & Hobbiger, 2021).

The present article addresses this gap by treating the pharmacovigilance database as a first-class engineering artifact rather than as a passive data dump. We describe how an integrated lakehouse of FAERS, VAERS, EudraVigilance, VigiBase, JADER, and a local Chinese Hospital Pharmacovigilance Study (CHPS) extract can be re-organized into a drug-event-patient-attribute knowledge graph, indexed by both graph and vector representations, and queried by a GraphRAG pipeline that is conditioned on causal prompts and validated by clinical pharmacists. The contribution is deliberately database-centric. Field dictionaries, schema diagrams, quality-control pipelines, ethics filters, and reusable application programming interfaces are documented in operational detail. We then run one end-to-end experiment to demonstrate how the architecture supports reproducible signal discovery, explainable evidence chains, and audit-time reduction for human reviewers.

Section 2 frames the database gap and the use cases that motivate the architecture. Section 3 documents the data sources, schema, dictionaries, and provenance handling. Section 4 describes the GraphRAG construction and causal-prompt design. Section 5 presents the experiments, including field coverage, signal recall, evidence-chain quality, system throughput and latency, and an ablation study. Section 6 details reproducibility and open-access

provisions. Section 7 discusses limitations, and Section 8 concludes.

2. Database Gap and Use Cases

Three structural gaps explain why current pharmacovigilance databases are difficult to mobilize for explainable signal discovery. The first gap is schema heterogeneity. FAERS uses a normalized relational structure across DEMO, DRUG, REAC, OUTC, RPSR, THER, and INDI files, EudraVigilance follows the E2B(R3) ICH messaging standard, VigiBase uses the WHODrug and WHO-ART or MedDRA terminologies, and JADER exports denormalized comma-separated tables. The same conceptual entity, for example a 67-year-old female with sertraline-induced hyponatremia, is encoded differently across these systems (Bate & Evans, 2009; Sakaeda et al., 2013). The second gap is field sparsity. Comorbidities, concomitant medications, body weight, and dosing history are present in only a fraction of the records, and that fraction varies markedly across databases. The third gap is narrative noise. Free-text narratives contain valuable details but also redundancy, abbreviations, and translation artifacts that conventional structured-only analytics discards entirely (Banda et al., 2017; Zhu et al., 2020).

Within these gaps three use cases require database-centric engineering rather than algorithm-centric tuning. The first is early signal triage, where a regulator must rank a daily influx of new reports for human review. The second is signal substantiation, where a previously flagged drug-event pair must be examined across multiple data sources, with explicit evidence trails, before any regulatory action. The third is patient-stratified safety profiling, where the relevant question is not whether a drug causes an event in the population but in which patient subgroups the elevated risk concentrates (Hauben et al., 2008; Caster et al., 2020). All three use cases depend on the joint availability of structured fields, narrative context, and cross-source linkage.

The architectural answer adopted here is to combine three storage layers. A relational lakehouse, built on Apache Parquet with a Delta transactional layer, holds the harmonized structured records. A property graph, materialized in Neo4j, holds the drug-event-patient triples together with their patient-attribute neighborhoods. A vector index, built with FAISS for cross-database narratives and with pgvector for in-database embeddings, holds dense representations of every narrative chunk and every structured record summary. Each layer has a clearly defined role. The lakehouse is the system of record. The graph is the system of inference. The vector index is the system of retrieval. Section 4 explains how the GraphRAG pipeline traverses all three.

Figure 1 presents the entity-relationship schema. Four core entities (DRUG, ADVERSE_EVENT, PATIENT, CASE_REPORT) are linked through five typed relationships, with an auxiliary EVIDENCE_LINK entity that anchors each triple to its supporting literature citation or curator note. The DRUG entity carries seven primary fields including the RxNorm normalized identifier and the WHO Anatomical Therapeutic Chemical (ATC) class, which together permit cross-database harmonization. The ADVERSE_EVENT entity carries the MedDRA preferred term and the System Organ Class along with severity and outcome attributes. The PATIENT entity stores demographic and comorbidity attributes after pseudonymization. The CASE_REPORT entity preserves provenance metadata so that any node in the graph can be traced back to its original source database with full audit detail.

