Paper SD-01

Accelerating Patient Safety with Knowledge Graphs: Lessons from a Successful Implementation

Partha Bhattacharjee, Altair, Chelmsford, US
Kirti Aggarwal, Altair, Chelmsford, US
Patrick Jackson, Altair, Chelmsford, US

ABSTRACT

Three challenges stymie the ability to maximize value generation from patient safety systems. First, not all relevant data such as in-house safety reporting applications, drug labels, industry taxonomies, clinical trials, and real-world data are integrated. This ecosystem of data siloes spawns inefficient analysis and reporting processes. Second, well-compensated Subject Matter Experts (SMEs) spend more time data wrangling than analyzing it for scientific insights. Third, data engineered in such information systems is suboptimal input to Large Language Models (LLMs) due to structural constraints such as the inability to model complex relationships, semantically represent data, integrate semi- and unstructured data, and explicitly state relationships between entities. Knowledge graphs are an innovative approach for semantically integrating heterogeneous data for analysis, reporting, and input into LLMs. We describe a Knowledge Graph Patient Safety solution that accelerates analysis and reporting by up to 50%. We also highlight key takeaways from deployment in a real-world GxP-validated environment.

INTRODUCTION

Pharmacovigilance, the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem, is a critical aspect of patient safety. Patient safety solutions are used to monitor, analyze, and report adverse drug reactions throughout a product's lifecycle. In addition, such systems are used for signal detection and risk management to identify and investigate potential new safety concerns or changes in known safety issues related to medications.

Legacy IT standards and approaches using relational systems result in the creation of suboptimal patient safety analytical systems. Such systems do not enable patient safety scientists and physicians to efficiently identify risk factors, detect signals, design future clinical trials, monitor cases, perform epidemiological studies, and more. The data is often disjointed, messy, and hard to harmonize. Relational systems struggle to integrate structured data like safety databases and industry vocabularies with semi-structured and unstructured data sources, including drug labels, case narratives, and medical histories. The absence of robust data integration often compels safety experts to resort to brittle workflows characterized by substantial reliance on IT teams, manual operations, and spreadsheet-driven analyses.

Individual case safety reports (ICSRs) and other data sources critical to the pharmacovigilance process are stored in a variety of formats in diverse systems, including relational databases and warehouses, big data repositories, and spreadsheets. Additionally, unstructured sources such as drug labels, case narratives, medical histories, journal articles, research papers, and structured sources such as business vocabularies, including the Medical Dictionary for Regulatory Activities (MedDRA), WHODrug Global, EMA EudraVigilance, and the FDA Adverse Event Reporting System (FAERS) are important for analyses.

A knowledge graph-based solution blends data from different sources and semantically enriches them into consistent formats with high degrees of accuracy and reliability. Data integration from structured, semi-structured, and unstructured sources are better integrated than relational systems and, because the data is structured as a knowledge graph, users can utilize innovative analytical techniques such as network analysis to improve signal detection in a self-service environment. For example, a knowledge graph-based system can:

- Handle adverse event terms from different versions of the MedDRA coding system.
- Harmonize different names for the same drugs and dosage amounts using different units or formulations.
- Merge data from clinical trials, spontaneous reports, electronic health records, and databases containing information on drug and drug-protein interactions.
- Integrate safety data from global affiliates of a pharmaceutical company, each of which may use different measurement and reporting systems.

- Unify data from randomized controlled trials with observational studies. Combine structured data from clinical trials with unstructured data from literature reviews and social media.

OVERVIEW OF PATIENT SAFETY KNOWLEDGE GRAPH

A knowledge graph is a flexible, graph-structured information system that represents and organizes information about entities and their relationships. It consists of nodes and edges. The nodes represent entities or concepts (e.g., people, places, events, or abstract ideas) and the edges depict relationships between nodes.

Knowledge graphs are characterized by their ability to:

- Store semantic information, allowing both humans and machines to understand and reason about the contents
- Integrate data from multiple sources and domains
- Evolve and adapt to incorporate new information and changing relationships
- Support complex queries and derive new insights as well as allow reasoning and inferencing

These structures are widely used in various applications, including search engines, question-answering systems, recommendation engines, and AI-powered technologies. Knowledge graphs enable organizations to create a unified, contextual view of their data, facilitating better decision-making, improved data integration, and enhanced analytical capabilities

The following illustration describes the creation of a knowledge graph from structured, semi-, and unstructured data sources.

