Paper SD_08

Reproducible RWE at Scale

Agnes Youn, Domino Data Lab, San Francisco, CA, USA

ABSTRACT

Real-world evidence (RWE) is increasingly important for regulatory and clinical decision-making, yet many organizations remain constrained by static reports, siloed tools, and limited reproducibility. This demonstration shows how life sciences teams transform fragmented RWE workflows into governed, end-to-end insights engines. We connect to large-scale claims and EHR data, define cohorts, and run scalable analyses using versioned datasets and controlled environments. Along the way, we highlight how traceability from data to output ensures regulatory readiness without slowing innovation.

Attendees will learn how to:

- Scale analyses across millions of patient records while ensuring reproducibility
- Govern data, code, and environments for auditability and transparency
- Seamlessly move from exploratory analyses to regulatory-ready reporting
- Extend RWE with predictive models and interactive outputs

This approach unifies data, code, and environments into governed, reproducible pipelines, ensuring insights are both scalable and defensible in regulatory contexts.

OVERVIEW

Real-world evidence (RWE) is increasingly central to regulatory and clinical decision-making, yet most organizations are still fighting the same upstream battles: fragmented data pipelines, limited reproducibility, and environments that were never designed to scale with regulatory scrutiny in mind. There is no shortage of guidance on what decisions to make in an RWE program — which data standards to use, which pitfalls to avoid, which governance checkpoints to follow. What is missing is a framework for what your infrastructure needs to be capable of before any of that guidance becomes actionable at scale. This paper addresses that gap through four technical properties that any RWE infrastructure has to satisfy: data standardization and versioning, reproducible computational environments, traceable cohort logic, and automated submission-ready outputs. The goal is not to prescribe a specific platform, but to work through what principled RWE infrastructure looks like — and to make the case that the discipline the statistical programming community has already built for clinical trials is exactly the discipline RWE needs.

INTRODUCTION

Imagine this: a regulatory question arrives about an analysis completed eight months ago. The analyst pulls up the job record, which contains the exact version of the cohort definition used, the specific data snapshot it ran against, the computational environment — Operating System (OS), package versions, dependencies — and the full output log. The analysis re-runs as intended and produces identical results. The audit trail is complete because it was generated automatically as part of the workflow containing all the relevant metadata.

That scenario is achievable with current technology. However, it is not the norm. The gap between that scenario and most current RWE workflows is not primarily a scientific gap or even a technology gap — the tools exist. It is an infrastructure and governance gap: the organizational discipline to use those tools consistently, and the workflow design that makes reproducibility a structural feature rather than a best-effort aspiration. Most RWE guidance treats that gap as someone else's problem.

The regulatory momentum behind RWE is real. The FDA's Real-World Evidence Program and the 21st Century Cures Act have created genuine pathways for RWE to support label expansions, post-market surveillance, and in some cases primary efficacy. Pfizer's 2019 IBRANCE label expansion — using Flatiron Health EHR data from 600 male breast cancer patients to support an indication a prospective RCT could not feasibly have studied — is the most cited example of what is possible [FDA, 2019]. Medical devices have been ahead of therapeutics here; the 2017 FDA device RWE guidance established

frameworks that drug developers are still catching up to. But successful submissions have not made the underlying technical problem easier — and the bar for rigor is rising, not falling.
The rest of this paper is a working-through of what good infrastructure actually requires — not a checklist of decisions to make, but a framework for the capabilities your environment needs before those decisions are reproducible at scale. First, how to frame the technical problem against the FDA's regulatory questions; then, the four specific properties any RWE pipeline needs to satisfy.

BUILDING INFRASTRUCTURE THAT ADDRESSES THE RIGHT PROBLEMS

The FDA's 2018 RWE Framework organizes the evaluation of RWE submissions around three questions that should also organize your infrastructure decisions — because they define exactly what a submission needs to demonstrate, and therefore what a pipeline needs to produce.

Data fit for purpose requires demonstrating that your sources can reliably identify your population, capture your exposures, and measure your outcomes. This is a data engineering question as much as a scientific one. Immutable, versioned dataset snapshots tied to OMOP-standardized vocabularies make the fit-for-purpose argument auditable.

Study design adequacy requires that your cohort logic be transparent and executable — version-controlled, pre-specified, and independently reviewable. Decisions made before the analysis runs carry more regulatory weight than decisions reconstructed after the output exists.

Regulatory requirements compliance means your environment is controlled and your audit trail is complete and automatic. Every job, every output, every environment configuration needs to be tied to an immutable, traceable record. This is where most current RWE workflows break down — not because people do not care about compliance, but because the infrastructure was not built with this requirement as a first-class concern.

FOUR TECHNICAL PROPERTIES THAT ACTUALLY MATTER

1. Standardized, Versioned Data Assets

The OMOP Common Data Model has become the most viable path to cross-source consistency in RWE analysis. It provides a standardized vocabulary for conditions, drugs, procedures, and measurements that makes cohort definitions portable across databases and legible to regulatory reviewers. When inclusion criteria reference OMOP-standard concept sets rather than source-specific ICD or NDC codes, the definition is auditable and reusable across studies.

Standardization alone is not sufficient. Every analysis must be tied to a specific, immutable snapshot of the underlying data — the RWE equivalent of database lock. Without it, your cohort definition and code can be perfectly documented and still produce different results on re-run if the underlying source has been refreshed. Data quality profiling — systematic checks for completeness, consistency, and plausibility — needs to be a documented, first-class step in the pipeline, not an afterthought, as it is central to answering the FDA's fit-for-purpose question.

