Accelerating and Extending the Benefits of Observational Research into Advanced Outcomes Analytics

Integrated Data for Analysis across the Drug Development Process Extends Outcomes Research

02.15 EB8704 LIFE SCIENCES

TERADATA.

Table of Contents

Executive Summary

While the Observational Medical Outcomes Partnership (OMOP) has been successful at providing common data models and standard algorithms for improving medical product population studies and safety surveillance using existing real-world data, there are ongoing data challenges and unrealized opportunities for extending the benefits of observational research. The sheer size and scope of data and the lack of processing efficiency in many of today's first-generation IT platforms limit the ability for life sciences companies to unlock the value from their internal and external data.

An integrated analytical environment that combines optimized solutions for analyzing structured and multi-structured data extends the benefits of OMOP and Mini-Sentinel so life sciences companies can perform new types of analytics that move beyond population and safety outcomes. By expanding outcomes analytics across the enterprise, benefits are extended to a wider range of users to fully realize the potential of data-driven real world insights throughout the business.

A World Filled with Data

With a goal of conveying the appropriate use of real-world observational databases for studying the effects of medical products, OMOP—covering more than 200 million patient lives across a network of disparate databases—can be considered a success. It has provided common data models and standard algorithms that previously did not exist. The result of both OMOP and Mini-Sentinel has been a better system for monitoring drugs, devices, and procedures to reliably identify risks and opportunities to improve patient care.

Life sciences companies can now continuously prove drugs are safe and effective using real-world evidence from massive heterogeneous datasets, while also ensuring that they are producing the greatest health benefit outcomes at the "right" cost. That's not to say there have not been and don't continue to be ongoing data challenges because of the extraordinary size and scope of data, as well as the lack of processing efficiency.

State-of-the-art statistical analysis software packages offer the ability to uncover new insights and boost scientific productivity, but their performance is limited by the resources available in the hardware platform. OMOP consists of many complicated, long running standard SAS routines. A lack of processing efficiency means OMOP routines can take hours or days to run. Plus, vast amounts of duplicate data spread across the enterprise in disparate systems add one more barrier to be able to rapidly analyze information without IT intervention.

Pushing the OMOP Boundaries

OMOP has delivered a promising set of standards and best practices to follow for outcomes analysis of very large datasets through standard data models, vocabularies, methods, metrics, and algorithms. With more and more real-world evidence data becoming available—clinical data from healthcare providers, claims data from insurers, electronic medical records, and even self-reported outcomes data from social media sources—there is tremendous opportunity to extend the benefits of OMOP and integrate new data sources that OMOP does not support.

These datasets are structurally diverse because they originate in different systems within very different hardware and software environments. They include structured, multi-structured, and potentially unstructured data types, not all of them well suited to relational databases or to SQL-based analytics.

Because of these inherent data management challenges and other technical barriers, it is difficult to load and integrate full datasets into a single repository or build linkages across them to create patient-level longitudinal data. To do so requires a comprehensive set of transformation rules, very large storage volumes, extremely efficient translation and loading engines, and flexible, scalable hardware and applications. It also requires accelerated analytical execution, which can be done by performing all analytical computation directly on the data repository platform.

By extending the benefits of OMOP, life sciences companies can perform new types of analytics that move beyond simply understanding the burden of a certain illness or conducting safety signal detection for one condition for one drug. And, move analytics across the enterprise to extend the benefits to a wider range of users to fully realize the potential of data-driven real world insights throughout the

business. The impact can reach into a number of new areas, such as market research, identifying gaps in therapy, determining how to price and market drugs, and better positioning products to all customers (payers, physicians and patients). It delivers the ability to prove that products are not only safe and effective, but also provide the greatest health benefit outcomes at a competitive price to gain and keep top-tier formulary status.

The Benefits of Advanced Outcomes Analytics

The ability to incorporate new data sources for a larger group of users across the enterprise to answer more questions in a shorter period of time is the payoff for extending OMOP and Mini-Sentinel models beyond outcomes and safety research. Epidemiologists, scientists, and analysts gain more analytic freedom to finally ask the hard questions, across all data and all variables, with the ability to iterate on data at the speed of thought. The result is rapid and accurate real-world evidence insights that confirm a health and/or economic outcomes-related hypothesis or provide input and direction for the next question. As OMOP and Mini-Sentinel continue to evolve and unknown regulatory changes continue to lurk, advanced outcomes analytics also allow life sciences companies to respond faster to changes and new regulations to better compete in a crowded market.

Know More to Do More

Teradata helps enterprises know more about their customers, products and business so they can do more of what really matters, better and faster than any alternatives. By unlocking the potential in large, diverse data, Teradata empowers life sciences companies with new capabilities that impact many areas, complementing and extending the work done by the OMOP consortium.

