Unlocking Biopharma Insights: A RAG-Enabled AI Framework for Rapid Access to Clinical and Nonclinical Study Data

Dr Sapna Chandran L, PointCross, Bengaluru, India

Lakshmi B K, PointCross, Bengaluru, India

ABSTRACT

This paper presents a Retrieval-Augmented Generation (RAG) framework designed to revolutionize access to biopharmaceutical study data. The system addresses critical challenges in drug development by enabling rapid, contextual data retrieval through natural language queries. Leveraging a Unified Data Model (UDM) repository, the system of record harmonizes clinical and nonclinical study data across all development phases, ensuring consistent indexing and analysis. Architecture employs securely embedded Large Language Models (LLMs) within pharmaceutical sponsors' networks to enable intuitive, chat-based interactions, thereby avoiding complex user interfaces. LLM-triggered agents facilitate automated data curation, standardization, visualization, and statistical output generation while minimizing the risk of AI hallucinations. The system serves diverse stakeholders, including data managers, researchers, and regulatory teams, by providing immediate access to required information. By addressing data sovereignty concerns and reducing platform complexity barriers, this RAG-enabled framework significantly enhances study data accessibility and utility, ultimately accelerating insight generation and value creation throughout the drug development lifecycle.

1. INTRODUCTION AND SCOPE

Artificial Intelligence (AI) and Large Language Models (LLMs) are being rapidly adopted across industries, and biopharma is no exception. These technologies are transforming how organizations access, interpret, and utilize data, streamlining development processes and enabling faster, more informed decisions as timelines continue to accelerate. Across pharmaceutical sponsors, biotechnology companies, and contract research organizations (CROs), the question is no longer whether AI should be adopted, but how it can be deployed safely, efficiently, and predictably within tightly governed, SOP-driven environments.

Early adoption of AI in biopharma has primarily focused on discovery-phase use cases such as high-throughput assays, microarray analysis automation, and transactional workflows including ticketing, documentation, and administrative coordination. While these use cases are relatively straightforward to prototype, they present significant risks when scaled. As autonomous, interconnected workflows increase to hundreds or thousands, comprehensive simulation becomes impractical. Even minor deviations can lead to SOP breaches, an unacceptable outcome in a highly regulated industry. These limitations have prompted organizations to reassess how AI can be responsibly deployed beyond early-stage prototypes.

The lifecycle of in-vivo studies and clinical trials in drug development poses a range of additional constraints. Ethical, legal, and regulatory restrictions are expected as various regions of the world enact laws to govern the expansion of AI. The absolute secrecy, security, and proprietary ownership of this expensive data mean that none of the study or clinical trial data can be applied on open public AI infrastructure. The remote possibility that the queries and prompts provided may leak proprietary information is a risk that stands in the way of quick adoption.

Biopharmaceutical sponsors increasingly maintain large, centralized repositories containing comprehensive clinical and nonclinical data assets. These include as-collected study and trial data, intermediate work-in-progress files on the path to standardized SDTM and ADaM datasets, and insights captured in TFLs for Clinical Study Reports (CSRs). Additionally, Trial Master File (TMF) documents, including protocols, Statistical Analysis Plans (SAPs), CSRs, monitoring plans, and other regulatory artifacts, are stored alongside these datasets.

While this wealth of information represents a significant investment, its practical utility is often constrained. Today, access to these assets is mediated through multiple specialized applications, each designed for a specific data domain. Users, whether data managers, statisticians, clinicians, or regulatory professionals, must possess deep familiarity with these systems to locate, analyze, and extract the information they need. This requirement is challenging for organizations facing limited staff capacity, high turnover, and evolving IT landscapes. Consequently, adoption suffers, and the full value of these expensive data assets remains unrealized.

This paper outlines the implementation of a RAG-enabled conversational AI interface, designed to transform how biopharmaceutical organizations access and analyze clinical and nonclinical study data. The solution integrates natural language interaction with governed data retrieval and advanced signal detection to support more timely and informed decision-making across the drug development lifecycle.

By combining the reasoning capabilities of large language models with secure, validated repository access, the RAG architecture enables rapid, traceable insight generation within controlled environments, without compromising compliance, data integrity, or operational trust. Importantly, this approach is not positioned as a "magic switch," but as a deliberately governed capability built on existing standards, transparent workflows, and phased adoption, intended to reduce navigation friction, encourage user adoption, and unlock enterprise data value responsibly.