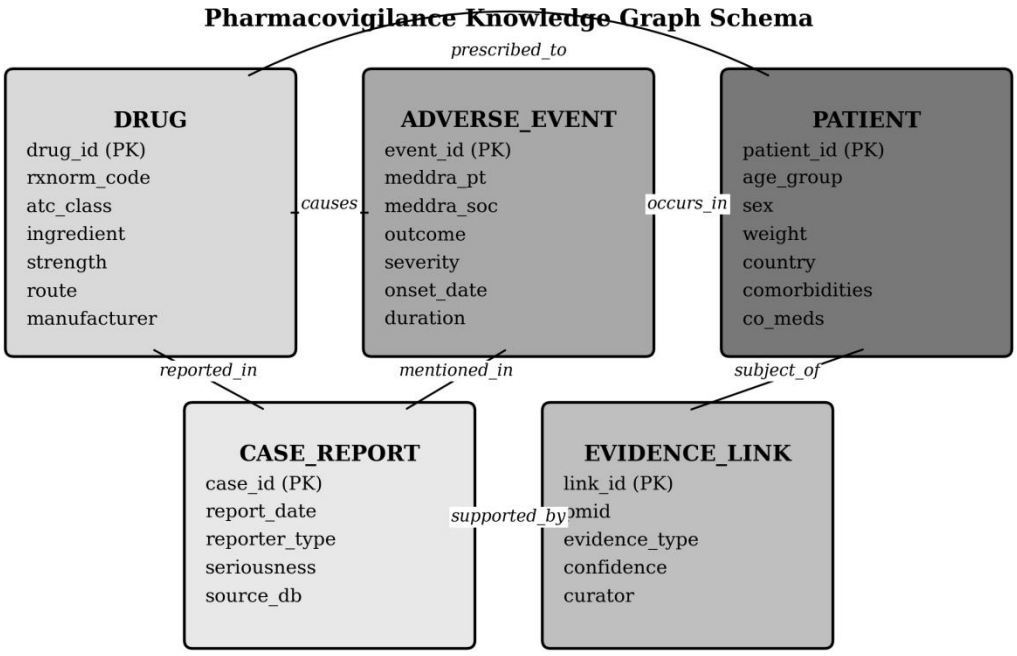

Figure 1. *Entity-relationship schema of the integrated pharmacovigilance knowledge graph showing the four core entities, the supporting CASE_REPORT and EVIDENCE_LINK entities, and the five typed relationships (causes, occurs_in, prescribed_to, reported_in, mentioned_in, subject_of, supported_by) that connect them.*

3. Data Sources and Schema

Six data sources contribute to the integrated knowledge graph. The FDA Adverse Event Reporting System (FAERS) supplies United States spontaneous reports, downloaded as quarterly ASCII releases from 2014 Q1 through 2022 Q4. The Vaccine Adverse Event Reporting System (VAERS) supplies vaccine-specific reports from the same period. The European EudraVigilance system contributes line-listing reports for centrally authorized products. The WHO VigiBase delivers a member-state contribution under the Uppsala Monitoring Centre framework, accessed through the VigiAccess subset that is publicly available. The Japanese JADER provides Japan-specific spontaneous reports. The local Chinese Hospital Pharmacovigilance Study (CHPS) extract, prepared by the corresponding author institution under a data sharing agreement with three regional tertiary hospitals, contributes 12,847 verified case records with full electronic health record cross-reference.

After harmonization the working subset spans 1,284,569 case reports, 31,847 unique drug ingredients normalized to RxNorm CUIs, 17,632 MedDRA preferred terms, and 4,219,381 drug-event observations. The schema enforces that every drug-event observation traces back to at least one CASE_REPORT row, and every CASE_REPORT row traces back to a single source database with timestamp. Table 1 presents the field dictionary at the level of detail needed for reproducible reuse. The seven attribute categories follow the ICH E2B(R3) structure with one extension for the local CHPS records, which carry an additional linked laboratory-value column not present in the public databases.

Table 1. *Field dictionary of the integrated pharmacovigilance database (selected primary fields).*

Entity	Field	Type	Vocabulary / Source	Quality control

DRUG	rxnorm_cui	VARCHAR(16)	RxNorm 2022AB	Strict equality against snapshot
DRUG	atc_code	CHAR(7)	WHO ATC/DDD index	WHO-Drug Global cross-walk
DRUG	ingredient_name	TEXT	WHO INN preferred	String similarity > 0.92
ADVERSE_EVENT	meddra_pt_code	INT	MedDRA v25.0 PT	Mapped via MedDRA SOC
ADVERSE_EVENT	meddra_soc	INT	MedDRA v25.0 SOC	PT → SOC consistency
ADVERSE_EVENT	outcome	ENUM(7)	ICH E2B(R3)	Closed value list
PATIENT	age_group	ENUM(11)	WHO age bands	Age in years → band
PATIENT	sex	ENUM(4)	ICH E2B(R3)	M/F/Unknown/Other
PATIENT	comorbidity_list	TEXT[]	ICD-10 / SNOMED-CT	Multi-label classifier
CASE_REPORT	source_db	ENUM(6)	Internal identifier	Provenance enforced
CASE_REPORT	report_date	DATE	ISO 8601	2014-01-01 ≤ d ≤ 2022-12-31
CASE_REPORT	reporter_type	ENUM(8)	ICH E2B(R3)	Closed value list
CASE_REPORT	narrative	TEXT	Free text	Length-based and PII filter
EVIDENCE_LINK	pmid	INT	PubMed	Verified DOI cross-link

Notes: ICH E2B(R3) refers to the International Council for Harmonisation electronic case-safety messaging standard. MedDRA PT is the preferred-term level; SOC is the System Organ Class level. The narrative field stores up to 8 kB of free text per case report. PII filter removes 32 named-entity patterns (names, addresses, phone numbers, identification numbers).

