workbench

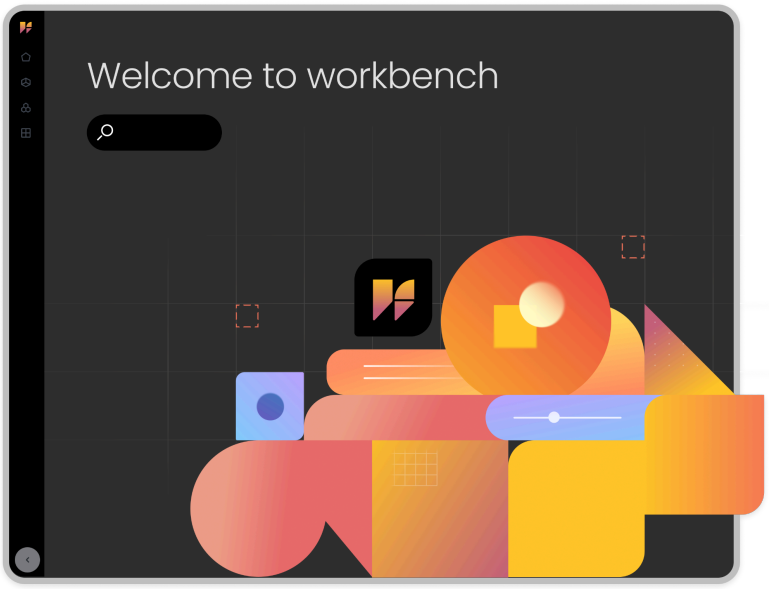
Accelerating time-to-discovery: A blueprint for collaborative, multimodal data exploration



Built on by verily

Executive summary

Biomedical research is often bottlenecked by the technical complexity of analyzing multi-layered, hierarchical data. Subject Matter Experts (SMEs) such as clinicians, possess the medical knowledge to ask the right questions but often lack the coding skills to navigate the hundreds of layers found in hierarchical data models like OMOP. Conversely, data scientists possess coding fluency but may lack the clinical nuance to interpret the data accurately. This disconnect, compounded by siloed systems and administrative friction, drastically slows the "time-to-discovery" that delays scientific inquiry.