The purpose of this paper is threefold:

→ Present a scalable approach to reducing friction in study data access, enabling rapid, compliant, and context-aware retrieval without compromising enterprise security or governance.

→ Demonstrate how RAG-enabled AI elevates productivity across stakeholder groups, including data managers, scientists, statisticians, clinical programmers, and regulators, through conversational, provenance-anchored insights and seamless navigation across analytical applications.

→ Outline a strategic roadmap for responsible AI adoption, ensuring innovation is matched with integrity, transparency, auditability, and long-term organizational readiness.

2. CURRENT CHALLENGE

The industry is now moving toward enterprise-scale applications, including protocol interpretation, predictive safety monitoring, automated data analysis, study report generation, and agent-driven scientific and regulatory workflows. Amid rapid technological evolution, biopharmaceutical organizations continue to face a fundamental operational challenge: consistent, efficient access to data and actionable insights across the drug development lifecycle.

- **Data volume and heterogeneity**
  Modern drug development generates large, diverse datasets spanning multi-omics data (e.g., cell phenotyping, cytokines, ADA), clinical, nonclinical, and real-world evidence, which are difficult to access cohesively using traditional query interfaces.

- **Platform complexity and limited usability.**
  Enterprise systems expose feature-dense interfaces with deep navigation paths, excessive controls, and steep learning curves, leading most users to access only a small fraction of available functionality.

- **Fragmented data and context**
  Clinical, non-clinical, biomarker, SDTM/ADaM, TLF, and TMF content remains siloed across disconnected platforms, limiting unified interpretation of safety, efficacy, and exploratory endpoints.

- **Portfolio-wide analytical demand**
  Strategic decisions increasingly require cross-study and cross-indication analysis, yet portfolio-level questions often require extensive data engineering and manual coordination, delaying insight generation.

- **Workforce and role constraints**
  Outsourcing to CROs and FSPs in large pharmaceutical companies, along with lean operating models in biotech, reduces direct data access for experienced in-house users, forcing scientists, statisticians, and decision-makers to rely on intermediaries for critical data.

- **Governance, security, and traceability requirements**
  GxP expectations, data sovereignty constraints, and regulatory inspection needs require AI-enabled access to remain fully permissioned, explainable, auditable, and confined to controlled sponsor environments.

- **Operational risk at scale**
  Highly interconnected SOP-governed workflows, uncertain return on investment, and loss of institutional knowledge limit the safe deployment of autonomous or agent-driven solutions across enterprise operations**.**

Together, these factors create significant friction for users and limit the industry's ability to deploy AI responsibly.

3. THE VISION: AGENTIC CONVERSATIONAL RAG ARCHITECTURE ENABLING SEAMLESS, GOVERNED DATA ACCESS

A RAG-enabled AI chat interface changes the data access paradigm by shifting users away from siloed, complex platforms toward seamless, conversational interaction anchored in validated repositories. The following sections detail a three-phase implementation of the RAG-enabled conversational AI interface designed to modernize access to and analysis of study data through a modular Conversational RAG architecture that supports intent-driven retrieval, governs access to enterprise data assets, and generates context-aware responses while preserving regulatory compliance and traceability.(Fig.1)

