Paper RE05

RWD PEARL

Pharma **E**nhanced **A**nalysis (and **A**I) **R**eady **L**ayer

A Standardized, Specialized, and Innovative Data Solution for Efficient Real-World Data Analysis

Navdeep Pal, Genentech, South San Francisco, CA

Nayan Chaudhary, Genentech, South San Francisco, CA

Brandon Arnieri, Genentech, South San Francisco, CA

Abstract

Real-world data (RWD) analysis is crucial for modern drug development but faces significant challenges due to inconsistent methodologies, fragmented data sources, and limited standardization. RWD PEARL (Pharma Enhanced Analysis and AI Ready Layer) addresses these challenges by providing a specialized, standardized framework for efficient real-world data analysis. This innovative solution combines domain expertise with advanced data science techniques to create analysis-ready datasets while maintaining scientific rigor. By streamlining data transformation, ensuring reproducibility, and enabling AI-readiness, PEARL enhances cross-functional collaboration and accelerates decision-making processes across clinical, analytics, and RWD teams. The framework's implementation demonstrates significant improvements in operational efficiency and data quality.

Introduction and The Gap

Real-world data (RWD) holds immense potential for generating actionable insights, supporting evidence-based decisions, and enhancing patient outcomes. The use of RWD has transformed drug development strategies, enabling biopharmaceutical companies to accelerate and enrich product development. RWD is generated from the routine delivery and administration of healthcare, encompassing sources like electronic health records (EHRs), claims data, disease registries, and even more novel data types like genomics and social determinants of health. As highlighted by Dagenais et al., RWD provides valuable insights throughout the drug development lifecycle, from pipeline strategy to clinical trial design and execution. By leveraging advanced analytics, biopharmaceutical companies can use RWD to refine disease prevalence estimates, select endpoints, and identify target populations, thereby enhancing both the efficiency and precision of their development programs[1].

Despite the promise inherent in RWD, several challenges arise when attempting to transition from raw data to meaningful, reproducible insights. First is the lack of standardization. While industry-standard frameworks such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model [2] provide a foundational structure for raw data—akin to the Study Data Tabulation Model (SDTM) [3] used in clinical trials—there is no widely accepted equivalent to ADaM (Analysis Data Model) [4] for generating submission-ready or analysis-ready RWD datasets. This missing link hinders the seamless movement of real-world data from its raw messy state to a final, refined form suitable for regulatory review or high-stakes clinical decision-making.

The second challenge is the inherent fragmentation of data sources. Because RWD can originate from multiple platforms, including EHR systems, labs, pharmacies, and patient registries, variations in data quality, level of detail, and methodological approaches to data capture can lead to inconsistencies. This heterogeneity not only challenges the integration and replication of analytical processes but also undermines confidence in the derived real-world evidence, particularly when variable derivations are inconsistently documented across studies.