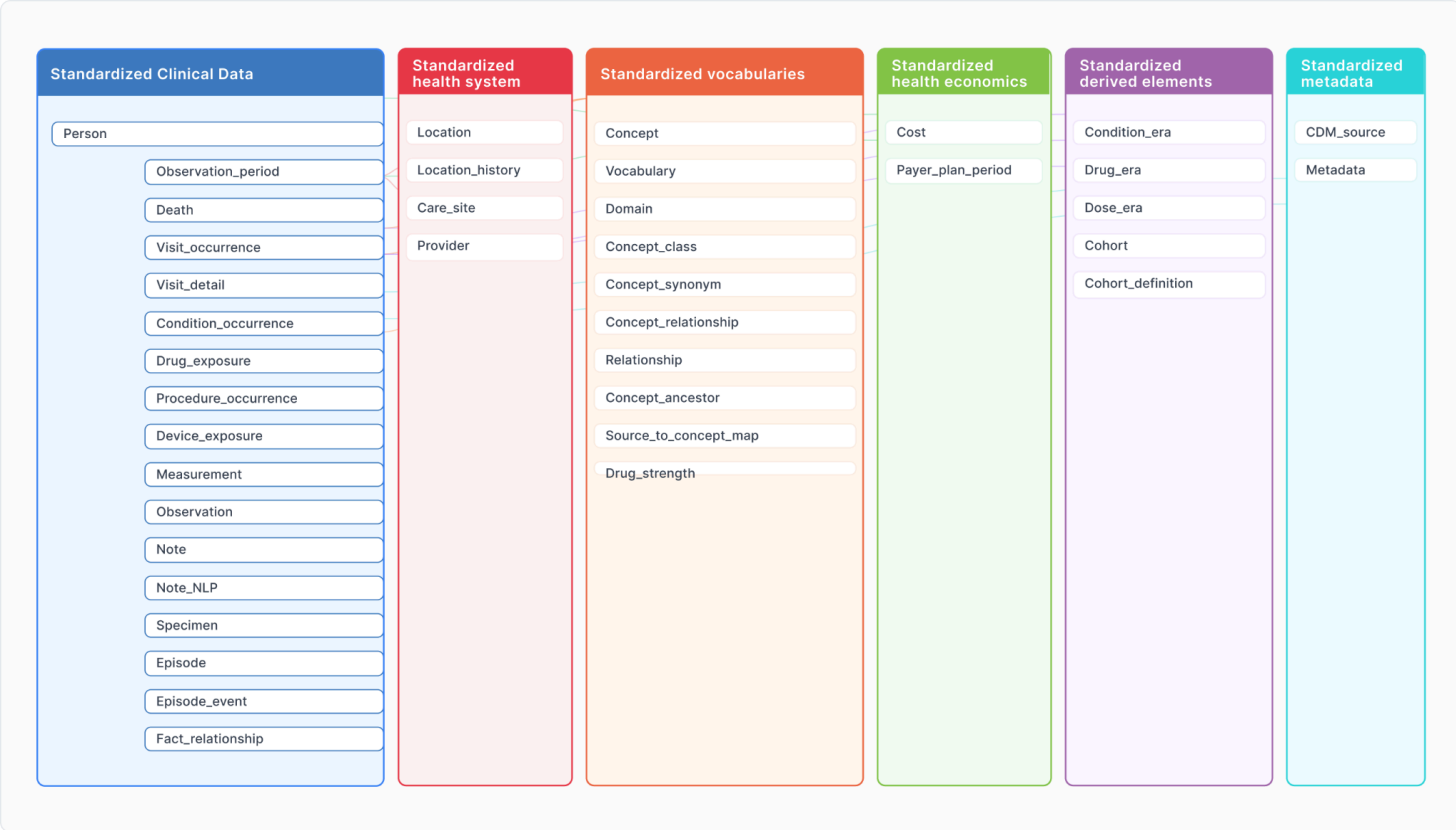


Biomedical science is a team sport. The right tools can bridge the divide between clinical and technical expertise. Verily Data Explorer offers a robust no-code interface for complex cohort discovery. It builds on work started at Vanderbilt Health and then scaled to support over 22,800 researchers in the NIH's *All of Us*Research Program. Verily Data Explorer is fully configuration-driven, allowing these capabilities to be deployed quickly for new datasets.

This whitepaper outlines a blueprint for distinct collaborative workflows: allowing researchers to visually define the "what" (cohort definition), validate it (interactively) and automatically exporting reproducible, standardized code to data scientists for the "how" (deep analysis). By transforming a bespoke academic tool into a modular, commercial-off-the-shelf (COTS) solution, Verily Data Explorer is democratizing access to petabyte-scale data and accelerating the path from inquiry to insight.

The challenge: navigating modern biomedical data

The *All of Us* Research Program represents a historic effort to gather health data from one million participants. To make this massive, multimodal resource useful, the dataset is mapped to the OMOP Common Data Model with additional customized data types including surveys and wearable sensors. While OMOP provides a powerful, research-oriented structure for clinical data from multiple sources, it introduces a steep learning curve. The model preserves the richness of real-world data through deep, hierarchical vocabularies, often resulting in complex nested parent-child relationships that are nearly impossible to navigate without an intuitive interface.



Visualization of OMOP complexities

The fragmented workflow of Real-World Data (RWD)

Before unified tools like the Data Explorer, the workflow for accessing RWD was fragmented and inefficient:

- **Siloed systems**: Data for research is often isolated across institutions and departments, hindering the compilation of comprehensive datasets. Researchers were previously required to navigate disparate systems with unique protocols and interfaces, imposing significant administrative burdens on both researchers and IT staff.
- **The coding barrier**: Accessing data frequently requires coding expertise (SQL/Python) to query databases directly, or reliance on paid services. Errors in coding and inconsistencies in retrieval often led to substantial delays and iterative "clean-up" efforts to construct usable datasets.
- **The reproducibility barrier**: When datasets were updated or expanded, researchers faced challenges adapting previous queries to include new data. This resulted in significant delays in cohort creation and threatened scientific rigor.
- **Inequity of access**: Limited institutional resources meant that only well-funded experts could access the specialized support needed to leverage high-quality datasets and computational tools, creating significant inequities in research opportunities..