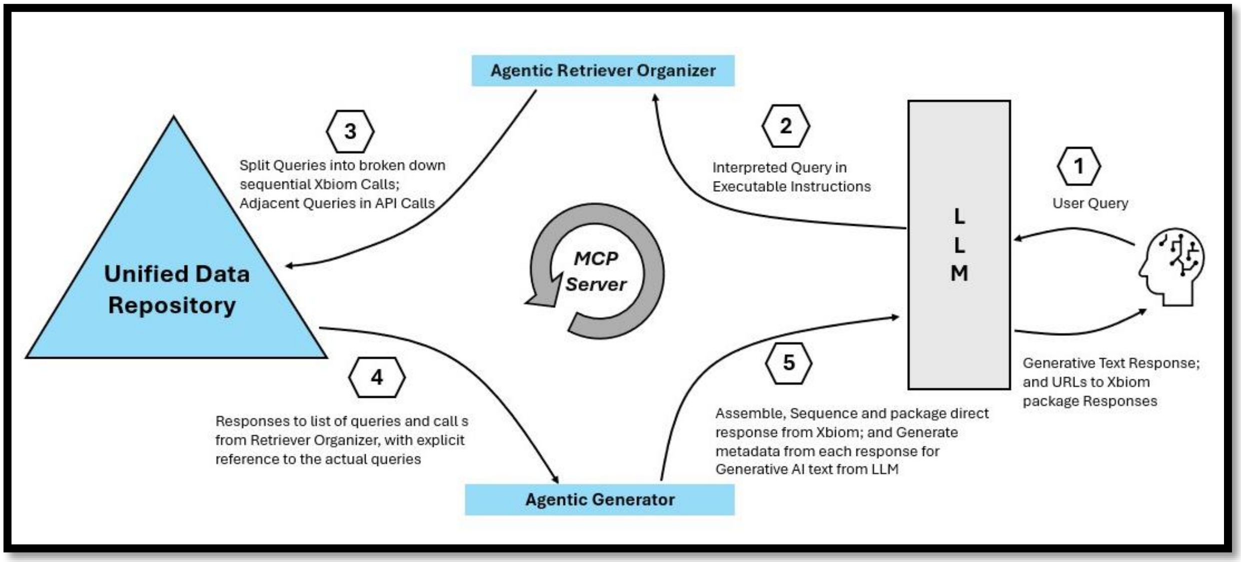

Agentic Retriever Organizer
3
Split Queries into broken down sequential Xbiom Calls; Adjacent Queries in API Calls
2
Interpreted Query in Executable Instructions
1
User Query
L L M
Unified Data Repository
MCP Server
4
Responses to list of queries and call s from Retriever Organizer, with explicit reference to the actual queries
5
Assemble, Sequence and package direct response from Xbiom; and Generate metadata from each response for Generative AI text from LLM
Generative Text Response; and URLs to Xbiom package Responses
Agentic Generator

Figure 1. LLM–RAG–System Interaction Architecture

3.1. CORE COMPONENTS

The Conversational RAG system comprises four primary components, each addressing a distinct stage of the interaction lifecycle.

1. **Large language model**

   The LLM component provides the natural language interface and performs intent interpretation. It converts unstructured user input into a structured semantic representation and captures the analytical intent, domain context, and constraints required for downstream processing.

2. **Agentic retriever organizer**

   The Agentic Retriever Organizer functions as the core engine for query decomposition and retrieval planning. It translates interpreted intent into a sequence of system-supported retrieval actions, ensuring that all data access remains authorized, traceable, and aligned with underlying data capabilities.

3. **Unified repository**

   The Unified Repository serves as the centralized, governed source of study data, including structured datasets, metadata, and derived analytical outputs. It responds exclusively to explicit, structured requests, ensuring data integrity and compliance.

4. **Agentic generator**

   The Agentic Generator assembles retrieved data into contextualized, human-readable responses. It synthesizes results while preserving data lineage, source references, compliance metadata and maintains conversational context across iterative interactions.

3.2 RAG THREE-LAYER ARCHITECTURE

The RAG three-layer architecture is an intent-driven layered model designed for rigor, traceability, transparency, and usability. RAG agents bridge the layers, interpreting user intent, generating validated API calls, and returning results as narrative summaries to end users. (Fig.2)