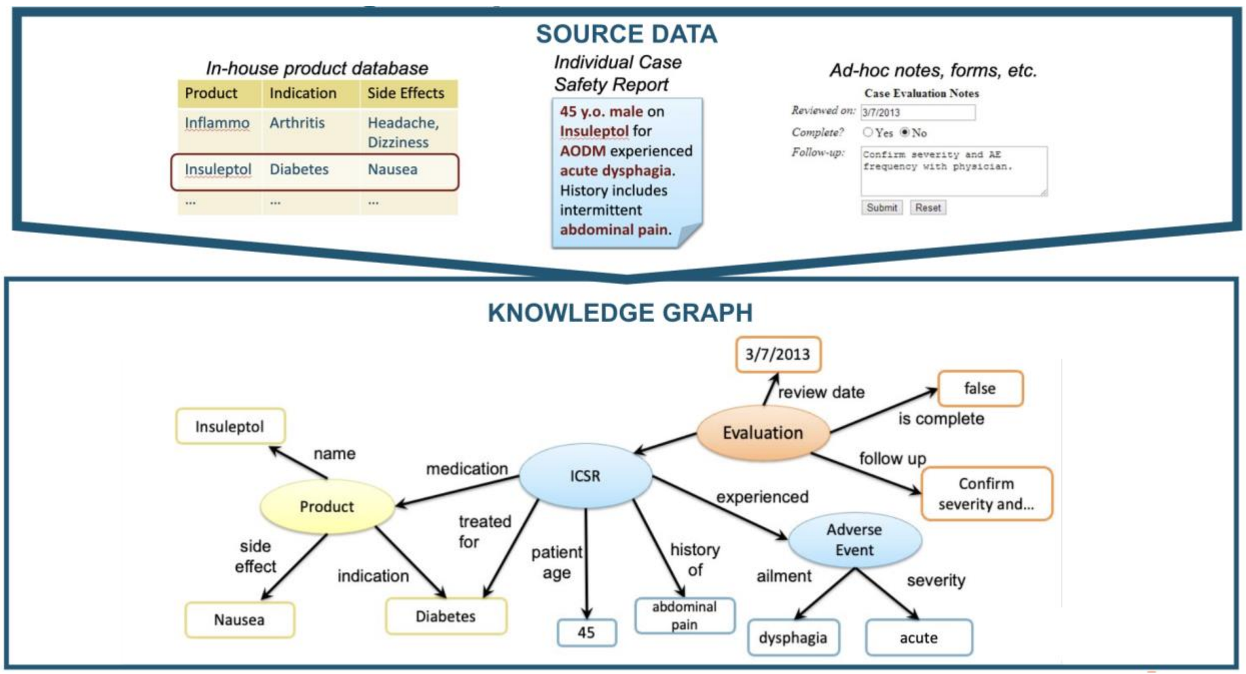

Figure 1. Overview of Knowledge Graph

The application of knowledge graphs to integrate diverse data in patient safety offers important benefits.

Safety experts can investigate adverse events in aggregate within the necessary scientific context:

- Patient demographics sourced from in-house and public safety databases.
- Drug information about suspect and concomitant products sourced from labels, industry standard taxonomies such as the WHO Drug Dictionary, and syndicated data.
- Adverse event details sourced from MedDRA.
- Important medical history information locked in unstructured case narratives.
- In-house "tribal knowledge" that may sprawl across presentations, spreadsheets, word processor files, emails, and databases.

Knowledge graphs lend themselves to exploratory analyses, statistical analyses, classical machine learning, graph and network science, and analyses driven by large language models (LLMs). Users can analyze data using low-code/no-code workflows as needed. Importantly, the explicit assertion of relationships between entities enables innovative approaches such as graph retrieval-augmented generation (RAG) and prompt-driven analysis. Safety experts can uncover relationships and identify risk factors between entities by "walking" the graph. They can spot trends and patterns that emerge only when they can analyze and visualize all relevant data.