The persona gap: why “one size” does not fit all

Navigating these multi-layered structures creates distinct challenges for the diverse players in the research ecosystem. The *All of Us* Researcher Workbench is currently utilized by 22,800+ researchers from 1,300+ institutions at every career stage.

- **Clinicians and SMEs**: Subject Matter Experts (SMEs) possess deep knowledge of disease pathology and patient care. However, they are typically trained to work with health data as individual narratives, not large-scale datasets meant to be analyzed in aggregate. They often face difficulties interpreting research-optimized data models (like OMOP) that differ significantly from the systems they use in everyday practice.
- **Data analysts and bioinformaticians**: While these users possess the coding fluency to manipulate large datasets, they typically lack exposure to the clinical environment. This knowledge gap can hinder their understanding of how patient data is actually collected - such as the difference between a “prescribed” drug and an “administered” drug - leading to potential misinterpretations of the data.
- **Early-career researchers**: Students and junior researchers often face a “double barrier”: they may lack sufficient familiarity with complex medical terminology, clinical data collection processes, research methodologies, and the advanced SQL or Python skills required to query petabyte-scale databases. Notably, 66% (13,570+) of current registered *All of Us* researchers are students and trainees. This creates a high barrier to entry, preventing the next generation of scientists from effectively engaging with the dataset’s full potential.

The scale of the problem

With *All of Us* data from over 633,000 participants, including 414,000 with whole genome sequencing (WGS), the dataset spans EHRs, high-frequency wearable sensors, and longitudinal surveys. This unprecedented breadth means that manual data exploration can be extremely overwhelming for users unfamiliar with standardized research data or large-scale omics datasets. For example, the OMOP Condition Domain alone encompasses 687,839 concepts organized in intricate hierarchies; and over 1.2 billion gene variants are identified. Without specialized tools to facilitate data querying, navigation, and analysis, users are forced to rely on code to assess the basic feasibility of a study.

When coding expertise becomes a prerequisite for asking questions, it creates a barrier to equitable data access. This hinders the democratization of scientific discovery, restricting the utility of a national dataset to a small subset of highly technical users.

The origin story: Built by clinicians, scaled by Verily

The Verily Data Explorer traces its roots to Vanderbilt Health, a pioneer in linking DNA banks with electronic health records. Motivated by the success of BioVU — VUMC 's massive biobank — clinician-researchers built the internal "Cohort Builder" to solve a specific, local problem: allowing researchers to test study feasibility independently, without needing to navigate the complex backend of raw clinical databases. This shifted the "Scientific North Star" back to the subject matter experts, enabling those with the clinical context to lead the data exploration process.

In 2016, to meet the unprecedented technical demands of the NIH's *All of Us* Research Program, VUMC led the partnership with Verily and the Broad Institute. The collaboration was strategic: VUMC provided the deep informatics expertise and the scientific "North Star", while Verily delivered the secure cloud infrastructure and enterprise engineering required to scale these capabilities to a national level.

Proven scale: The legacy of the researcher workbench

The partners initially integrated the VUMC Cohort Builder into the Legacy *All of Us* Research Workbench. This system proved the power of democratized access, and its success is evidenced by its massive footprint today:

- **A global community:** The Researcher Workbench is currently used by more than 22,800 researchers from over 1,300 institutions. Its reach is international, with over 310 institutions based outside the United States.

- **Massive adoption across modalities:** The Researcher Workbench has been essential for hypothesis generation, utilized in:
    - Nearly 15,500 workspaces for EHR data (Conditions, Drugs, Procedures).
    - Over 10,360 workspaces for Survey data (mapping complex question-response relationships).
    - Over 5,780 workspaces for large scale Genomics data.
- **Beyond academia:** The ecosystem has expanded to include nearly 250 commercial entities involving over 500 researchers, proving that the Workbench serves the needs of both academic and industry-based research.