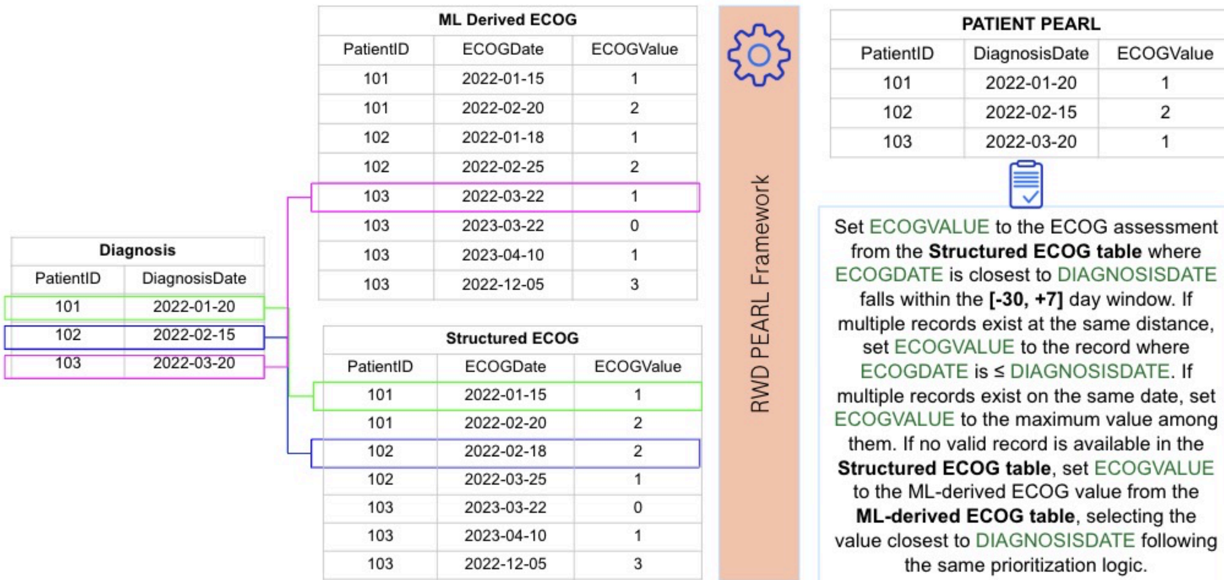
A third challenge concerns the distinct operational contexts of real-world data teams. Unlike specialized clinical trial data science teams, RWD teams often work with limited resources while addressing a broad spectrum of research questions—from rapid internal exploratory analyses to complex regulatory submissions. Therefore, the necessity for efficiency, reproducibility, and traceability is paramount, without the intensive processes typical of clinical trial workflows.

To fill these gaps, the concept of the Pharma Enhanced Analysis (and AI) Ready Layer (RWD PEARL) emerges as a vital innovation. Rather than introducing a rigid new standard, RWD PEARL strives to create a flexible, provenance-preserving framework that facilitates the consistent production of analysis-ready datasets. This approach is especially pertinent in today's era of rapidly expanding data sources, heightened regulatory scrutiny, and rising expectations for data-driven decisions. Crucially, RWD PEARL lays the groundwork not only for traditional statistical analyses but also for the growing domain of artificial intelligence (AI) and machine learning, enabling these advanced methods to operate on high-quality, pre-structured data.

By bridging the gap between raw, heterogeneous data and scientifically rigorous, analysis-ready datasets, RWD PEARL seeks to bring transparency, reproducibility, and trust to real-world data science. The resulting benefits include faster generation of insights, greater alignment between internal exploratory needs and external regulatory requirements, and an overall improvement in how organizations utilize the full potential of real-world evidence.

What Is RWD PEARL and How Does it Bridge The Gap?

RWD PEARL stands for Pharma Enhanced Analysis Ready Layer, and it represents a deliberate rethinking of how real-world data should be prepared and utilized for research and decision-making across the life sciences. Rather than serving as a simple data pipeline or set of static datasets, RWD PEARL constitutes a dynamic, holistic framework designed to unify various streams of real-world data into a consistently interpretable, analysis-ready state. This reimagined layer draws upon established practices in clinical data management—where frameworks such as SDTM and ADaM define clear standards for data structure and derivations—and adapts them to the more fluid, variable context of real-world evidence.



One defining feature of RWD PEARL is domain-specific specialization. Recognizing that a “one-size-fits-all” approach is inadequate, RWD PEARLs are built to address specific therapeutic areas. This focus ensures that each RWD PEARL mirrors the unique clinical nuances, terminologies, and research priorities relevant to a particular disease or treatment category. By embedding domain expertise directly into the design of variables, derivations, and documentation, researchers and analysts can derive more meaningful, context-rich insights. Generic data models may fail to capture these subtleties, whereas a tailored approach fosters a deeper level of scientific understanding.

Another central aspect of RWD PEARL is its role as a presentation layer. In other industries such as finance, commerce, and logistics, presentation layers are designed to offer rapid, consistent access to large datasets, thereby supporting swift decision-making. Additionally, presentation layers are created by teams who have deep expertise in the raw data and can build in the right derivations. Similarly, RWD PEARL ensures that the transformation from raw data to analysis-ready datasets is not merely a matter of standardizing formats; it is a carefully orchestrated process that preserves data provenance, enforces methodological consistency, and documents every step. This overarching architecture allows for both high-throughput analyses and the meticulous traceability required in regulatory environments.