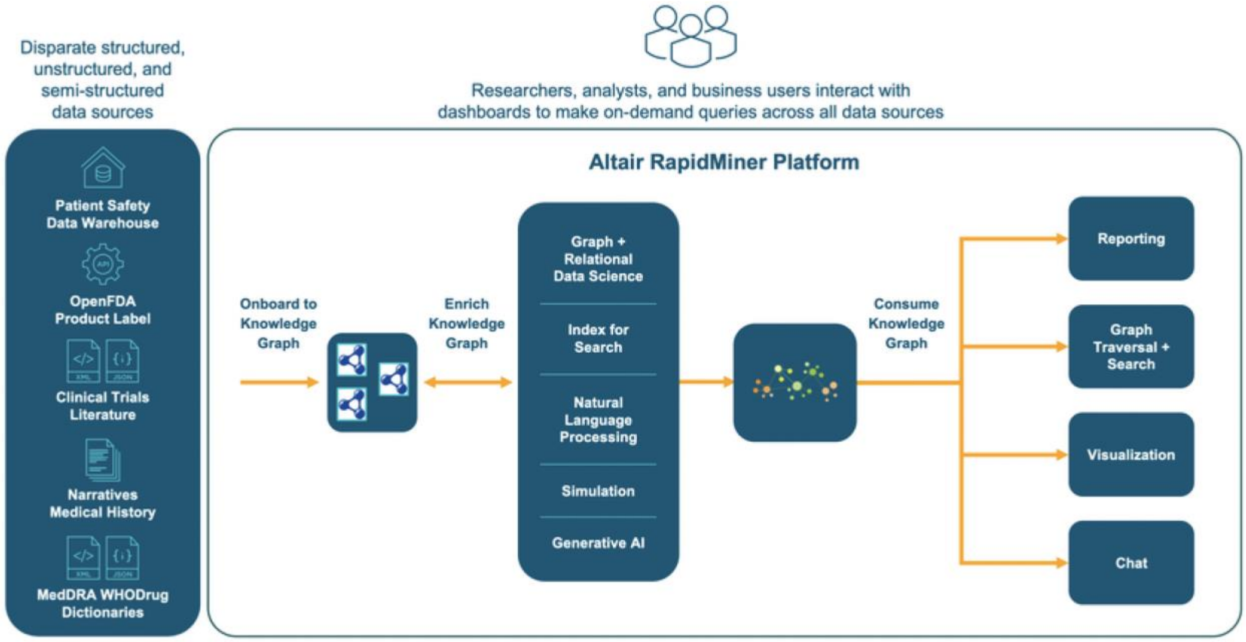

Figure 2. Data Flow

For example, safety experts can enhance efficiency by automatically navigating to relevant sections of annotated drug labels of concomitant and suspect products while investigating risk factors. Furthermore, experts can analyze risk factors by reviewing billions of data points to generate aggregate statistics based on a variety of user-defined criteria spanning connected drug, patient, adverse event, medical history, narrative, and drug label data. Estimates indicate that this solution can potentially halve the time required to analyze aggregate cases for risk factors and generate reports.

Traditional rules-based automation and AI/machine learning systems alone perform suboptimally on disparate data involved in pharmacovigilance applications. They also require a significant amount of manual work to maintain. Knowledge graphs support statistical approaches along with AI/machine learning. They also unlock avenues for other innovative approaches such as graph algorithms and network analysis.

IMPLEMENTATION DETAILS

The following visual represents the various data sources integrated as part of the solution.

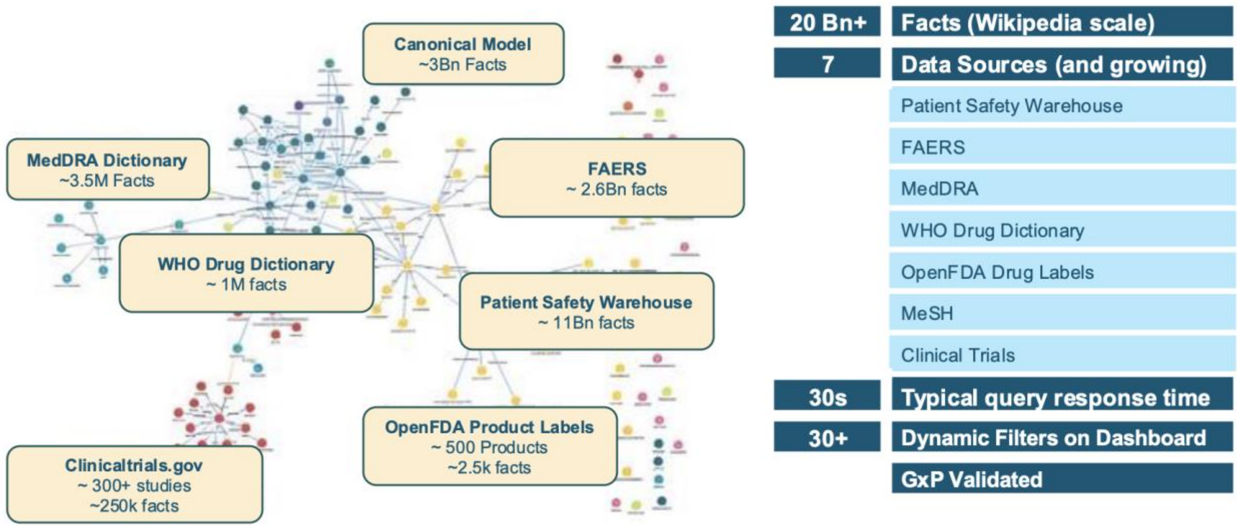

Figure 3. Data Sources

The unit of scale of a knowledge graph is a triple describing a subject-predicate-object relationship. For example, following are three triples about Tylenol describing the information in English as well as Resource Description Framework (RDF), the technical representation standard of information in a knowledge graph.

#	Fact - English	Fact - RDF
1	Tylenol is a Drug	<http://example.com/Tylenol> a <http://example.com/Drug>
2	Fever is a Disease	<http://example.com/Fever> a <http://example.com/Disease>
3	Tylenol treats fever	<http://example.com/Tylenol> <http://example.com/treats> <http://example.com/Fever>

For context, DBpedia, a collection of structured data from Wikipedia, has 21 billion triples[i].