2. Reproducible Computational Environments

A script that produces different results depending on which package version is installed is not a reproducible analysis. In clinical trial programming, validated SAS environments with locked macro libraries solve this by construction. RWE analysis — spanning R, Python, SQL-based cohort tools, and increasingly ML frameworks — needs an equivalent: container-based environment management that captures the complete computational stack at the point of analysis and versions it alongside the code and data.

The challenge is organizational: establishing the norm that an analysis without a locked, reproducible environment is an incomplete analysis.

3. Traceable, Pre-Specified Cohort Logic

Cohort definition is where the most consequential methodological decisions in an RWE study are made, and where documentation is most often informal. Which ICD codes define the condition? How many claims confirm a diagnosis? What is the washout period? These choices materially affect results and need the same rigor as statistical methods — pre-specified, version-controlled, and peer-reviewed before the analysis runs.

Tools like ATLAS (OHDSI) provide structured interfaces for building OMOP-standard cohort definitions that are both machine-executable and human-readable. The regulatory precedents support this approach: the FDA's rare disease natural history guidance explicitly requires pre-specified, transparent methods, and the medical device precedents — MitraClip, SAPIEN TAVR — both succeeded in part because patient selection criteria were clearly documented upfront.

4. Automated, Submission-Ready Outputs

The largest efficiency opportunity in RWE infrastructure is eliminating the gap between analysis and documentation. In most current workflows, the analysis runs first and the submission package is assembled separately — introducing version mismatches, requiring reconstruction of informal decisions, and creating the conditions for the kind of reproducibility failure described at the outset of this paper.

A pipeline that generates audit trails, dataset lineage, environment provenance, and cohort documentation as automatic outputs of the analysis run eliminates this gap by construction. The compliance record is a byproduct of doing the work, not a separate effort layered on afterward.

THE GOVERNANCE LAYER: MAKING IT STICK

The four technical properties above are achievable. What prevents them from being the norm is not a technology gap — it is a governance gap. Governance frameworks are most effective when compliance is the path of least resistance: when data locking, environment provenance, and cohort documentation are structural features of the workflow that analysts do not have to remember to invoke.

Role clarity matters: who owns the fit-for-purpose data assessment? Who signs off on cohort definition review? Who is responsible for environment control? In clinical trial programming these roles are defined by SOPs refined over decades. In most RWE programs they are informal, which means they are inconsistent and do not survive deadline pressure or personnel turnover.

Governance champions — people who understand both the scientific and regulatory dimensions of RWE and can translate organizational norms into practice — are the mechanism by which infrastructure choices become sustained habits. Technology enables the practice; people establish and maintain it.

The specific application points worth prioritizing for the PHUSE community:

- Treat cohort definitions as analytical code: version-controlled, peer-reviewed, pre-specified before the analysis runs. The ICD code lists and inclusion/exclusion logic that define an RWE study population are as consequential as the statistical model.
- Extend SAS environment discipline to the full tool stack: R, Python, and SQL-based cohort tools need the same environment control that validated SAS environments provide. This is a governance expectation, not just a technical configuration.
- Identify governance champions early: people who can translate regulatory requirements into engineering norms and hold the line on reproducibility standards when timelines compress.
- Start where the regulatory risk/reward is clearest: label expansions for well-characterized populations, rare diseases where RCTs are impractical, and post-market safety surveillance. These use cases have the strongest regulatory precedent and offer the best environment in which to build and validate reproducible infrastructure before taking on more methodologically contested questions.

CONCLUSION

The technology required for reproducible, scalable RWE analysis is not the bottleneck — and it hasn't been for some time. Immutable data versioning, containerized environments, version-controlled cohort logic, and automated audit trails are all achievable with current tooling. The bottleneck is recognizing that these are engineering and governance requirements, not optional enhancements — the same way validated environments and independent QC are non-negotiable in clinical trial programming. The RWE field has been slow to make that shift. It doesn't have to be.

The PHUSE community is well-placed to lead this. The underlying discipline is the same; the application is new. The FDA's RWE program is expanding the scope of evidence it will accept, but the bar for rigor is rising with it. The organizations best positioned to take advantage of that expansion are those building reproducible, governed infrastructure now — before the regulatory question arrives, not after. If the

framing here resonates — or if you think the priorities are wrong — that's exactly the conversation this community should be having.

REFERENCES

FDA. (2019). FDA approves new treatment for certain advanced or metastatic breast cancers. https://www.fda.gov/news-events/press-announcements/fda-approves-new-treatment-certain-advanced-or-metastatic-breast-cancers

FDA. (2017). Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. https://www.fda.gov/media/99447/download

FDA. (2018). Framework for FDA's Real-World Evidence Program. https://www.fda.gov/media/120060/download

FDA. (2019). Rare Diseases: Natural History Studies for Drug Development. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-natural-history-studies-drug-development-guidance-industry

FDA. Real-World Evidence Program. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

Franklin, J. M., & Schneeweiss, S. (2017). When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clinical Pharmacology & Therapeutics, 102(6), 924–933. https://pubmed.ncbi.nlm.nih.gov/28817160/

OHDSI. The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/

Tendler, M. (2025). Staying Out of the Ditch — How Governance Can Provide Your Team Guardrails. PHUSE Connect Paper DS17. Domino Data Lab.

CONTACT INFORMATION

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

Author Name: Agnes Youn
Company: Domino Data Lab
Email: agnes.youn@dominodatalab.com
Website: https://domino.ai