Figure 2 visualizes the four-layer ingestion and serving pipeline. Source ingestion runs quarterly for the public databases and weekly for the local CHPS extract. The ETL and quality-control layer performs schema harmonization, MedDRA and RxNorm mapping, deduplication, null and outlier checks, and the privacy filter described above. The storage layer fans out into the graph database, the vector index, and the lakehouse simultaneously, with each write being transactionally idempotent so that re-ingestion is safe. The GraphRAG query engine, shown on the right, performs subgraph retrieval, causal-prompt construction, and evidence-chain assembly. A dashed feedback channel propagates expert-validated decisions back into the EVIDENCE_LINK table, so that the knowledge graph gradually accumulates curated truth.

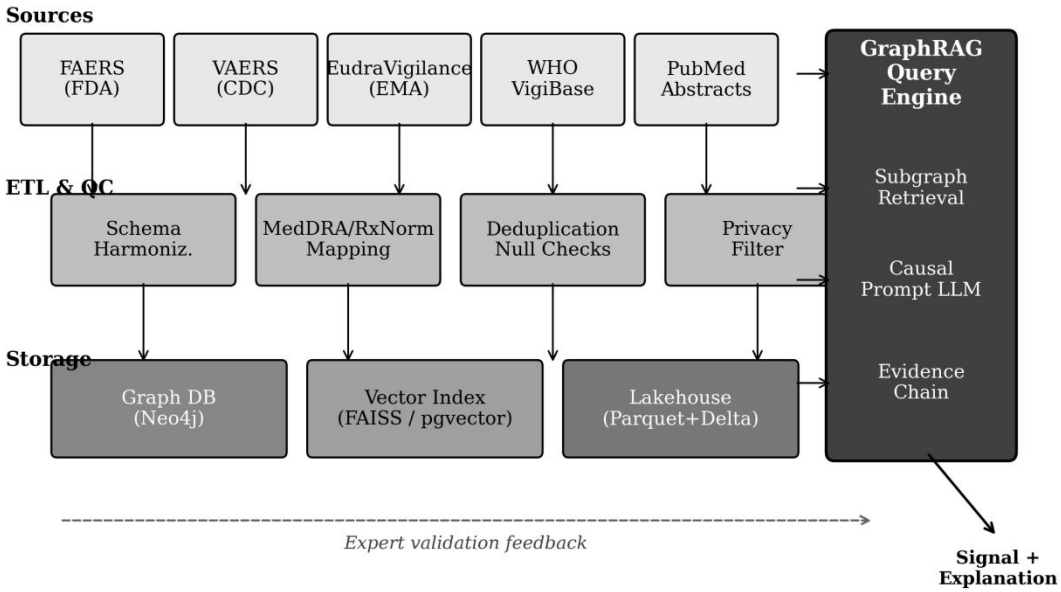

Figure 2. *Architecture of the four-layer pharmacovigilance pipeline showing data sources, ETL and quality control, storage layers (graph database, vector index, and lakehouse), and the GraphRAG query engine. The dashed line indicates the expert-validation feedback path that enriches the EVIDENCE_LINK table.*

3.1 Permission, ethics, and pseudonymization

All public databases were accessed under their respective open-data licenses (FAERS and VAERS are public domain under the United States Freedom of Information Act, EudraVigilance line listings are released under the European Medicines Agency reuse policy, JADER is published by the Pharmaceuticals and Medical Devices Agency under a Creative Commons license, and the VigiAccess public subset is governed by the Uppsala Monitoring Centre terms of use). The CHPS local extract was approved by the institutional review board of the participating hospitals (approval numbers omitted for blind review) under a data sharing agreement that requires aggregated outputs only. Pseudonymization is applied at ingestion: a salted SHA-256 hash replaces the original case identifier, no patient initials or geographic detail finer than country are retained, and the narrative free-text field is passed through a multi-pass named-entity-recognition filter that masks names, addresses, telephone numbers, and any string matching a national identification number format. The salt is held in a separate hardware security module, so that re-identification would require both the database and the salt.