The architecture of the solution is illustrated below. Fundamentally, there are three components of the platform using which the knowledge graph is built:

1. **Altair® Graph Studio™:** This server manages all the components of the solution and serves as a metadata catalog. It also has a distributed component for processing unstructured data. The component processes text fields for case narrative and medical history as well as documents such as product labels at scale to extract entities of interest that can then enrich the knowledge graph.

2. **Altair® Graph Lakehouse™:** A Massively Parallel Processing (MPP)-scale distributed graph warehouse to process complex analytical queries on large data volumes at speed and scale for optimal user experience.

3. **ElasticSearch:** A distributed open-source search and analytics engine which is used for indexing data from the knowledge graph to facilitate search + graph traversal data retrieval workflows for users.

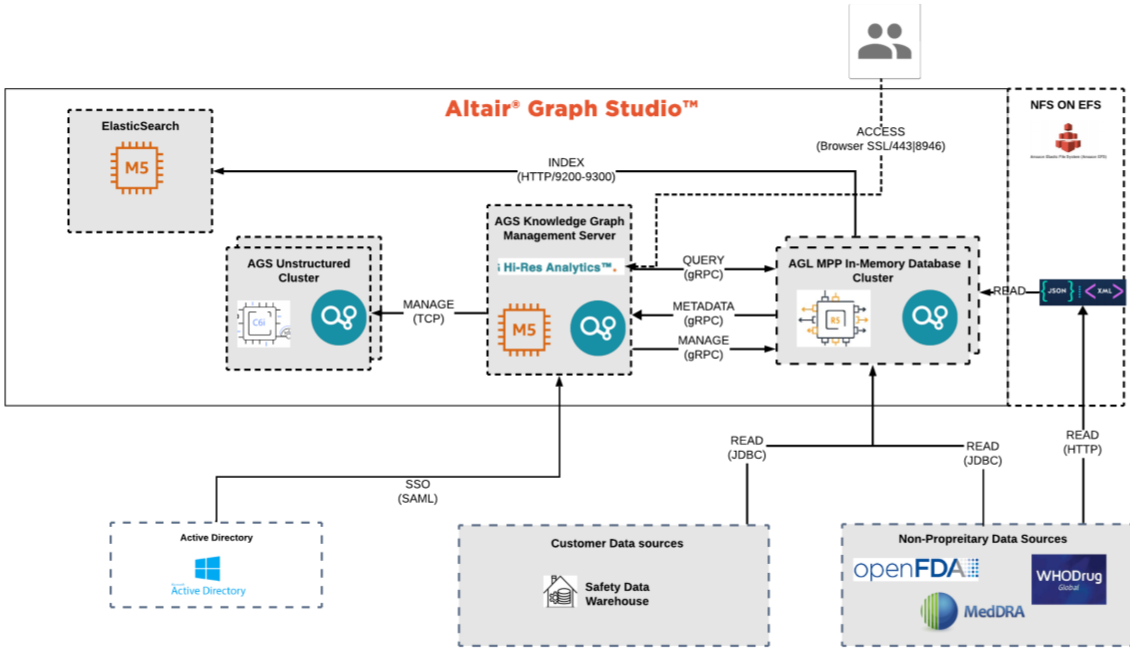

Figure 4. Solution Architecture

These components function synergistically to onboard data from the above-mentioned structured and unstructured sources, integrate them using shared business-friendly concepts, and make the connected data available for self-service consumption. The typical user experience is illustrated below:

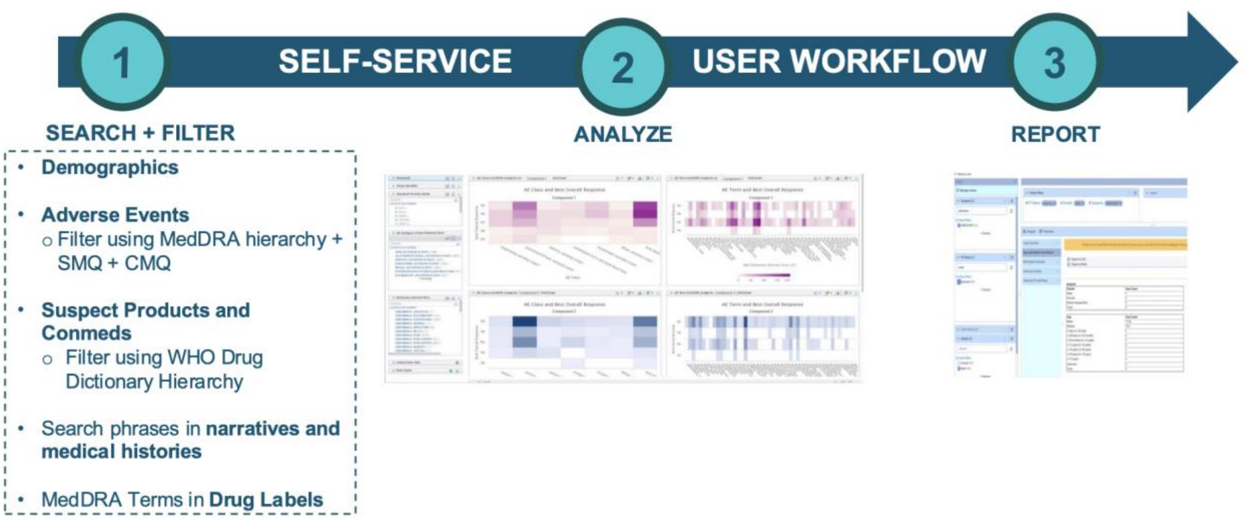

Figure 5. User Workflow

Following is a representative user query for shortlisting cases

REPRESENTATIVE QUERY

Find all cases meeting the following criteria (on a record set of 10 Million cases):