Furthermore, reusability underpins the value proposition of RWD PEARL. By incorporating standardized derivations and metadata, once a PEARL is built for a particular therapeutic area, it can be employed across multiple studies or research questions without requiring investigators to recreate the same transformations or logic repeatedly. This level of efficiency is particularly critical for RWD teams that often must answer numerous queries rapidly, from internal feasibility or comparative effective studies to external demands for submission-ready evidence. In addition, reusability helps mitigate potential inconsistencies that arise when different teams or analysts define the same variables differently, thus ensuring reliability and reproducibility of insights across the organization.

Finally, in the era of AI and machine learning, data readiness is crucial. RWD PEARL transforms raw, heterogeneous data into a clean, “plug-and-play” format that accelerates model development and deployment, reducing preprocessing time and allowing data scientists to focus on advanced analytics.

Robust data provenance—ensured through automated logging and transparent derivation logic—enables any variable in the final dataset to be traced back to its source. This transparency not only strengthens scientific rigor but also meets evolving regulatory expectations for reproducibility and reliability in real-world evidence generation.

Table: Benefits of RWD PEARL

Benefit	Description
Enhanced Transparency & Reproducibility	Every variable’s derivation and transformation steps are documented, improving reproducibility and building trust among stakeholders and regulators.
Scientific Rigor	Collaboration with domain experts (clinicians, data scientists, statisticians) ensures that definitions, methods, and algorithms align with real-world clinical practices and evidence-based standards.
Improved Efficiency	Automated data pipelines, standardized derivations, and detailed documentation reduce redundant coding and speed up analyses, freeing teams to focus on interpretation and innovation rather than data wrangling.
Scalability & Reusability	Once a derivation is validated and included, it can be applied across multiple studies or therapeutic areas, supporting higher throughput and consistent insights.
AI-Readiness	Clean, standardized data expedites model development and deployment for advanced analytics, including machine learning and generative AI approaches.
Regulatory Alignment	Preserving data provenance and ensuring traceability at each step facilitates regulatory submissions and compliance with emerging RWD-focused guidelines.

Cross-Functiona l Synergy	Fosters collaboration among diverse teams—clinical, data, regulatory—promoting knowledge-sharing and aligning organizational goals.