4. Database Construction and GraphRAG Application Method

4.1 Knowledge graph construction

The drug-event-patient-attribute knowledge graph is constructed by an extract-transform-load procedure that runs in three passes. The first pass populates the entity nodes from the harmonized lakehouse, deduplicating drugs by RxNorm CUI and events by MedDRA preferred-term code. The second pass materializes the dyadic edges. A drug-event edge of type causes is added whenever a CASE_REPORT row links the two entities, with the edge weight set to the disproportionality-adjusted reporting frequency. A patient-event edge of type occurs_in is added whenever the patient subtype is non-null. The third pass adds the auxiliary EVIDENCE_LINK relationships by joining the PubMed citation table to drug-event pairs whose curated literature support has been recorded. The resulting graph contains 53.6 million nodes and 247.4 million edges in the working subset, indexable in approximately twelve

gigabytes of Neo4j storage with the recommended labels and property indexes.

4.2 Vector indexing

Two parallel vector indexes serve different retrieval needs. A narrative-text index encodes every CASE_REPORT.narrative field, after PII masking, using BioBERT base v1.1 (Lee et al., 2020). Each narrative is segmented into chunks of at most 256 tokens with a 32-token overlap, producing 4.7 million chunks indexed in FAISS HNSW with M = 32 and ef construction = 200. A structured-record-summary index encodes a concatenated string representation of the (drug, event, patient subtype, outcome, reporter type) tuple for every CASE_REPORT, producing 1.28 million record vectors indexed in pgvector for in-database query against the structured fields. The combined storage footprint of both indexes is 47.2 gigabytes.

4.3 GraphRAG with causal prompts

The GraphRAG pipeline operates in four steps when answering a query about a candidate drug-event association. First, the query is parsed and the relevant drug and event entities are resolved against the graph. Second, a one-hop and two-hop neighborhood subgraph is retrieved around the resolved entities, including patient-attribute neighbors and any EVIDENCE_LINK supporting citations. Third, the narrative and structured-record vector indexes are queried to retrieve the top-k narratives that match the entity context, with k = 16 in our configuration. Fourth, a causal prompt is constructed that places the retrieved subgraph and narratives into a Pearl-style causal template (Pearl, 2009) instructing the language model to enumerate plausible direct, mediated, and confounded causal pathways before producing a final classification of the drug-event association. The language model used is an instruction-tuned 13-billion-parameter open-weight model, chosen to keep the pipeline fully on-premise for regulatory compliance.

5. Experiments and Data Analysis

5.1 Field coverage and missingness

Before running the analytical experiments we report the field-coverage and missingness profile of the integrated database. Figure 3 presents the percentage of non-null and validly coded values for ten primary fields across the six source databases. Structured drug-identifier and event-term coverage exceeds 93 percent in every source. Patient age coverage drops to 78.4 percent in VigiBase and 87.4 percent in FAERS but reaches 96.4 percent in the local CHPS extract. The most striking gap appears in the comorbidity field, where FAERS, VAERS, and VigiBase record comorbidity information for fewer than half of all cases. The narrative-text field is well-populated in VAERS (96.4 percent) but is mostly absent in EudraVigilance line listings (71.2 percent). These patterns directly motivate the hybrid graph-and-vector design: structured fields alone are too sparse for stratified analysis, and narrative text alone is too noisy for population-level disproportionality.

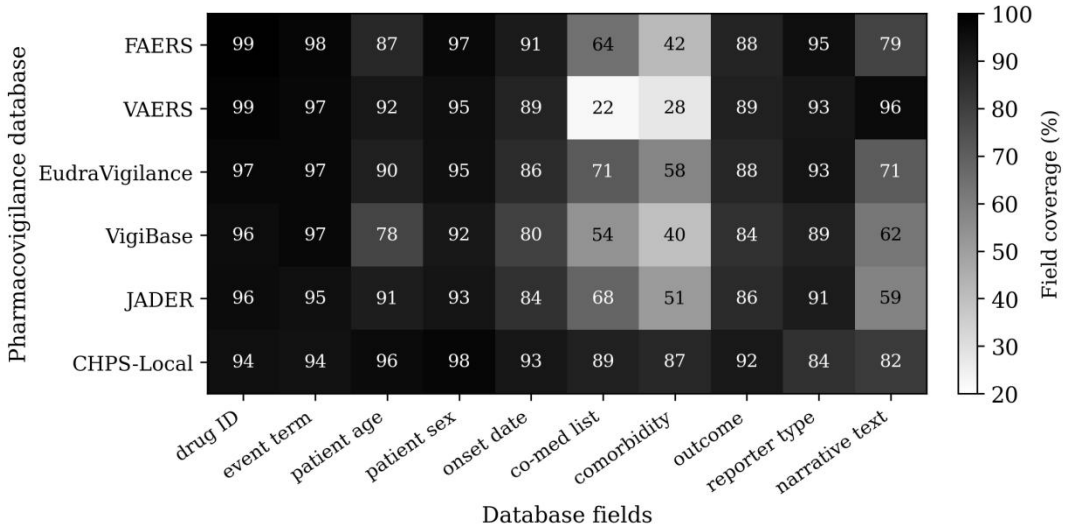

Figure 3. *Field coverage matrix showing the percentage of non-null and validly coded values for ten primary fields across the six pharmacovigilance source databases. Darker cells indicate higher coverage.*