1. Taking drug X
2. Experiencing any of the adverse events characterized it as part of SMQ Z in MedDRA
3. Age Group = Adult
4. Gender = Male
5. Only the latest version of the case
6. Between 01/01/2010 and 08/31/2024
7. Case Approval Status = Approved
8. Only valid ICSRs
9. Meeting certain Product Flag
10. Exclude Nullfiied/Deleted Case

Queries typically get executed in 30 seconds using MPP-scale architecture of Altair Graph Lakehouse

11. Meets Serious/non-serious event
    11.1 If serious then which criteria:
    - Hospitalization required
    - Death
    - Life threatening
12. Country
13. Dechallenge/Rechallenge
14. Time to Onset of event
15. Indication

Figure 6. Representative User Query

LESSONS LEARNED

ENHANCED RISK FACTOR ANALYSIS IS A TRUE DIFFERENTIATOR

A key benefit of the Patient Safety Knowledge Graph solution is the ability of safety experts to analyze risk factors and view attendant case listings at speed and scale. Following is an illustration of search and filter executed for Risk Factor analysis

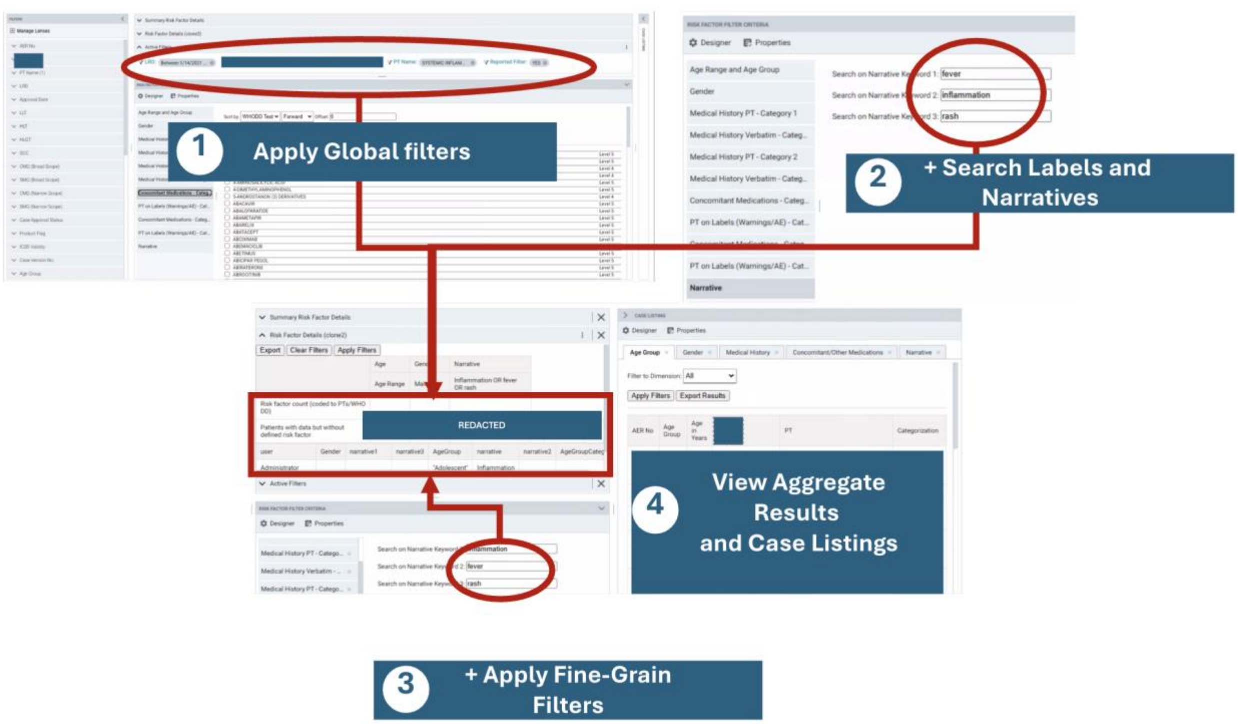

Figure 7. Risk Factor Analysis Workflow

These criteria are applied across billions of data points with results typically rendered in under two minutes. This speed and scale offers significant improvement over legacy workflows for a couple of reasons:

1. This search offers innovative capabilities for combining searches across unstructured documents such as drug labels and case narratives with filters across other structured data

2. The query performance is significantly faster as output are rendered in seconds and minutes instead of dozens of minutes or hours of manual effort

Such exhaustive filter, search, and summarization are made possible due to the connected nature of the data, the performance of the underlying database, and the ability to integrate structured and unstructured data using knowledge graphs.