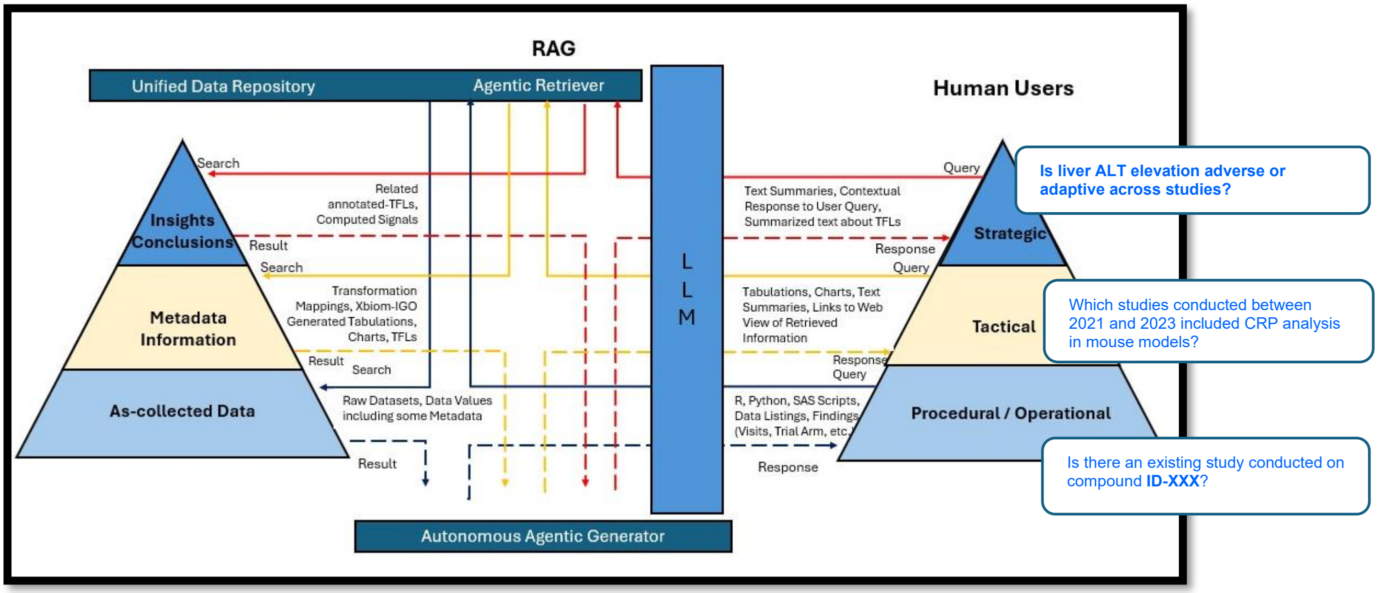

RAG
Unified Data Repository
Agentic Retriever
Human Users
Search
Related annotated-TFLs, Computed Signals
Insights Conclusions
Result
Search
Transformation Mappings, Xbiom-IGO Generated Tabulations, Charts, TFLs
Metadata Information
Result
Search
Raw Datasets, Data Values including some Metadata
As-collected Data
Result
L L M
Query
Text Summaries, Contextual Response to User Query, Summarized text about TFLs
Strategic
Response
Query
Tabulations, Charts, Text Summaries, Links to Web View of Retrieved Information
Tactical
Response
Query
R, Python, SAS Scripts, Data Listings, Findings (Visits, Trial Arm, etc.)
Procedural / Operational
Response
Autonomous Agentic Generator
Is liver ALT elevation adverse or adaptive across studies?
Which studies conducted between 2021 and 2023 included CRP analysis in mouse models?
Is there an existing study conducted on compound ID-XXX?

Figure 2. Intent Driven Retrieval with Augmented LLM Generative Response

Hierarchical data organization

Three-Layer Architecture organizes content across three maturity levels that align with different user inquiry types:

1. Procedural and Operational (Data layer / Foundation)

- The Data Layer intent and decisions are grounded in raw data in its most primitive form. This includes as-collected patient (21 CFR Part 11–compliant) or subject data, as well as assay-level data such as biomarker measurements or omics sequencing outputs, which serve as the foundational evidence for downstream analysis and decision-making.
- Data are retained in their original structure, provenance, and context to meet the needs of Clinical Data Managers, data governance, and regulatory stakeholders, ensuring auditability, reproducibility, inspection readiness, and support for downstream validation.

2. Tactical (Information layer / Curation)

- The Information Layer consists of curated, standardized, and analysis-ready datasets produced through controlled transformation pipelines. This layer integrates study metadata, applies standard data models and terminologies, and generates derived datasets suitable for statistical analysis and visualization.

- Designed to support tactical queries, this layer is primarily used by data managers, statisticians, and clinicians who require reliable, interpretable outputs such as tabulations, visual summaries, and mapped terminologies. By abstracting raw complexity, the information layer ensures consistency, comparability, and efficient reuse across analyses and studies.

3. Strategic (Knowledge layer / Insights)

- The Knowledge Layer captures synthesized insights and evidence, including integrated analyses, TFLs, and study reports, representing the highest level of abstraction where curated data are translated into interpretable scientific conclusions.

- It supports exploratory, hypothesis-driven inquiry by scientists, clinicians, pharmacologists, and toxicologists, enabling cross-study signal detection and integrated efficacy and safety interpretation, with outputs suitable for regulatory, scientific, and executive decision-making.