Table: Use Cases for RWD PEARL

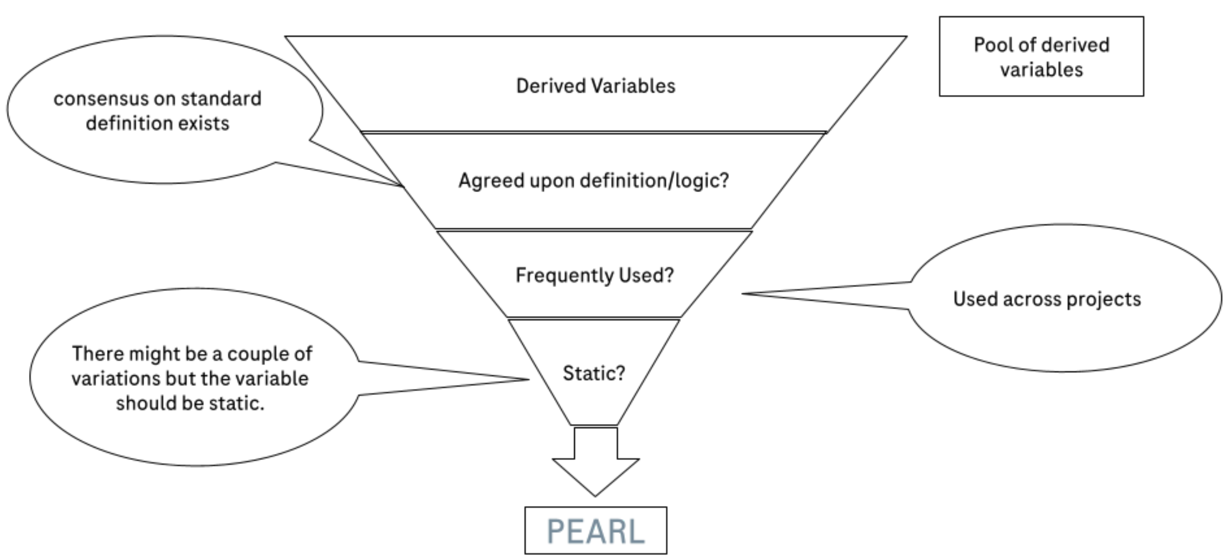
Use Case	Examples	How PEARL Helps
Regulatory	- RWD Regulatory Submissions, External Controls	Ensures transparent, traceable data suitable for regulatory filings. Standardized derivations and robust documentation reduce rework and accelerate the review process, instilling confidence in real-world evidence.
Software Apps	- Interactive Trial Exploration (eg. TEAL)	Powers interactive tools with consistent, curated data, enabling reliable cohort selection and trial planning. PEARL's standardized logic ensures stable outputs across multiple user queries and scenarios.
RWD Analyses	- Feasibility Studies to Comparative Effectiveness	Maintains clinically relevant, validated variables, allowing researchers to quickly conduct feasibility checks and compare treatment outcomes. This leads to reproducible, trustworthy findings for diverse study designs.
Machine Learning	- Data Preprocessing & Cleansing, Feature Engineering	Delivers ML/AI-ready, high-quality data that require minimal cleanup, enabling data scientists to focus on advanced modeling. PEARL's consistent schemas boost model performance and accelerate insight generation
Generative AI & LLM Agents	- Large Language Model Q&A, Agents	Brings raw data closer to actual business use cases by providing structured, high-fidelity information. PEARL's organization and transparency enable LLM-based systems to reliably answer questions that would be difficult with unstructured or inconsistent RWD.

How to Develop RWD PEARL in Your Organization?

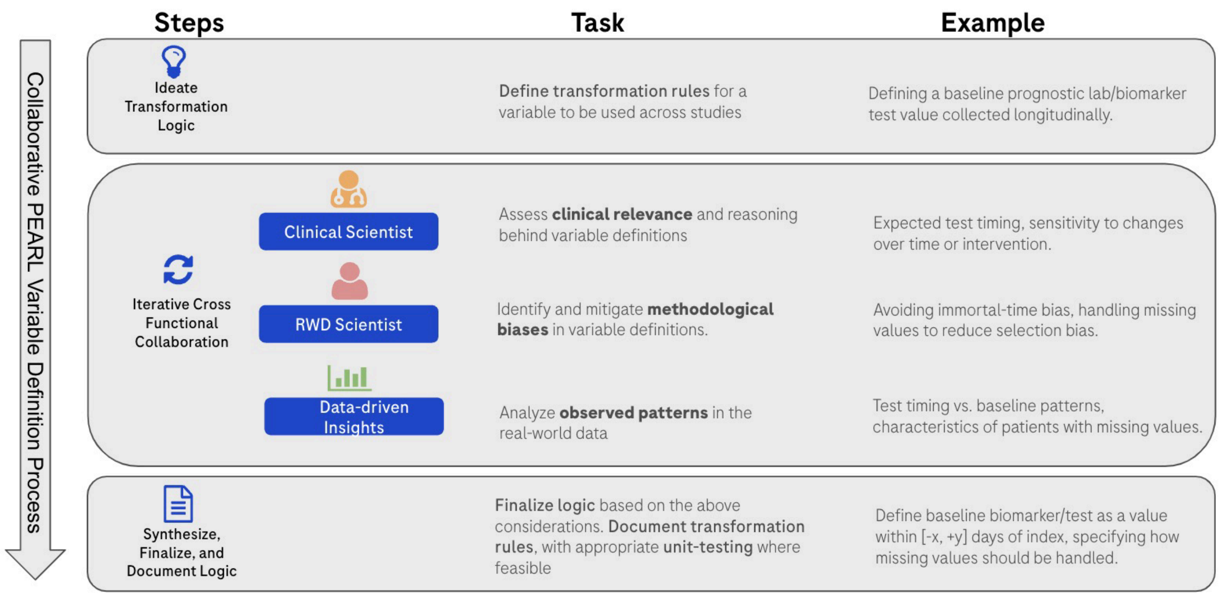
Developing RWD PEARL requires a synergistic approach that places collaboration, scientific rigor, and domain knowledge at the forefront. From the outset, teams must work in unison—clinicians, data scientists, statisticians, and AI agents—to identify which derived variables truly add value for real-world data analysis. By tapping into their collective expertise, organizations ensure that each variable or table reflects genuine clinical relevance, robust methodological underpinnings, and alignment with strategic research goals.