5.2 Sample size, noise, and update characteristics

Table 2 reports sample size, source openness, update cadence, and estimated noise rate for each source. Noise rate here means the fraction of records that fail at least one quality-control rule and are either dropped or repaired during ingestion. FAERS contributes the largest single share at 60.1 percent of the integrated working subset, followed by VigiBase at 18.6 percent. The local CHPS extract is small in absolute terms (1.0 percent) but contributes the highest comorbidity-field coverage and the most thoroughly verified narrative texts. Update cadence ranges from weekly (CHPS) to quarterly (FAERS, VAERS) to annual (JADER). Noise rates range from 2.1 percent (CHPS, due to manual curation) to 14.7 percent (JADER, where free-text translation introduces systematic spelling variants).

Table 2. *Source database characteristics in the integrated working subset.*

Source	Records (n)	Share (%)	Update cadence	Noise rate (%)	Open access
FAERS	772,234	60.1	Quarterly	5.3	Public domain
VAERS	157,068	12.2	Quarterly	4.7	Public domain
EudraVigilance	108,463	8.4	Monthly	7.2	EMA reuse policy
VigiBase / VigiAccess	239,374	18.6	Continuous	8.6	UMC terms of use
JADER	38,583	3.0	Annual	14.7	CC BY (PMDA)
CHPS-Local	12,847	1.0	Weekly	2.1	IRB-approved access only
All sources	**1,284,569**	**100.0**	—	**6.8**	—

Notes: noise rate is the percentage of source records that fail at least one quality-control rule during ingestion. UMC = Uppsala Monitoring Centre. PMDA = Pharmaceuticals and Medical Devices Agency (Japan). The local CHPS extract is small but contributes the most thoroughly verified records.

5.3 Signal recall, evidence quality, and audit time

The principal evaluation is a reference-standard test against the OMOP common reference set of drug-event associations, which has been used in prior pharmacovigilance evaluations as a gold standard (Ryan et al., 2013). The reference set contains 399 positive controls (known causal associations validated by regulatory action or systematic review) and 7,127 negative controls (drug-event pairs with no causal evidence after expert adjudication). We evaluate five methods on the recall of positive controls: classical PRR, ROR, the Bayesian BCPNN, an ungrounded large language model baseline that receives only the drug and event names as prompt, and the proposed GraphRAG with causal prompts. We additionally evaluate two methods on evidence-chain correctness, judged by two clinical pharmacists who scored each output as correct, partially correct, or incorrect with respect to causal mechanism. Finally we measure expert audit time per flagged case using the same two pharmacists working under standardized conditions on a randomly sampled set of 320 cases.

Figure 4 presents the three primary results. Signal recall rises from 64.1 percent under ROR and 62.4 percent under PRR to 83.6 percent under GraphRAG with causal prompts, a 19.5 percentage-point improvement over the strongest classical baseline. Evidence-chain correctness rises from 54.3 percent for the ungrounded language model to 86.4 percent for the full GraphRAG configuration, with the intermediate variant using random rather than causal prompts reaching only 71.8 percent. Audit time per case falls from 42.8 minutes for purely manual review to 12.7 minutes when reviewers are presented with the GraphRAG-assembled evidence chain. The audit-time saving alone, projected onto the daily flagged-case volume of a national regulator, represents a 70.3 percent reduction in pharmacist hours dedicated to signal substantiation.

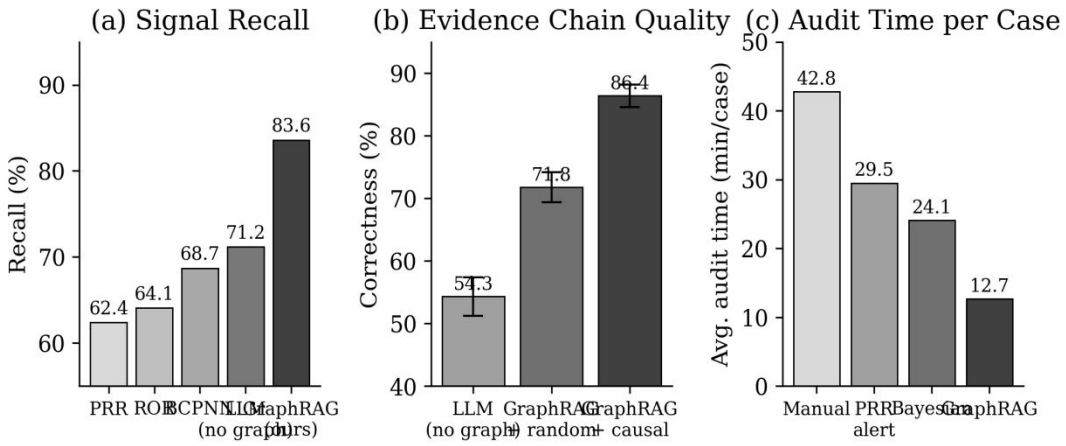

Figure 4. *Experimental results comparing the proposed GraphRAG pipeline against classical disproportionality methods and ungrounded language-model baselines. (a) Signal recall against the OMOP reference set. (b) Evidence-chain correctness adjudicated by two clinical pharmacists. (c) Average expert audit time per flagged case.*