At the core of this architecture is an agentic RAG system that functions as an intelligent translation layer between human inquiry and enterprise data systems. Rather than exposing users to application-specific query languages or workflow logic, this layer interprets user intent, resolves ambiguities, and maps requests to validate system capabilities. It orchestrates governed interactions with underlying data repositories, ensuring that every response is reproducible, auditable, and role-appropriate. By aligning a hierarchical user-intent model with a multi-layered data and application ecosystem, the architecture enables secure, provenance-anchored access across varying levels of analytical complexity. The result is accelerated insight generation delivered with rigor, transparency, and trust.

RAG extends the static "memory" of a large language model by actively connecting it to a unified enterprise data store. It serves as a bridge between the LLM's language understanding and the organization's data-rich repositories, enabling the retrieval of datasets, metadata, and derived insights, as well as the generation of combined outputs in response to user queries.

3.3 RAG DUAL-MODE QUERY PROCESSING

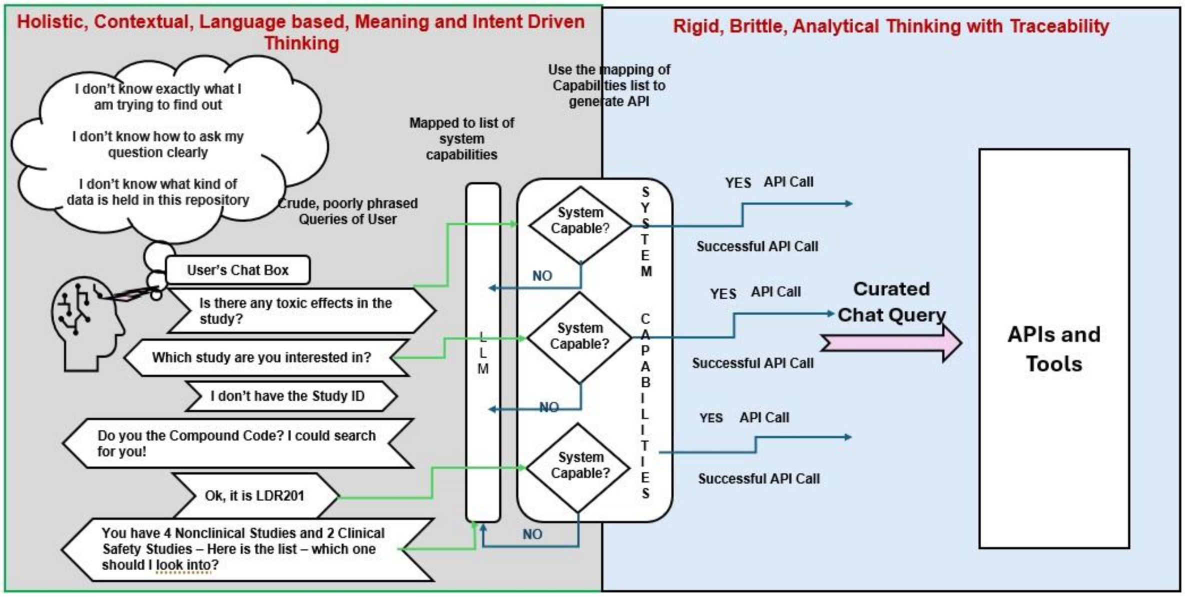

Holistic, Contextual, Language based, Meaning and Intent Driven Thinking
Rigid, Brittle, Analytical Thinking with Traceability
I don't know exactly what I am trying to find out
I don't know how to ask my question clearly
I don't know what kind of data is held in this repository
Crude, poorly phrased Queries of User
Mapped to list of system capabilities
Use the mapping of Capabilities list to generate API
User's Chat Box
Is there any toxic effects in the study?
Which study are you interested in?
I don't have the Study ID
Do you the Compound Code? I could search for you!
Ok, it is LDR201
You have 4 Nonclinical Studies and 2 Clinical Safety Studies – Here is the list – which one should I look into?
LLM
System Capable?
NO
System Capable?
NO
System Capable?
NO
SYSTEM CAPABILITIES
YES API Call
Successful API Call
YES API Call
Successful API Call
YES API Call
Successful API Call
Curated Chat Query
APIs and Tools

Figure 3. Holistic vs Analytical Query Processing

RAG query processing operates in two complementary modes to support both exploratory and regulated analytical use cases: **Holistic Processing** and **Analytical Processing**. This dual-mode approach ensures flexibility for natural language interaction while maintaining traceability and control for structured analysis. (Fig.3)

Holistic processing

Holistic processing addresses contextual, language-based, and meaning-driven queries using natural language understanding and semantic retrieval. It is designed to manage ambiguous or underspecified questions, domain-level inquiries, queries lacking explicit study identifiers, and complex relationship-based questions that require contextual interpretation rather than direct data extraction. This mode leverages retrieval augmentation to assemble relevant context across data layers before generating coherent, narrative responses.