Documentation of each derivation forms the bedrock of this process and is the “secret sauce” that drives success over the long term. Documenting how and why a variable is created—from initial raw data points to final logic—allows teams to maintain transparency, instill confidence, and streamline future updates or reuses of the same derivations. While this might seem like additional work, it pays substantial dividends by enabling new team members, regulators, and collaborators to readily grasp how variables were generated, and under which assumptions.

The practical journey of building PEARL often begins with a funnel-like approach to selecting derived variables. At the widest point of the funnel, you consider a broad pool of potential variables. As you move downward, candidates are narrowed based on whether there is consensus on standard definitions, agreed-upon derivation logic, frequent usage in multiple studies, and relative stability (i.e., minimal need for ongoing changes). Only those that pass these thresholds become part of PEARL, ensuring that the included variables are meaningful, consistent, and widely applicable across different projects.



Once the funnel identifies which variables merit inclusion, intensive **cross-functional collaboration** kicks in. This step is where scientific rigor and domain knowledge shine. Teams discuss or debate raw data sources, transformation logic, any underlying evidence-based guidelines, and how the resultant variables fit into the broader analytics ecosystem. Input from clinical experts ensures that definitions align with real-world practice, while data scientists confirm technical feasibility and best practices for reproducibility. AI agents capable of automated data exploration may also guide decisions on how to shape derivations so that they are ready for advanced modeling or machine learning algorithms.