Recognizing the need for a more sustainable and flexible architecture, VUMC and Verily evolved the Cohort Builder into the Verily Data Explorer - a modular component of the new Verily Workbench, built on Verily Pre, a powerful precision health AI platform designed to securely harmonize, model, and activate healthcare workflows. This modernization offers key advantages:

- **Configuration-driven architecture**: Moving away from bespoke engineering, the Verily Data Explorer is now fully configuration-driven. Institutions can onboard new datasets in days rather than months, drastically reducing the burden on IT departments and lowering maintenance costs for the customer.
- **Seamless collaborative handoff:** Point-and-click exploration results are saved as standardized SQL queries. This allows a clinician to define a cohort (the "what") and a data scientist to immediately execute that code in an analysis workspace (the "how").
- **Standardized learning**: Because the Data Explorer can be used across multiple datasets, researchers only need to learn one interface to work with data from diverse sources, significantly lowering the barrier to multi-institutional study.
- **Access to advanced AI tools:** Researchers can leverage the natively integrated NVIDIA AI tools and GPUs in Workbench.

By combining the proven success of the *All of Us* Research Program with a modernized, modular architecture, the Verily Data Explorer ensures that the path from data to discovery remains open to the widest possible range of researchers.

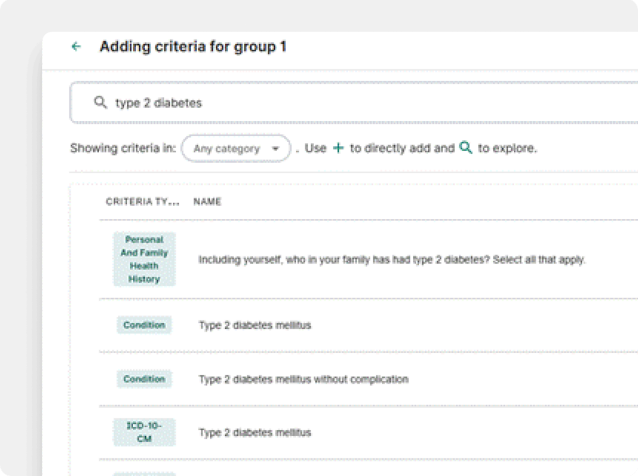
Core pillar: Democratizing access and supporting team science

The Verily Data Explorer functions not just as a viewing solution, but as a collaborative engine. It helps researchers visualize complex standards (like OMOP) without requiring an informatics degree.

Here's how the workflow looks for a researcher who is interested in studying Type 2 Diabetes but not sure what's available.

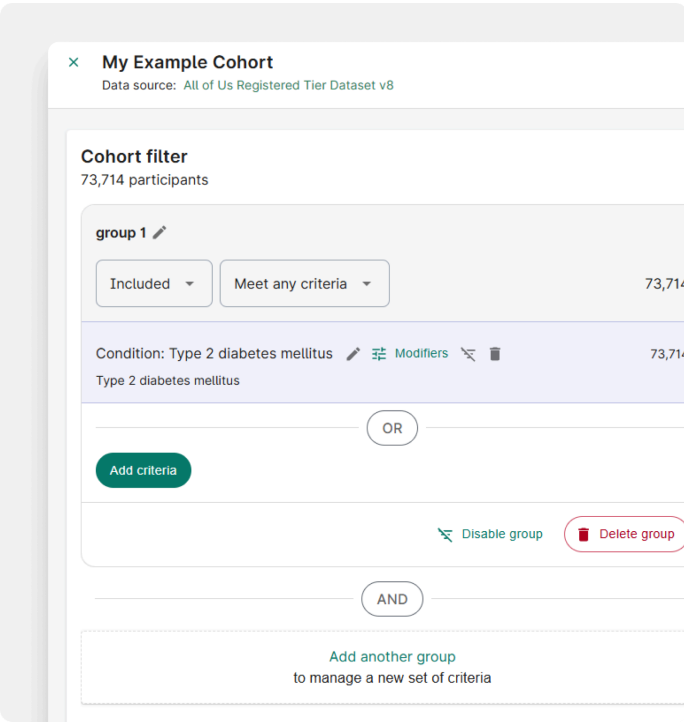
Visually browse and explore

Researchers can use the universal search function to explore available data types, such as Type 2 Diabetes, while receiving real-time updates on record counts.