Analytical processing

Analytical processing supports rigid, traceable, and reproducible analytical queries through structured execution pathways. Queries are translated into explicit API calls with precise capability mapping, sequential API orchestration, and curated query workflows. This mode enables direct access to validated data services, ensuring deterministic results, auditability, and alignment with regulated analytical requirements.

3.4 QUERY CLASSIFICATION FRAMEWORK

The system categorizes every user query into four distinct categories, each with specific handling protocols:

Category a: Out-Of-Scope Query

- Response Type: Out-of-Scope
- Acknowledges query, explains system purpose, and provides valid question examples.

Category b: Plausible but currently Unsupported Query

B1 - planned features

- Response Type: Planned Feature.
- Acknowledges request, states capability coming soon, suggests alternatives.

B2 - genuinely unsupported

- Response Type: Unsupported Feature.
- Acknowledges relevance, states not supported, records feedback.

Category c: System Capabilities and Services

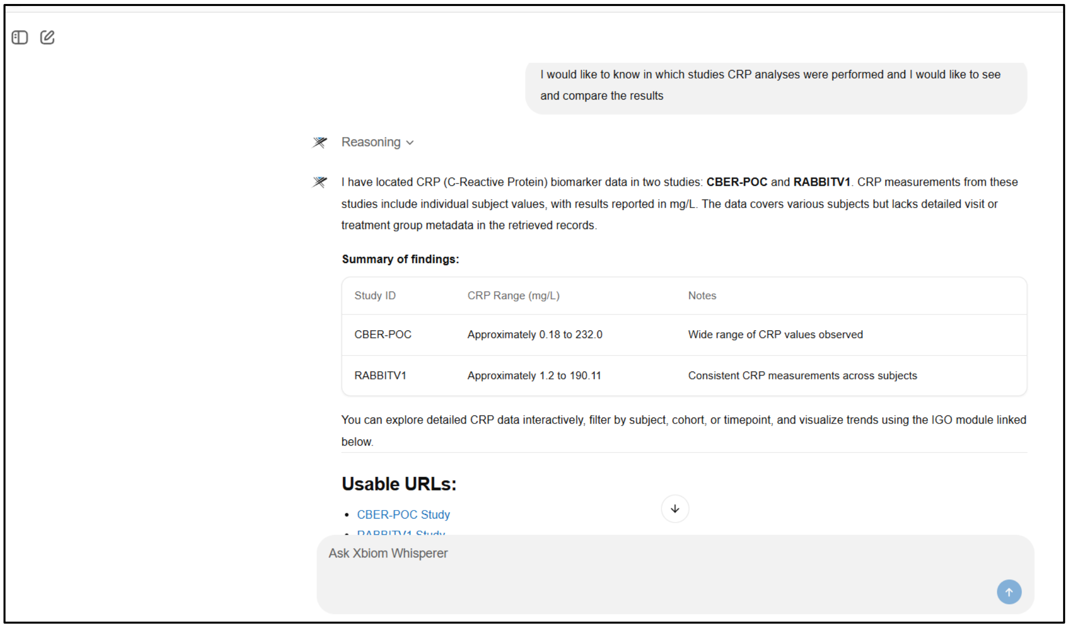
- Response Type: Document Response.
- Generic queries about platform, business solutions, security, regulatory compliance, and Data Standardization.
- Answered from FAQs and documented capabilities using retrieval-augmented generation over document content.

Category d: Supported Query requiring RAG Tool

Queries that can be answered using tools require systematic execution plans. (Fig.4,Fig.5)

Response protocol:

- **RESPONSE_TYPE:** Supported Tool Query.
- **TOOLS:** Semi-colon-separated list of required tools and entities.
- **GOAL:** One sentence summary of user intent.
- **PLAN:** Numbered, step-by-step action list with intermediate user feedback points.
- **URL:** Application URL with routing application URLs.