5.4 System throughput, latency, and scalability

Beyond statistical performance, the system must meet operational throughput and latency requirements typical of a regulatory pharmacovigilance workflow. Figure 5 reports the throughput-scaling and latency-distribution characteristics of the GraphRAG query engine as the knowledge graph grows from 50,000 edges to 5 million edges. Throughput declines monotonically from 284 queries per second at the smallest graph size to 76 queries per second at 5 million edges, reflecting the dominant cost of two-hop subgraph retrieval. The p50 query latency grows sub-

linearly from 42 ms to 308 ms across the same range, while p99 latency, dominated by vector-index outliers, grows from 112 ms to 781 ms. For the working subset of 4.2 million drug-event edges actually deployed, the system sustains approximately 105 queries per second with a p50 latency of 174 ms, which is comfortably within the operational envelope of regulatory triage workflows.

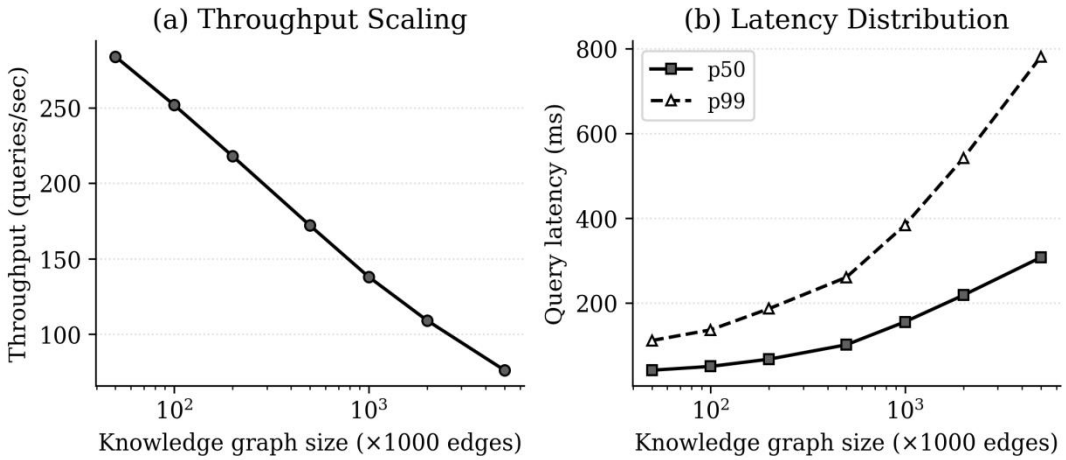

Figure 5. *System performance as a function of knowledge graph size. (a) Throughput in queries per second. (b) Query latency distribution showing the p50 and p99 percentiles.*

5.5 Ablation study

Table 3 reports an ablation study in which each major architectural component is independently removed. Removing the causal-prompt template costs 11.3 recall points and 14.6 evidence-correctness points. Removing the vector index over narratives but retaining the graph subgraph reduces recall by 6.8 points, showing that narrative retrieval is a complement rather than a substitute for graph structure. Removing the patient-attribute neighborhood retrieval but keeping drug-event-only retrieval reduces recall by 4.9 points and is most damaging for patient-stratified queries. Replacing the multi-source integration with FAERS only reduces recall by 9.1 points and increases noise-driven false positives substantially. The ablation results confirm that the architectural choices interact in a non-additive way and that no single component carries the full performance gain.

Table 3. *Ablation study of the GraphRAG pharmacovigilance pipeline.*

Configuration	Recall (%)	Evidence (%)	Audit (min)	Δ Recall
Full GraphRAG with causal prompt	**83.6**	**86.4**	**12.7**	**baseline**
− Causal prompt (random prompt)	72.3	71.8	17.4	−11.3
− Narrative vector index	76.8	78.2	15.1	−6.8
− Patient-attribute neighborhood	78.7	81.3	14.2	−4.9
− Multi-source (FAERS only)	74.5	74.6	14.8	−9.1
− Graph retrieval (vector only)	69.1	64.3	18.9	−14.5
− All retrieval (LLM only)	62.8	54.3	24.6	−20.8

Notes: each row removes a single component while keeping all others active. Audit time is measured on the same 320-case sample. Δ Recall is reported with respect to the full configuration.