Teradata's analytic platform can consume, manage, and iteratively analyze large and diverse data sources, overcoming many of the data management and technical barriers life sciences companies have encountered with OMOP and Mini-Sentinel. Teradata's solutions offer a high level of software-based parallelism and hardware-based scalability. They also support in-database analytical execution for SAS software and other third-party analytics providers for dramatically improved processing efficiencies and much faster analysis.

In place of obstacles are opportunities—the ability to stay ahead of consumer-facing issues, drive product use, identify superior performing patient sub-segments, optimize R&D portfolios, gain or defend market access, and maintain formulary status.

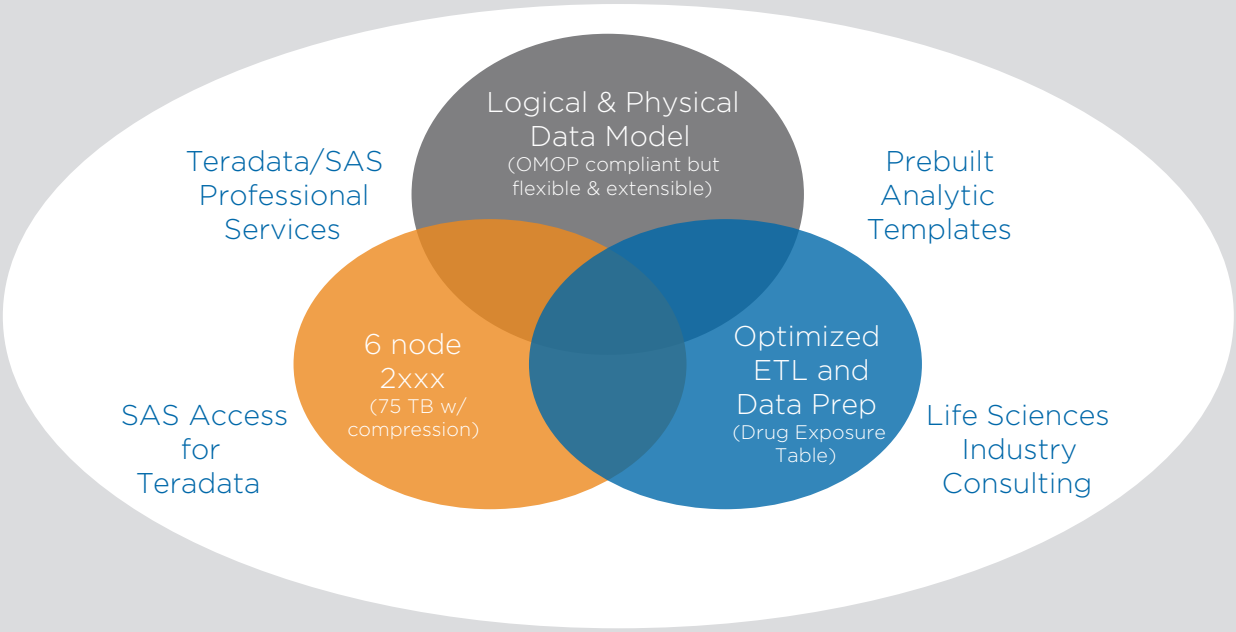
The Teradata Outcome Analytic Accelerator

Teradata's Outcome Analytic Accelerator removes barriers to insight, resulting in lower "scientific discovery to insight" cost, dramatically improved scientific productivity, reduced data replication, and an established foundation for R&D data integration above and beyond purchased RWE/outcomes data.

The Outcome Accelerator package consists of several components, including OMOP or Mini-Sentinel Common Data Model (CDM) mapping, data loading, and custom ETL to dramatically increase data loading performance and throughput.

- **Teradata Appliance/Integrated Data Warehouse** – provides a shared environment for strategic and operational analytics, and a single centralized source of data for reuse, as well as a data lab so analysts can quickly and easily self-load experimental new data with existing data to identify new trends or insights.
- **Teradata Life Sciences Data Model (LSDM)** – includes both a logical data model (LDM) and a physical data model (PDM) to support the migration from the OMOP

Teradata OMOP Analytic Offer



CDM. It delivers flexibility by offering a view that is identical to OMOP or Mini-Sentinel or a view with additional data sources integrated. The LDM is an integrated, cross-functional model that provides an enterprise-wide view of data. It's a comprehensive, flexible blueprint of how data is organized and provides structure that makes it easier to manage information. The PDM is generated from the LDM and provides life sciences companies with a starting point to speed the review and application of customer requirements and adjust the model to match their way of doing business.