I would like to know in which studies CRP analyses were performed and I would like to see and compare the results

Reasoning

I have located CRP (C-Reactive Protein) biomarker data in two studies: **CBER-POC** and **RABBITV1**. CRP measurements from these studies include individual subject values, with results reported in mg/L. The data covers various subjects but lacks detailed visit or treatment group metadata in the retrieved records.

Summary of findings:

Study ID	CRP Range (mg/L)	Notes
CBER-POC	Approximately 0.18 to 232.0	Wide range of CRP values observed
RABBITV1	Approximately 1.2 to 190.11	Consistent CRP measurements across subjects

You can explore detailed CRP data interactively, filter by subject, cohort, or timepoint, and visualize trends using the IGO module linked below.

Usable URLs:

- CBER-POC Study
- RABBITV1 Study

Ask Xbiom Whisperer

Figure 4. Chat agent

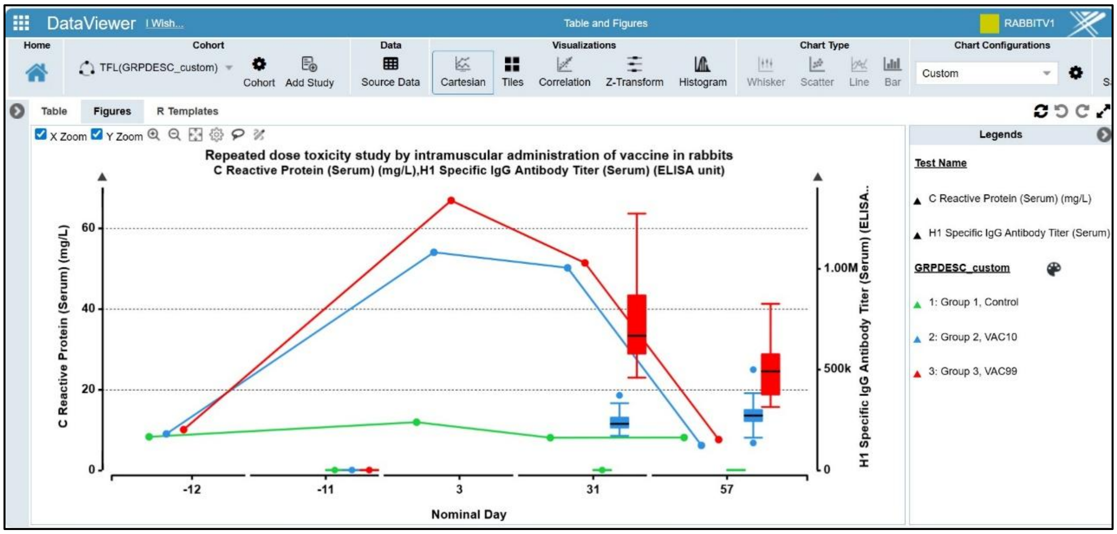


Figure 5. Data Visualization

3.5 DECISION LOGIC FLOW

The system employs intelligent routing logic with multiple decision nodes:

1. **Scope Assessment:** Check if the query is within the application's defined domain.

2. **Feature Availability:** Verify the requested capability is supported and implemented.

3. **Information Type Classification:** Determine whether the query requires data-driven evidence versus static system information.

4. **Tool Requirement Determination:** Identify whether RAG tool execution is required to answer the query.

5. **Topic Alignment Validation:** Confirm the query aligns with supported application-specific topics.

6. **Query Context Analysis:** Extract explicit and implicit constraints such as study identifiers or scope.

7. **Resource Availability Check:** Validate that required indexed documents or datasets are accessible.

8. **RAG Execution Routing:** Route the query to the RAG pipeline for retrieval and response generation.

4. CORE GOVERNANCE PRINCIPLES FOR REGULATED CONVERSATIONAL AI

- **Identity-Bound, Role-Aware Access:** Conversational AI capabilities are strictly confined to post-authorization use, enforcing identity-based access and dynamically tailoring responses to user roles and permitted data scope.
- **Auditability and Provenance by Design:** Every natural language query is translated into explicit, traceable API calls with complete lineage, ensuring reproducibility and inspection readiness.
- **Validated Processing Boundaries:** The conversational layer performs no transformations, with all derivations and computations executed exclusively within validated, SOP-controlled pipelines.
- **Human-in-the-Loop and Fail-Safe Controls:** High-impact or ambiguous queries require expert review, and the system defaults to partial results or deferral when validation or confidence thresholds are not met.
- **Scalable Adoption Without Retraining:** Context-aware natural language access reduces dependence on specialized training, enabling broad adoption with minimal operational disruption.