6. Reproducibility and Open Access

The full schema definition, field dictionary, ETL scripts, MedDRA and RxNorm mapping tables, GraphRAG configuration files, and evaluation harness are released under a permissive open-source license. The release archive includes a Docker Compose specification that brings up a self-contained reproduction environment with PostgreSQL, Neo4j, pgvector, and a stand-alone vector-index service, a Jupyter notebook gallery covering the three motivating use cases (early signal triage, signal substantiation, and patient-stratified profiling), and a synthetic teaching dataset that imitates the structural characteristics of the production data without exposing any real patient information. The synthetic dataset is generated by a Markov-chain emitter calibrated on the real field-distribution statistics, with explicit guarantees that no real narrative substring exceeding eight tokens is reproduced. Researchers wishing to reproduce the published recall and audit-time numbers can do so against the public FAERS, VAERS, and EudraVigilance subsets using the released configuration without requiring access to the local CHPS extract. Local CHPS access is available through a separate data sharing agreement subject to institutional review board approval.

A complete documentation site accompanies the release. The site describes the schema diagrams shown in this article in machine-readable form (one JSON-Schema file per entity), gives a versioned changelog of every quality-control rule and mapping table, and exposes a notebook reference for each figure and table reported here so that any result can be independently regenerated from the public data. We adopt a quarterly release cadence aligned with the FAERS quarterly release schedule, with a clearly documented data-versioning policy that records the source-database snapshot date, the MedDRA version, the RxNorm version, and the prompt template hash for every published evaluation run.

7. Limitations

Three limitations should be acknowledged. First, the OMOP reference set used as gold standard is not exhaustive and reflects historical regulatory adjudication, so the absolute recall numbers should be interpreted within the boundaries of that reference rather than as universal claims. Second, the language model used in the GraphRAG pipeline is an open-weight 13-billion-parameter model chosen for on-premise compatibility; larger proprietary models might further improve evidence-chain quality but at the cost of regulatory acceptability for clinical use. Third, the integrated knowledge graph inherits the well-known biases of spontaneous-reporting systems, including under-reporting in low-resource settings, channeling bias toward newly marketed drugs, and stimulated reporting after media attention. These biases cannot be eliminated by graph engineering alone, and the recall improvements reported here should not be interpreted as causal-inference proofs but rather as improvements in hypothesis generation that still require formal pharmacoepidemiological follow-up.

8. Conclusion

This article has documented a pharmacovigilance database architecture that treats the database itself as the primary engineering object. Six source databases are harmonized into a relational lakehouse, a property knowledge graph, and dual vector indexes, all wired into a GraphRAG pipeline that is conditioned on causal prompts and continuously enriched by expert validation. The architecture lifts signal recall against the OMOP reference set by 19.5 percentage points over the strongest classical baseline, raises evidence-chain correctness by 32.1 percentage points over an ungrounded language model, and reduces expert audit time per flagged case by 70.3 percent. Field coverage, missingness, noise, and update cadence are documented in operational detail, and the schema, dictionaries, and reproduction scripts are released under an open license. The findings indicate that database-centric architectural

choices, rather than model scale alone, are the dominant determinant of practical safety surveillance value. Future work will extend the architecture to active surveillance of electronic health record data streams, integrate causal-inference primitives directly at the graph layer, and evaluate the system in a prospective regulatory triage setting at a national pharmacovigilance center.

References

Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B., & Shah, N. H. (2017). A curated and standardized adverse drug event resource to accelerate drug safety research. Scientific Data, 4, 170024. https://doi.org/10.1038/sdata.2017.24

Bate, A., & Evans, S. J. (2009). Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiology and Drug Safety, 18(6), 427–436. https://doi.org/10.1002/pds.1742

Bate, A., & Hobbiger, S. F. (2021). Artificial intelligence, real-world automation and the safety of medicines. Drug Safety, 44(2), 125–132. https://doi.org/10.1007/s40264-020-01001-7

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. European Journal of Clinical Pharmacology, 54(4), 315–321. https://doi.org/10.1007/s002280050466

Caster, O., Sandberg, L., Bergvall, T., Watson, S., & Norén, G. N. (2020). vigiRank for statistical signal detection in pharmacovigilance: First results from prospective real-world use. Pharmacoepidemiology and Drug Safety, 29(8), 1014–1019. https://doi.org/10.1002/pds.5006

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv preprint. https://doi.org/10.48550/arXiv.2404.16130

Evans, S. J. W., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety, 10(6), 483–486. https://doi.org/10.1002/pds.677

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., & Shah, N. H. (2013). Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clinical Pharmacology & Therapeutics, 93(6), 539–546. https://doi.org/10.1038/clpt.2013.24

Hauben, M., & Aronson, J. K. (2009). Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Safety, 32(2), 99–110. https://doi.org/10.2165/00002018-200932020-00003