MAPPING TECHNICAL ENABLERS TO FUNCTIONAL ACCELERATORS DRIVES BUY-IN

While knowledge graphs offer multiple benefits and Altair® Graph Studio™ is a broad platform, we distilled a few key technical enablers for functional accelerators that drive positive business outcome as summarized below.

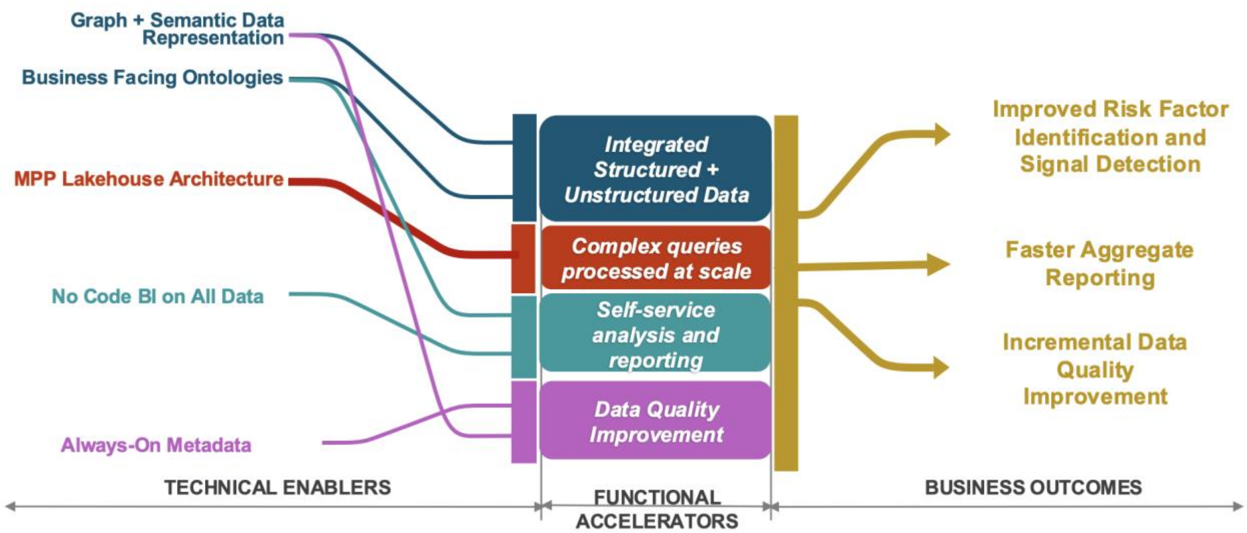

Figure 8. Technical Enablers and Functional Accelerators

The representation of domain knowledge as a knowledge graph using semantic data standards not only facilitated the integration of structured and unstructured data but also led to data quality improvements. The improvement in data quality is driven by three factors. First, data discrepancies become apparent when users avail self-service access mechanisms to evaluate all data at speed and scale. Second, the integration of diverse data sources uncovers data discrepancies that might otherwise not be discovered. Third, the utilization of metadata such as schema and table information combined with flexibly connecting data in different patterns aids discovery of data gaps and issues.

The technical capabilities of the product, such as MPP-scale performance on complex queries across large data volumes and no-code Business Intelligence (BI) access on all data, helps users analyze data at unprecedented scale and flexibility. The technical enablers, complemented by the fact that a wide swath of interconnected data is made available, allow users to identify and analyze signals more efficiently. In addition, the capabilities make aggregate reporting faster and more accurate by eliminating manual processes.

ENGAGING SAFETY EXPERTS EARLY IN DEVELOPMENT DRIVES INNOVATION

While it is considered good practice in technical development to engage users early for ensuring accurate requirements gathering and regular feedback during development, we found that engaging them also helps drive innovation. For example, a critical challenge to developing the solution was reconciling drugs from different sources---customer's patient safety warehouse, OpenFDA, and WHO Drug Dictionary---to their labels sourced from OpenFDA. The SMEs helped us largely solve this challenge by generating heuristics for mapping drugs to labels. Their feedback also shaped iterative dashboard improvements, reducing training overhead.

Another major advancement that benefited from extensive user engagement was designing the Risk Factor analysis workflow and dashboard. Experts' insights into usage patterns, current pain points, and desired analytical capabilities informed the design and development of this feature. Such inputs, completed by product capabilities, have resulted in the development of a novel innovative capability for the customer.