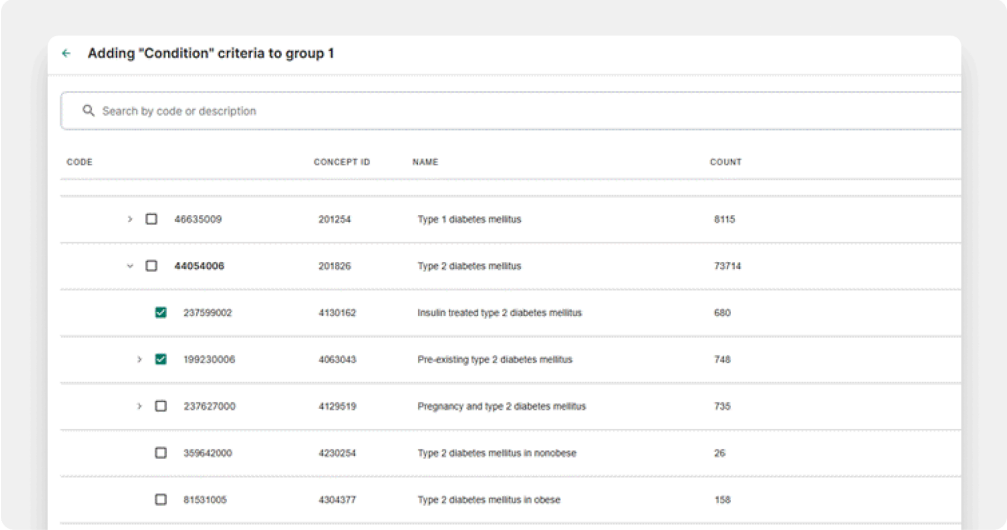


Build custom cohorts

Users can apply inclusion and exclusion criteria across multiple complex domains.



Users can also navigate through the cohort hierarchies, like Condition, Drug, or Personal and Family Health History, to explore and identify the criteria needed to further refine their study population.

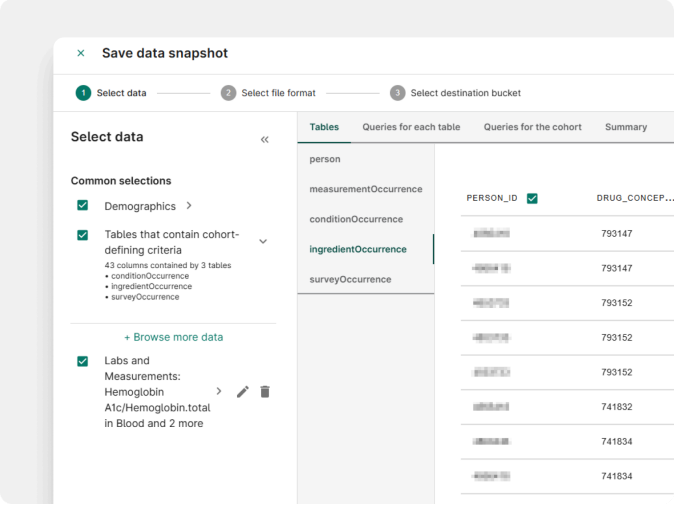


Review and annotate participants

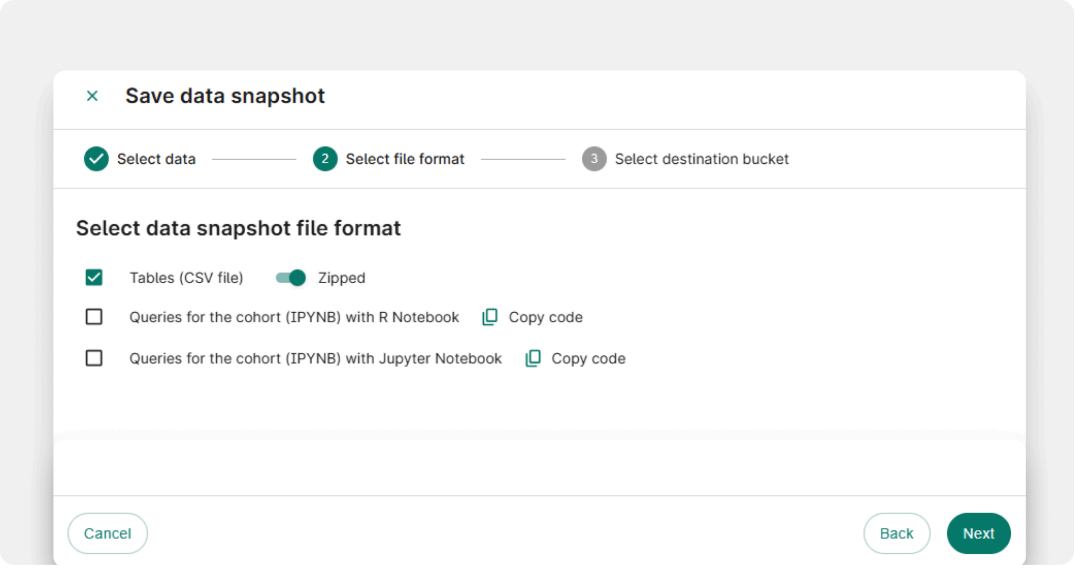
Check your cohort definition by reviewing a sample of individual participants, and add annotations to capture findings of interest.