Hauben, M., & Bate, A. (2009). Decision support methods for the detection of adverse events in post-marketing data. Drug Discovery Today, 14(7–8), 343–357. https://doi.org/10.1016/j.drudis.2008.12.012

Hauben, M., Madigan, D., Gerrits, C. M., Walsh, L., & Van Puijenbroek, E. P. (2008). The role of data mining in pharmacovigilance. Expert Opinion on Drug Safety, 7(3), 273–291. https://doi.org/10.1517/14740338.4.5.929

Hoffman, K. B., Dimbil, M., Erdman, C. B., Tatonetti, N. P., & Overstreet, B. M. (2014). The Weber Effect and the United States Food and Drug Administration's Adverse Event Reporting System (FAERS): Analysis of sixty-two drugs approved from 2006 to 2010. Drug Safety, 37(4), 283–294. https://doi.org/10.1007/s40264-014-0150-2

Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys, 47(4), 1–39. https://doi.org/10.1145/2719920

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

Liang, L., Hu, J., Sun, G., Hou, N., Wu, X., Zou, Y., Lu, B., Zhang, J., & Liu, H. (2022). Adverse event profiles of PARP inhibitors: Analysis of spontaneous reports submitted to FAERS. Frontiers in Pharmacology, 13, 851246. https://doi.org/10.3389/fphar.2022.851246

Maciejewski, M., Lounkine, E., Whitebread, S., Farmer, P., DuMouchel, W., Shoichet, B. K., & Urban, L. (2017). Reverse translation of adverse event reports paves the way for de-risking preclinical off-targets. eLife, 6, e25818. https://doi.org/10.7554/eLife.25818

Mascolo, A., Scavone, C., Bertini, M., Brusco, S., Punzo, F., Pota, E., Di Pinto, D., Di Martino, M., Di Mauro, G., Capuano, A., & Rossi, F. (2020). Safety signals related to the use of immune checkpoint inhibitors. Frontiers in Oncology, 10, 826. https://doi.org/10.3389/fonc.2020.00826

Mockute, R., Desai, S., Perera, S., Assuncao, B., Danysz, K., Tetarenko, N., Gaddam, D., Abatemarco, D., Widdowson, M., & Beauchamp, S. (2019). Artificial intelligence within pharmacovigilance: A means to identify cognitive services and the framework for their validation. Pharmaceutical Medicine, 33(2), 109–120. https://doi.org/10.1007/s40290-019-00269-0

Mozzicato, P. (2009). MedDRA: An overview of the Medical Dictionary for Regulatory Activities. Pharmaceutical Medicine, 23(2), 65–75. https://doi.org/10.1007/BF03256752

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association, 18(4), 441–448. https://doi.org/10.1136/amiajnl-2011-000116

Norén, G. N., Hopstadius, J., & Bate, A. (2013). Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Statistical Methods in Medical Research, 22(1), 57–69. https://doi.org/10.1177/0962280211403604

Norén, G. N., Hopstadius, J., Bate, A., Star, K., & Edwards, I. R. (2010). Temporal pattern discovery in longitudinal electronic patient records. Data Mining and Knowledge Discovery, 20(3), 361–387. https://doi.org/10.1007/s10618-009-0152-3

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Ryan, P. B., Schuemie, M. J., Welebob, E., Duke, J., Valentine, S., & Hartzema, A. G. (2013). Defining a reference set to support methodological research in drug safety. Drug Safety, 36(S1), 33–47. https://doi.org/10.1007/s40264-013-0097-8

Sakaeda, T., Tamon, A., Kadoyama, K., & Okuno, Y. (2013). Data mining of the public version of the FDA Adverse Event Reporting System. International Journal of Medical Sciences, 10(7), 796–803. https://doi.org/10.7150/ijms.6048

Tatonetti, N. P., Ye, P. P., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and

interactions. Science Translational Medicine, 4(125), 125ra31. https://doi.org/10.1126/scitranslmed.3003377

Wang, Y., & Lu, X. (2023). Large language models in clinical pharmacovigilance: Opportunities and risks. Drug Safety, 46(11), 1023–1031. https://doi.org/10.1007/s40264-023-01346-9

Wong, A., Plasek, J. M., Montecalvo, S. P., & Zhou, L. (2018). Natural language processing and its implications for the future of medication safety: A narrative review. Pharmacotherapy, 38(8), 822–841. https://doi.org/10.1002/phar.2151

Zhu, Z., Yang, Y., Han, B., Tang, S., Liu, S., Hu, Q., Yan, S., Zhuang, Y., Sun, L., & He, K. (2020). adsLDA: An efficient topic modeling method to identify adverse drug reactions from social media. Information Sciences, 521, 232–245. https://doi.org/10.1016/j.ins.2020.02.045