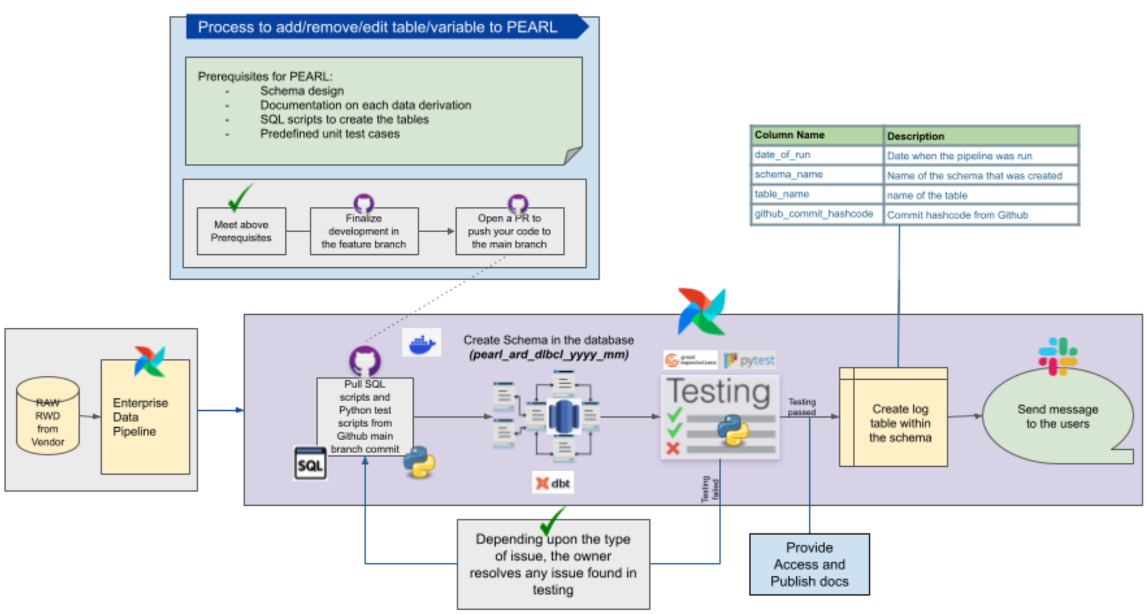


Training and capacity building are also critical components of the development process. Organizations should invest in upskilling their teams to ensure they are proficient in the tools and methodologies required for RWD PEARL development. This not only enhances efficiency but also fosters a culture of innovation and continuous improvement.

After consensus is reached, the **technical implementation** phase begins. Here, developers create or modify SQL scripts, Python code, and schema definitions in a feature branch within the organization's repository. Prerequisites—such as a clear schema design, thorough documentation of derivation steps, and predefined unit test cases—ensure that any new additions align with PEARL's overarching data model and maintain the high-level requirements for transparency and reproducibility. Teams then open a Pull Request (PR), where peers and subject matter experts can review the proposed changes, offer feedback, and validate the logic against existing PEARL standards.

Once the PR is approved and merged into the main branch, **automated pipelines** (often orchestrated by tools like Apache Airflow and dbt) take over. These pipelines pull the latest code from the repository, create or update the database schema (for instance, using a naming convention such as `pearl_ard_dbcl_YYYY_MM`), and run a suite of tests. Common testing frameworks include Pytest or Great Expectations, enabling automatic checks on data integrity, value distributions, and alignment with reference logic. If any anomalies are detected, the pipeline flags them, and the responsible developer makes the necessary fixes before re-running the tests. Once the tests pass, the system logs key metadata—such as the commit hash, run date, and newly created table name—in a dedicated log table for traceability. Users are then notified (via Slack or another messaging platform) that the updated PEARL data assets have been published and are ready for analysis.

Throughout this lifecycle, **iterative feedback** remains crucial. Users or downstream analytics teams might discover further refinements or additional derived variables that warrant inclusion in PEARL. Over time, the funnel process repeats as new candidate variables come into scope, pass through the collaboration phase, and enter the automated pipeline for validation. By uniting standardized definitions, thorough documentation, and automated testing, organizations can successfully build and maintain a trustworthy, AI-ready PEARL that reduces time-to-insight while preserving scientific rigor.



Best Practices for RWD PEARLs

Implementing RWD PEARLs effectively requires adherence to a set of best practices that ensure their usability, reliability, and scalability. One of the most important best practices is maintaining a clear framework for data traceability. By embedding metadata and automated logging at every stage of the data transformation process, researchers can ensure that every step is fully documented and reproducible. This traceability is critical for building trust in the data and facilitating its use in regulatory and clinical decision-making.

Another best practice is the focus on scientific rigor. This involves designing variables that address key challenges in real-world data analysis, such as missing data and confounders. Collaborating with therapeutic area experts is essential for ensuring that the variables are aligned with clinical practices and meet the specific needs of the research community. Additionally, the use of robust statistical methods and reproducible workflows helps to ensure the validity and reliability of the derived insights.

The design of RWD PEARLs should also prioritize reusability and scalability. By standardizing derivations and workflows, researchers can save time and reduce errors, enabling faster insights across projects. This standardization also makes it easier to integrate RWD PEARLs into existing analytics pipelines, enhancing their value as a research tool.

Looking to the future, RWD PEARLs should be optimized for next-generation analytics tools, including those powered by generative AI. These tools have the potential to revolutionize real-world data analysis by enabling more sophisticated exploratory analyses and predictive modeling. By designing PEARLs with these capabilities in mind, researchers can ensure that they remain at the forefront of scientific innovation.

Efficiency is another key priority for RWD PEARLs. By reducing redundant coding and streamlining workflows, they significantly enhance data accessibility and reduce the time required to generate insights. This efficiency not only benefits individual researchers but also contributes to the broader goals of the research community by enabling more organized and collaborative studies.