Define and export for analysis

Once a cohort is defined, the system automatically generates a precise SQL query to retrieve that data for future analysis.



Users have the flexibility to export as zipped CSV tables or as integrated R/JupyterNotebooks



Deep analysis in Verily Workbench

- Deep analysis within a customized workspace: Cohort appears as a resource within the Verily Workbench, allowing it to be seamlessly shared, duplicated, referenced and modified by collaborators designated
- This workflow allows a clinician to define the what (the study population) while automating the how (the complex SQL syntax), effectively "democratizing" access for users who understand clinical nuances but aren't experts in RWD parsing or EHR database schemas.

Technical architecture: Built for rigor

Algorithmic transparency and standardization

A primary critique of “no-code” tools in a clinical setting is the “black box” effect, where the logic behind cohort creation is opaque or non-exportable. Data Explorer solves this by ensuring that the interface is intuitive, while the output is technically rigorous.

- **Automated standardization:** The application automatically generates standardized code (e.g. SQL, CSV, python) that corresponds precisely to the visual logic defined by the researcher.
- **Eliminating logic bias**: Automating the generation of cohort-defining code eliminates the variability caused by human error or differing levels of technical expertise
- **Ensuring reproducibility**: These code-based definitions serve as citable digital artifacts. They can be executed directly against the database to replicate findings, providing full transparency for peer review and compliance audits.

Precision querying

To enable precise cohort creation on massive petabyte-scale datasets, the architecture utilizes a configuration layer known as the “Underlay”. Functioning as a semantic bridge that sits between raw data and the user, the Underlay maps the native model of a dataset into a unified model that drives the Data Explorer indexer, query builder, and user interface. This allows researchers to define cohorts through robust logic searches that incorporate:

- **Complex logic:** Cohort definitions can incorporate inclusion and exclusion criteria, cross-domain cohort selection, and advanced filtering without writing a single line of manual code.
- **Temporal precision**: Point-in-time limiters that establish chronologies of care (e.g. distinguishing between a past history of a condition vs a current diagnosis).
- **Multimodal flexibility:** While optimized for the OMOP Common Data Model, the system is built for the inherent “messiness” of real-world research. The architecture supports arbitrary schemas and custom extensions for diverse modalities, including genomic extracts (SNP variants), imaging (DICOM), wearables, and longitudinal surveys.

Performance and data integrity

To maintain scientific integrity without sacrificing performance, the architecture separates data storage from the exploration interface, ensuring that exploration and hypothesis testing never alter or burden the production-grade source tables:

- **High-performance indexing**: An optimized index layer is used to meet the need for real-time browsing across petabyte-scale datasets. This ensures rapid response times and fluid visualization, allowing researchers to iterate on cohorts at the "speed of thought" without impacting the model, performance or costs associated with the source data.

- **Rigorous QA:** To ensure total scientific accuracy, rigorous quality assurance measures are applied to all source data and their corresponding intermediate representations — including the semantic Underlay and optimized indices. This systematic verification & validation ensures that the data researchers interact with is technically sound and aligned with the authoritative source.

Seamless handoff and collaboration

The architecture is designed to support the full research lifecycle, moving beyond exploration into deep analysis in the researcher's preferred toolset.