- **Optimized ETL and data preparation**
- **SAS access for Teradata**
- **Teradata Unified Data Architecture**™ – a comprehensive enterprise approach that seamlessly integrates multiple technologies into a cohesive and hybrid architecture for all data, data processing, and analytic needs.
- **Data staging platform** – provides a cost-effective environment for loading, storing, and refining services to prepare all incoming data structured, unstructured, or multi-structured—for analysis based on the value of the data. Apache™ Hadoop® is highly optimized for precisely these functions.
- **Prebuilt analytic template** – Teradata Aster Discovery Platform, TIBCO® Spotfire®, FuzzyLogix and R. The discovery platform makes it faster and easier for a wider group of users to generate powerful, high impact business insights from big data. Rapid discovery analytics deliver iterative exploration using a variety of analytic techniques to support dynamic decision making.
- **Professional services and industry consulting** – Teradata's industry expertise is provided by senior industry consultants who have worked at top 10 pharmaceutical companies and have a combined 40 years of industry experience. Areas of expertise specific to observational analytics include:
  - OMOP knowledge and experience
  - CDISC modeling
  - R&D logical and physical data modeling (including observational)
  - Documenting user requirements, developing functional specifications, mapping requirements to functional specifications and developing end-user test plans

Teradata in Life Sciences

Eight of the top 10 life sciences companies choose Teradata to lead their way to a better understanding of their business and their customers, and position their products to be superior through data. Many of the world's most successful life sciences companies use Teradata for such critical tasks as:

- Conducting safety signal detection and analysis on all products for any condition/event within hours or days versus weeks or months.
- Ongoing rationalization of global corporate-wide spend to significantly reduce costs.
- Analyzing success of marketing channels for all products within the portfolio and being able to rapidly redirect efforts as needed.

Teradata has more than 30 years of experience working with business and government leaders, creating and implementing data-driven solutions with the flexibility and scalability to effectively handle changes now and in the future. Teradata solutions and applications transform big data sets into actionable knowledge and valuable insights so life sciences companies can know more to do more. Teradata delivers complete, best-of-breed solutions that include data platforms, discovery platforms, integrated data warehouses, analytics, and marketing applications. Our integrated solutions are engineered to analyze massive amounts of data in microseconds, yet designed for everyday use by the broadest constituency of epidemiologists, scientists, analysts, and business users.

Case Study #1

Query Time Cut from Days to Minutes

The epidemiology department in a top-10 pharmaceutical company runs the OMOP standardized Univariate Self-Controlled Case Series (USCCS) macro to determine cause-and-effect relationships between drugs and conditions in large populations, as well as precise surveillance windows. The department's original platform was SAS software running on shared UNIX servers.

Very large input tables caused extremely slow run times. Only one analysis could be run at a time, taking four hours to complete with only a limited data sample. To solve the performance challenge, the department took a hybrid approach running the SAS USCCS macro in-database on a Teradata platform. The move paid major dividends, with an average of 70-percent data compression, and processing time for one analysis being reduced from four hours to five minutes. Processing time for the longest-running query dropped from 4.3 days to 3.8 minutes—a 1,600 times performance improvement.

Case Study #2

Four-Hour Safety Signal Detection Analysis Decreased to Five Minutes

Pharmacoepidemiology investigators have typically been limited in their ability to complete studies and explore new methods because of constraints with the size of databases and the capacity of the servers and server software to allow the manipulation of data for more complex statistical tools to be evaluated. A major pharmaceutical company turned to Teradata to conduct a proof of concept (POC) to prove that its hardware/software solutions could solve the first two constraints so that more complex simulation studies could be carried out in the pharmacoepidemiology space. The Teradata solution allowed for novel signal detection and more complex analytics to be used. The benefit came in two steps through significant improvements in data manipulation and in the implementation of novel methods.

Using GE EMR data that consisted of medications, diagnoses, procedures, and medical encounters for approximately 25 million U.S. lives followed for a median of three years. The data was arranged to conform to the common data model (FDA Sentinel/OMOP). The largest tables are the medications and the conditions eras with >30 million rows each.