Challenges

Despite their many advantages, implementing RWD PEARLs is not without challenges. One of the biggest hurdles is aligning stakeholders. Ensuring buy-in from diverse teams—including data scientists, clinical experts, and IT professionals—requires clear communication and a shared vision for the value of RWD PEARLs.

Another challenge is the complexity of data sources. Real-world data often originates from disparate systems with varying formats, making integration a technically demanding process.

Standardizing these data sources while preserving their clinical nuances requires meticulous planning and execution.

Resource limitations also pose a challenge. RWD teams are often smaller than clinical trial teams but handle a significantly larger volume of questions. Developing RWD PEARLs requires upfront investments in technology, training, and personnel, which can strain limited resources.

Furthermore, maintaining data quality and accuracy across large datasets is a continuous challenge. Rigorous quality assurance processes are necessary to ensure the reliability of RWD PEARLs, but these processes can be time-intensive and require specialized expertise.

Summary

RWD PEARL represents a paradigm shift in the analysis of real-world data. By establishing a specialized, standardized, and efficient data layer, it addresses many of the challenges associated with traditional RWD analysis. The combination of scientific rigor, domain-specific expertise, and advanced technology ensures that RWD PEARLs deliver high-quality, analysis-ready datasets that empower researchers to generate actionable insights with confidence.

The collaborative BeeHive model underlying the development of RWD PEARLs fosters a sense of shared ownership and continuous improvement, ensuring that the datasets remain relevant and reliable. By reducing redundancy, enhancing traceability, and preparing for next-generation analytics, RWD PEARLs enable researchers to focus on answering pressing clinical questions rather than grappling with data preparation challenges.

In conclusion, RWD PEARLs democratize best practices in real-world data analysis, paving the way for scalable, sustainable, and impactful research. By bridging the gap between raw data and actionable insights, they represent a transformative solution for the life sciences community, empowering researchers to drive innovation and improve patient outcomes.

Case Studies:

Case Study 1: Real-World Prognostic Score (RoPRO)
PEARL enables the reproduction of prognostic scores using the RoPRO [4] model, ensuring consistency and transparency. By standardizing the complex model outputs and making them accessible across the organization, PEARL democratizes predictive analytics, allowing teams to leverage advanced prognostic insights without requiring specialized expertise in model development.

Case Study 2: Trial Pathfinder

Trial Pathfinder [5] leveraged ARDs created by PEARL to rapidly build and refine patient cohorts for clinical trials. By integrating standardized lab values, biomarkers, and clinical variables, it streamlined cohort selection, reducing manual effort and ensuring reproducibility. This approach improved trial feasibility assessments and accelerated patient matching across therapeutic areas.

References

[1] Dagenais, S., Russo, L., Madsen, A., Webster, J. and Becnel, L. (2022), Use of Real-World Evidence to Drive Drug Development Strategy and Inform Clinical Trial Design. Clin. Pharmacol. Ther., 111: 77-89. https://doi.org/10.1002/cpt.2480

[2] Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, Londhe A, Zhu V, Ryan PB. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553-64. doi: 10.1093/jamia/ocu023. Epub 2015 Feb 10. PubMed PMID: 25670757; PubMed Central PMCID: PMC4457111.

[3] CDISC, "CDISC SDTM, CDISC ADaM " [Online]. Available: https://www.cdisc.org/standards/foundational [Last Accessed Feb 2025]

[4] Becker T, Weberpals J, Jegg AM, So WV, Fischer A, Weisser M, Schmich F, Rüttinger D, Bauer-Mehren A. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Ann Oncol. 2020 Nov;31(11):1561-1568. doi: 10.1016/j.annonc.2020.07.013. Epub 2020 Jul 31. PMID: 32739409.

[5] Liu, R., Rizzo, S., Whipple, S. et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 592, 629–633 (2021). https://doi.org/10.1038/s41586-021-03430-5

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Navdeep Pal

Company: Genentech

Address: 1 DNA Way, South San Francisco, CA 94080

Work Phone: (650) 457-9380

Email: pal.navdeep@gene.com

Website: https://www.linkedin.com/in/navdeeppal/