A LITTLE NATURAL LANGUAGE PROCESSING (NLP) GOES A LONG WAY

The ability to implement workflows where users can easily view unstructured data such as case narratives and relevant sections of product labels, with annotations for entities of interest such as MedDRA terms in the context of their standard workflow in the solution is a major benefit. Such operations are often performed using web searches in legacy processes.

A signal analysis workflow often entails a safety expert needing to review labels of dozens of drug labels for concomitant and suspect products.

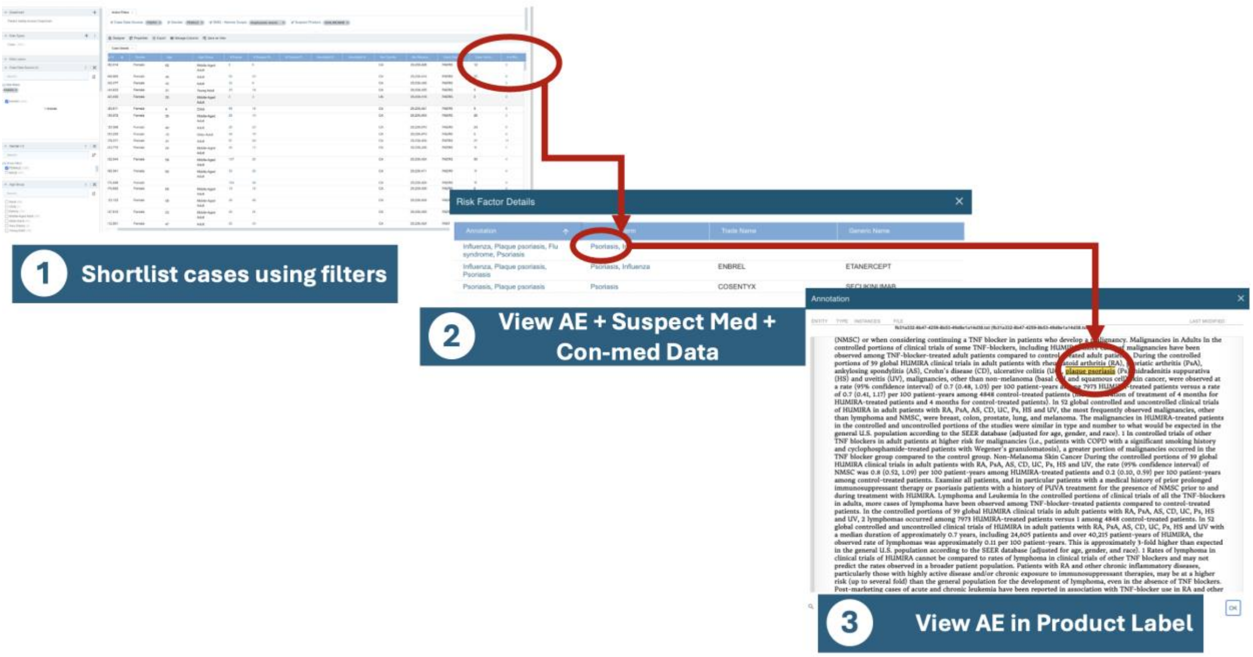

Figure 9. Text Analytics Integration

While innovations such as Large Language Models (LLMs) offer deep NLP capabilities, our experience with this implementation provides evidence that deterministic approaches for text matching backed by context in the form of a knowledge graph, combined with user-friendly user experience, offer significant value to users.

BUILDING WITH GxP-VALIDATION INTENT FROM OUTSET IS EFFICIENT

The sensitive operations performed by the solution necessitated its GxP-validation. One of our lessons learned is to be more intentional about such validation from the outset and plan for it. In particular, it is important to develop shared understanding of the validation perimeter for the solution and design operations to meet the Installation-, Performance- and Operational Qualification requirements. The ability to rely on Out of the Box capabilities of the product is a key accelerant during the validation effort.

SIMPLE BUSINESS-FACING ONTOLOGIES DRIVE VALUE

Ontologies are the knowledge representation models to which data are conformed to create a knowledge graph. A business-friendly ontology—avoiding overly technical constructs—enables SMEs to explore data independently. An illustration of a simple ontology is below.