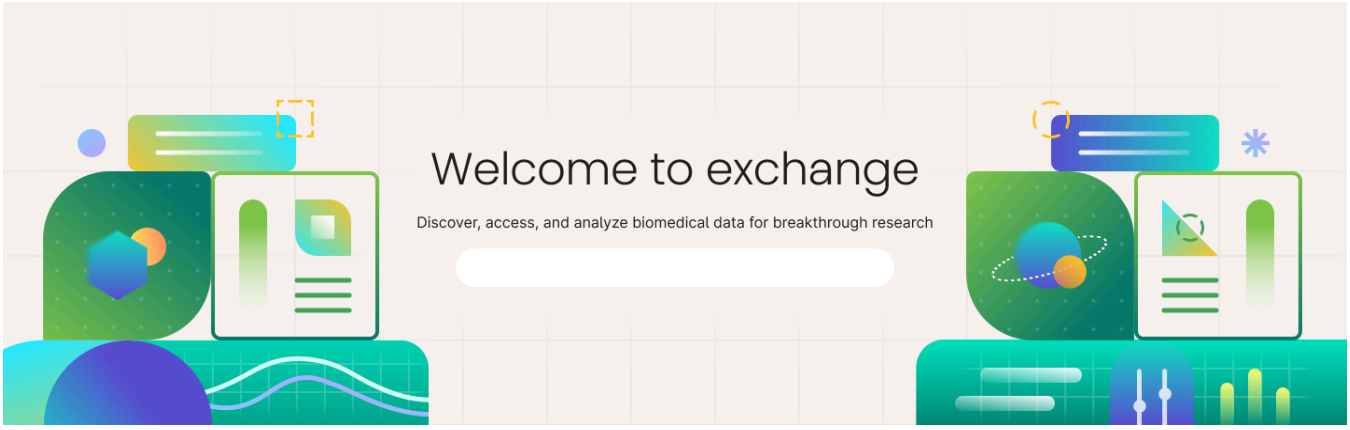
- **Extensibility:** Researchers can select concepts and data to be exported. Exports can produce tabular data files (.csv) or code artifacts containing queries (e.g. R and Jupyter Notebooks) within Verily Workbench. This allows each researcher to continue working with their data outside of the tool, fitting seamlessly into existing analysis workflows
- **Team science:** Shared workspaces promote collective rigor, allowing multiple researchers to collaborate on cohort definitions in real-time. When code-proficient users use Verily Workbench to conduct novel analyses or develop visualizations, those can also be shared back to scientific experts through Workbench for review.

The Verily Data Explorer is defined by its inherent flexibility. Rather than being tethered to a single technology stack, the system is engineered for multi-cloud environments. This modular, multi-cloud approach provides institutions with the long-term freedom to scale and adapt their data strategies as the research landscape evolves.

Technical specification table

Feature	Specification	Research benefit
Data model	Native OMOP CDM (v5.x); extensible to custom schemas	Standardized cross-institutional analysis
Scalability	Optimized for 1M+ participants / Petabyte-scale	Performance degradation stays strong as datasets grow
Latency	Sub-second response via optimized Index Layer	Real-time hypothesis testing and feasibility
Auditability	Automated SQL/JSON export per cohort	Provenance for regulatory/peer review
Interoperability	Export cohorts to R, Python (Jupyter), and SAS	Seamless handoff to deep-coding environments
Security	HIPAA/FedRAMP high-compliant architecture	Secure handling of PII/PHI at scale

The future: AI and enterprise scale



The Verily Data Explorer is a blueprint for how academic medical centers and technology leaders can co-innovate to accelerate discovery. As we move towards the next generation of AI-driven, collaborative medicine:

Conversational cohort building: The integration of LLMs to support natural language querying is under active exploration. Instead of navigating filter menus, researchers will soon be able to simply ask “Show me a cohort of female participants over 50 with Type 2 Diabetes,” and the system will automatically generate the precise reproducible SQL required to build that cohort.

Verily Exchange: Future capabilities will allow for interactive pre-purchase exploration. Researchers can anonymously visualize cohort feasibility and data density before provisioning a secure workspace, lowering the risk for data consumers and allowing data generators to securely showcase the value of their data.

By evolving from a bespoke cohort builder to an intelligent, enterprise-grade solution, the Verily Data Explorer is accelerating 'time-to-insights.' This path of innovation is leading toward a reality where technical complexity dissolves, enabling discovery at the speed of thought, where researchers can iterate from hypothesis to insight in minutes, not months.

Contact us to learn more at verily.com/workbench-request