5. INSIGHTS AND LEARNINGS FROM PILOT DEPLOYMENTS

Insights were derived from controlled pilot deployments of the RAG–enabled conversational interface executed within validated environments. The evaluation assessed usability, system performance, and alignment with regulatory expectations.

Key Learnings

- Measurable reduction in time to insight for common informational and exploratory queries.
- Strong benefits observed for queries related to study context, protocol, and CSR navigation, regulatory, and standards documentation.
- Conversational access lowered barriers to data discovery.
- Improved adoption among non-technical stakeholders, including Clinical scientists, Regulatory and Quality users, operational, and decision-making roles.
- Generated responses consistently aligned with regulatory expectations due to preservation of source provenance, end-to-end traceability, auditable retrieval, and response generation pathways.
- Outputs remained inspectable, justifiable, and reproducible within regulated workflows.

Performance Factors and Observed Challenges

- Semantic mismatches occurred between user intent and retrieved content due to variability in medical terminology, differences in regulatory phrasing, and cross-domain language usage.
- These mismatches highlighted the need for domain-tuned retrieval strategies, terminology-aware indexing, and normalization.
- Response accuracy was constrained by index freshness and content completeness.
- Outdated or partially curated repositories increased the risk of retrieving superseded documents.
- Findings reinforced the importance of robust document lifecycle management and dynamic and governed index update mechanisms.

Key Considerations for Regulated AI Adoption

- **Usability must be coupled with governance.**
  - Conversational interfaces are acceptable in regulated environments only when embedded within enforceable governance frameworks.
- **Mandatory governance guardrails**
  - Clearly defined and enforced scope boundaries aligned to validated use cases.

    - End-to-end data provenance and traceability, ensuring every response is attributable to governed sources.
    - Role-based access controls (RBAC) to restrict data visibility and functionality based on user authorization.
- **Model selection aligned to regulatory risk**
    - Domain-specific, right-sized models demonstrate superior performance over large, generic models by minimizing hallucination and unsupported inference, preserving semantic fidelity to regulated and curated content, supporting predictable and explainable outputs.

6. FUTURE DIRECTIONS

Future enhancements will focus on extending governed AI capabilities while preserving regulatory rigor.

- Key areas of development include:
    - Dynamic knowledge updates to reflect evolving regulatory guidance
    - Improved longitudinal and cross-study analytics
    - Deeper support for regulatory review workflows
    - Enhanced inspection readiness and submission-support use cases
- As these capabilities mature, AI systems are expected to:
    - Function as compliant analytical assistants
    - Accelerate insight generation
    - Remain auditable, reproducible, and ready to be inspected within validated environments

7. CONCLUSION

In summary, this Retrieval-Augmented Generation framework redefines biopharmaceutical data access by integrating a Unified Data Model (UDM) with securely embedded LLMs and intelligent agents, overcoming longstanding barriers in data harmonization, retrieval, and analysis. By delivering rapid, hallucination-resistant insights through intuitive natural language interfaces, it empowers data managers, researchers, and regulatory teams to navigate complex clinical and nonclinical datasets with unprecedented efficiency while upholding data sovereignty within sponsor networks. Ultimately, this system not only accelerates the drug development lifecycle but also unlocks new avenues for innovation, driving faster breakthroughs in therapeutic discovery and positioning RAG as a cornerstone of next-generation pharmaceutical intelligence.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact:

Author Name: Dr. Sapna Chandran L
Company: PointCross Life Sciences
Address: 3rd Floor, BVR Lake Front, No. 1/4,
Hebbal Outer Ring Road,
Bangalore, Karnataka,
Work Phone: +91 9496749846
Email: sapna@pointcross.com

Website: https://pointcrosslifesciences.com

Co-Author Name: Lakshmi B K
Company: PointCross Life Sciences
Address: 3rd Floor, BVR Lake Front, No. 1/4,
Hebbal Outer Ring Road,
Bangalore, Karnataka,
Work Phone: +91 9972141833
Email: lakshmi@pointcross.com

Website: https://pointcrosslifesciences.com