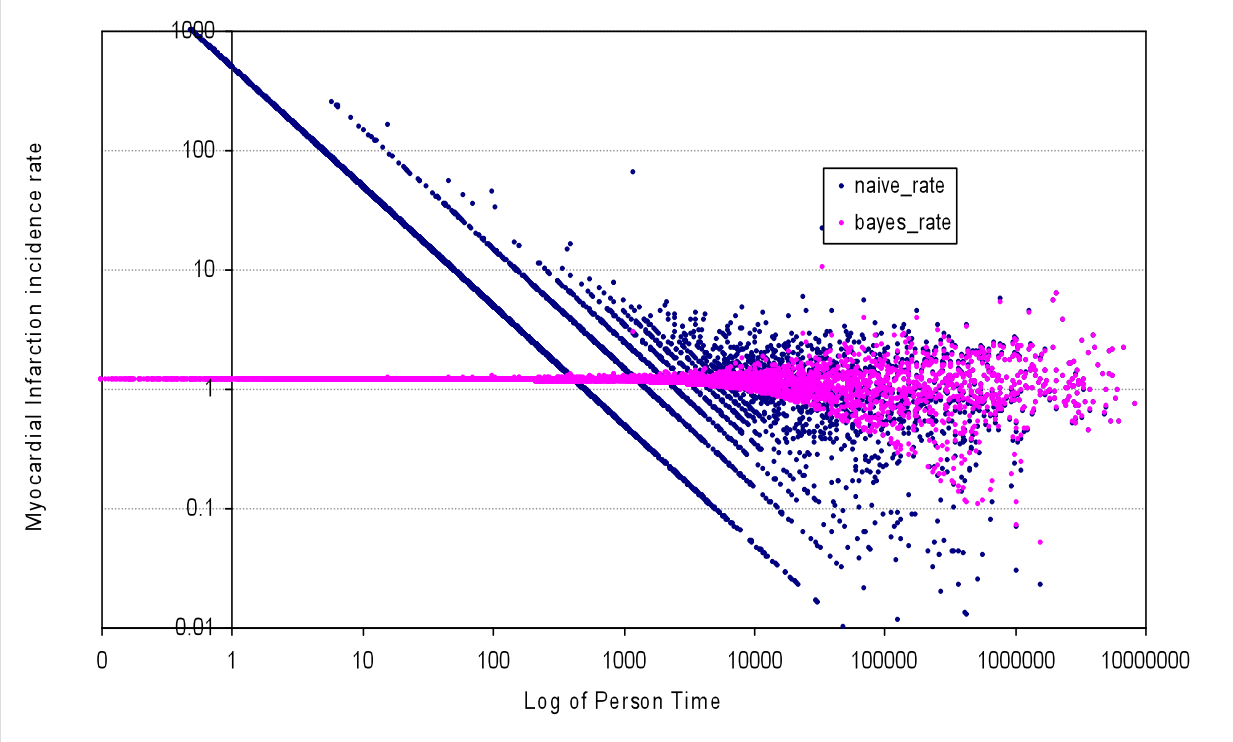
The POC involved the computation of Bayesian Poisson rates, naïve Poisson rates and Bayesian rate ratios in the context of detection of a safety signal. The main problem of signal detection is that it requires the use of the full dataset in order to compute the background rate (prior distribution) of a given signal (i.e., myocardial infarction). Using the Teradata solution, the pharmaceutical company was able to discover the rate at which the signal occurred in a given amount of time for any and all medications. Once it was obtained, the pharmaceutical company computed/contrasted Bayesian/naïve rates for before/during/after the exposure happened (posterior distribution). The POC also involved an exact matching of variable ratio using propensity score for two cohorts (drug 1 and drug 2) in the context of a safety signal evaluation procedure (R or SAS).

The Results

The POC results proved that the Teradata solution would enable more complex pharmacoepidemiology simulation studies to be carried out in record time. The customer commented, "The Teradata hardware/software combination provides a significant advantage in analyzing large databases compared to SQL Server/Oracle, with loading and manipulating large amounts of data being relatively painless." The customer added, "Teradata statistical functions have incredible potential."

Tasks	Sim Rows	Prior	Method 1 in R	Method 2 Teradata in DB	Success
MatchIt #1	200,000 (2x the cohort specified)	7 hr (100,000)	3 Hours	2-3 Minutes	<4 hr; <1 hr
MatchIt #2	1.2 million (12x the cohort specified)	N/A	Did Not Finish	5 Minutes	Extra Credit

MI Rate for any Medicine in the ER



Case Study #3

Processing Time for Active Epidemiology Surveillance Drops from Hours to Seconds

The Epidemiology team at a top pharmaceutical company has an active surveillance deliverable, which is to test three active surveillance methods (PRR, USCCS and HDPS) on the THIN and Optum Data Mart claims databases. These methods have been implemented (with varying success) by OMOP. The Epidemiology team's implementations are to varying extents built on the OMOP and Mini-Sentinel implementations. Previously, the Epidemiology team was running the methods using SAS code on UNIX servers, which proved to be extremely slow and used large amounts of storage. The time for analyses also imposed a challenge to completing all necessary method testing for informing appropriate active surveillance methodological strategies around supporting the portfolio.

As a first step, the PRR active surveillance method was run on the Teradata platform during the proof-of-concept phase, greatly increasing speeds and decreasing storage needs. Processing time for a single compound was reduced from 46.5 hours to 37 minutes in initial testing. The second step was to test the USCCS method, which resulted in a reduction in processing time from 2 hours 56 minutes to 18 seconds.

For more information, contact your Teradata representative or visit **Teradata.com/healthcare**.

10000 Innovation Drive, Dayton, OH 45342 **Teradata.com**

02.15 EB8704

TERADATA.