

Trusting the Evidence: Objective Frameworks to Ensure Data is “Fit-for-Purpose” for Clinical Discovery

Yong Chen, PhD¹

¹Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania.

Contact: ychen123@upenn.edu



Penn Medicine

Background

- Real-world data can accelerate AD/ADRD research at a scale not possible with trials or cohort studies alone.
- Common RWD sources include electronic health records, CMS claims, and deeply phenotyped aging cohorts.
- However, RWD is often observational, heterogeneous, incomplete, and prone to bias.
- The same dataset can support either credible or misleading findings depending on whether it is appropriate for the research question.
- Therefore, the key question is not only “What does the data say?” but “Is this data fit for the purpose at hand?”

Objective

- Define data readiness across multiple dimensions before evidence generation.
- Demonstrate how the ReCARDIO Research Network operationalizes this framework for AD/ADRD research.
- Shift trust from a general claim to an auditable, evidence-based property of the data ecosystem.

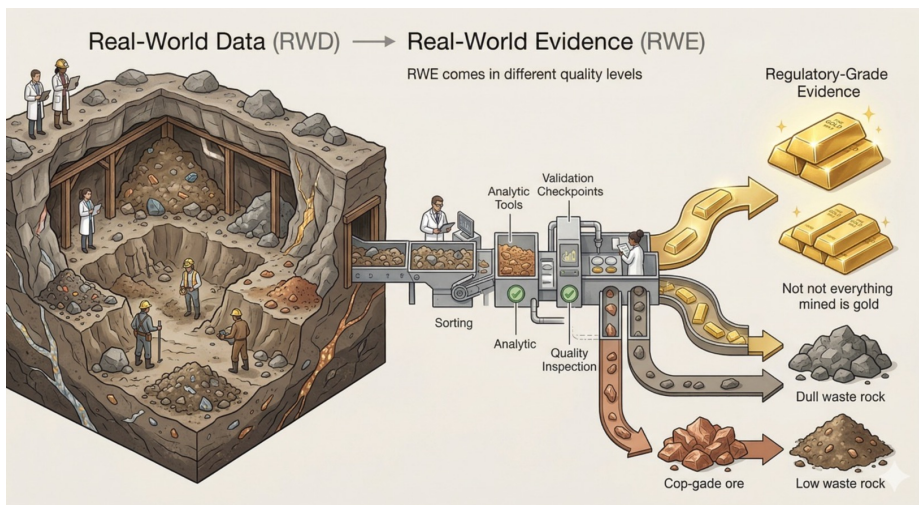


Figure 1. From Real-World Data to Regulatory-Grade Real-World Evidence: A Fit-for-Purpose Quality Pipeline

Core Principles

- Declare before you discover:** Define readiness criteria and thresholds before examining results.
- Diagnose, don't just describe:** Each dimension should produce an explicit pass/fail or “inestimable” verdict with a reason.
- Treat trust as a lifecycle property:** Readiness must be tied to specific data versions, standards, and provenance records.
- Use sequenced assessment:** Earlier readiness dimensions must be satisfied before later analytic conclusions are meaningful.

Conceptual Framework: Five Dimensions of Data Readiness

Dimension	Question it answers	What is verified (pre-specified)	ReCARDIO realization
D1 Structural	Is the data present, current, and traceable enough to use at all?	Completeness, coverage of the target population, recency, and end-to-end provenance of every record and transformation.	Ingestion QA/QC integrated with the data pipeline; NHash provenance identifiers.
D2 Semantic	Do variables mean the same thing across sources?	Harmonized definitions, permissible values, and units; explicit, auditable mappings to shared standards.	CDE-Engine, CDE Knowledge Graph, and curated AD-CDEs in seven domains.
D3 Relational / temporal	Are records linked and time-aligned so the question is answerable?	Record linkage at the individual level and temporal alignment of events to common reference points.	RWD2RWE Phase I linkage and temporal alignment; CDE-Data.
D4 Analytic	Does this data support this specific question without disqualifying bias?	Sufficiency, comparability, and bias assessed against criteria declared before results are seen.	Systematic readiness / fitness-for-purpose diagnostics in CDE-Data.
D5 Sustained	Will the verdict still hold as data and standards evolve?	Versioning, change tracking, and monitoring for drift relative to the validated population.	Ecosystem versioning, provenance chains, quarterly CDE releases.

Figure 2. Five Dimensions of Data Readiness

ReCARDIO as the Operational Use Case

- ReCARDIO is a ten-institution, NIA-funded U24 network focused on deriving AD/ADRD Common Data Elements from RWD.
- ReCARDIO's RWD2RWE pathway translates real-world data into real-world evidence.
- The framework is implemented through two linked ecosystems:
 - CDE-Engine: supports CDE development and semantic harmonization.
 - CDE-Data: supports extraction, readiness diagnostics, provenance tracking, and evidence generation.
- Together, these systems operationalize the five readiness dimensions across the full data lifecycle.



Figure 3. Collaborating sites of the ReCARDIO meeting

How ReCARDIO Implement Five Dimensions

- Structural:** CDE-Data integrates ingestion QA/QC into the data pipeline and tracks provenance using NHash identifiers.
- Semantic:** CDE-Engine and the CDE Knowledge Graph harmonize source data elements into standardized AD/ADRD CDEs.
- Relational / temporal:** RWD2RWE Phase I links individual-level records and aligns events to common reference points.
- Analytic:** CDE-Data evaluates sufficiency, comparability, and bias using pre-specified readiness diagnostics.
- Sustained:** ReCARDIO maintains versioning, provenance chains, quarterly CDE releases, and drift monitoring.

Conclusion and Take-Home Message

- Fit-for-purpose assessment should be explicit, pre-specified, sequenced, and evidence-based.