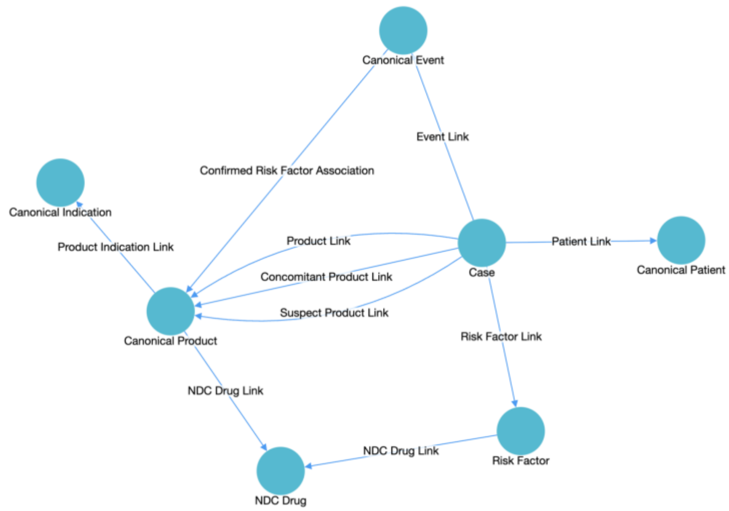

Figure 10. Representative Data Model

In addition to facilitating self-service access by users, business-facing ontologies with readable entity and relationship labels also help the development of GenAI-based applications such as Question Answering systems. Such ontologies are both machine- and human- readable. Large Language Models (LLMs), trained on natural language text, are adept at processing such ontologies to answer user questions. Altair® Copilot, an AI-powered support assistant that understands natural language and can answer questions, is a manifestation of this principle.

DATA QUALITY IMPROVEMENTS ARE AN EMERGENT PROPERTY OF KNOWLEDGE GRAPH SYSTEMS

Knowledge graphs ensure data always exists with its context i.e., metadata. Such metadata are modeled using ontologies. The solution makes such data available for ontology-driven self-service analysis to not only business users but also data engineers. The solution offers data engineers the ability to execute automated tests over three-hundred thousand data points daily after loading incremental values. These tests, performed using self-service dashboards, help support teams quickly identify data discrepancies, mismatches, and irregular patterns.

Such ability to profile and analyze data at scale is typical of knowledge graph solutions built using the Altair® Graph Studio™ platform. While knowledge graph implementations are typically directed towards solving business problems, the organic improvements in data quality through easy exposure of issues and flexible mechanisms of resolution are a valuable emergent property of the system. Proactively highlighting improvements has built stakeholder trust in the solution.

LEGACY SYSTEMS ARE FALLIBLE

The development of the solution entailed an involved process of comparing the data and output from it to legacy systems. The process of resolving such mismatches surfaced gaps in legacy systems and workflows. Such gaps were largely attributable to undetected data discrepancies and manual processes. Transparent root-cause analyses not only improved such processes but also strengthened the case for data integration using knowledge graphs for this use case as it helps drastically minimize manual processes and surfaces data quality issues.

FUTURE WORK

Future work on the solution can be classified into three categories – Operational, Functional, and Innovation. Details of these categories are described below.

OPERATIONAL

The current solution comprises data modeling using ontologies, data engineering pipelines using SPARQL, and data access using dashboards. Future work will entail improvements to these components to enhance user experience. A key requirement from users is to be able to use the WHO Drug Dictionary as a hierarchy of terms in a selectable filter when analyzing risk factors. The current configuration makes such terms available as a flat list of terms. Other operational efforts will include optimization of input data volume, continual data quality assessment, and review of existing integrations for improving performance of complex cross-modal queries.

FUNCTIONAL

The availability of integrated data, combined with the codeless self-service access capabilities of the platform, makes addressing new use cases seamless. Given the baseline of a knowledge graph, future work will entail quickly assembling dashboards to address other use cases such as monitoring cases in specific geographies and comparing safety signals for drug-event combinations of interest between clinical trial and real-world data. Additionally, efforts are underway to integrate other data to address use cases in related domains such as Medical Affairs.

INNOVATION

In addition to operational and functional use cases, an R&D track aims to design new solutions by combining knowledge graphs with innovations in AI. Some examples of efforts being considered in this category include:

- **Conversational BI:** Setting up a Question Answering system for patient safety using Altair® Copilot, an AI-powered support assistant that understands natural language and can answer questions based exclusively on data of interest without hallucinations
- **AI-Driven Patient Journeys**: Extracting temporal events (e.g., *hospitalization*, *dose change*) from unstructured narratives to build longitudinal patient timelines using knowledge graphs and AI
- **Real-World Evidence (RWE) Integration**: Incorporating claims data to assess post-market safety signals.
- **Predictive Analytics:** Predict drug-drug and drug-protein interactions by analyzing a “bionetwork” of connected entities drawn from the knowledge graph
- **Search:** Apply techniques such as Graph RAG and embeddings to improve search and context-aware conversational Business Intelligence

CONCLUSION

In conclusion, the Patient Safety Knowledge Graph is an industry-leading solution that accelerates patient safety workflows. It integrates structured and unstructured data to improve case analysis, offers unprecedented comprehensiveness in signal and risk factor analysis, and automates reporting. Its integrated data can further be connected to other sources for novel analytical capabilities. Finally, such integrated data forms an efficient input for AI models to drive use cases such as conversational BI, sophisticated document searches, and predictive analytics.

REFERENCES

1. DBpedia Release Dashboard. release-dashboard.dbpedia.org. Retrieved February 9, 2025, from https://release-dashboard.dbpedia.org/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Partha Bhattacharjee
Altair, 300 Apollo Dr., Chelmsford, MA 01824
+1 (248) 614-2400
partha@altair.com
https://altair.com/healthcare-life-sciences