RWD Signal Integrity Modelling

A Framework for Evaluating the "Health" of Real-World Data Before Use

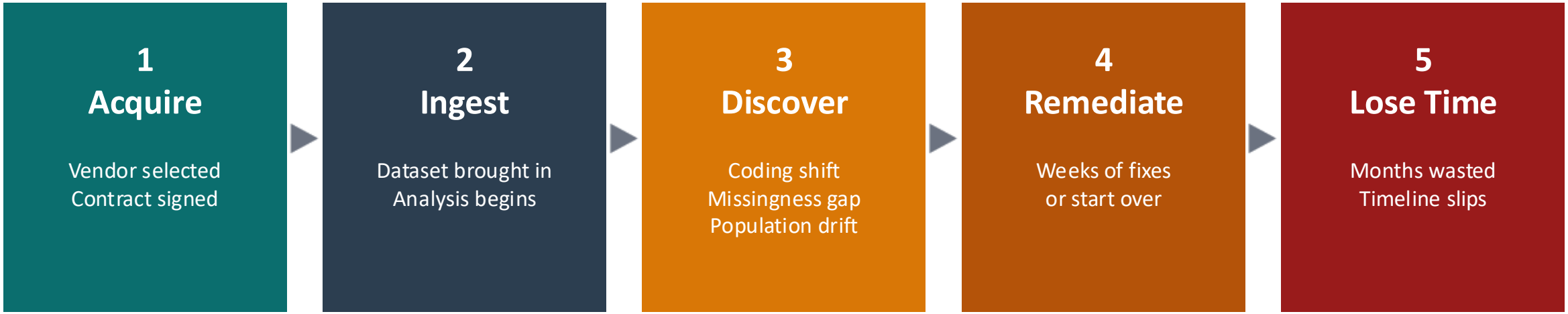
Yogesh Kumar Gupta · Independent Researcher

PHUSE Real World Data Autumn Event 2026 · bioyog@gmail.com · ORCID 0009-0001-2635-3619

Disclosure

Conflicts of interest	None to declare.
Funding	No external funding. This is independent research.
Institutional affiliation	Views expressed are my own and do not represent any employer or institution.
Framework status	At proposal stage. I will be explicit about what is established and what is still being developed.
Published basis	Related peer-reviewed work: BATPC (Current Problems in Cancer, 2026, PMID 42435516) and PNDI (Diagnosis, 2026, PMID 42446947).

The Reactive Cycle — A Pattern Most of You Will Recognise



"The team discovers the problem halfway through the analysis — not before committing to the dataset."

This reactive cycle is not rare. It is the default.

The Cost of Finding Out Late



Timeline slip

Weeks to months of remediation after commitment. Alternative datasets deprioritised.



Weak submission rationale

"We chose this dataset because it was available." Not a defensible regulatory answer.



Sunk analytical cost

Two months of team effort spent on the wrong dataset before the problem surfaces.



Reactive rather than proactive

Problems addressed when discovered — not anticipated, not prevented.

A Rich Landscape — With a Shared Blind Spot

Field-Level QC Tools	Regulatory Frameworks	Distributed Networks
✓ OMOP Achilles	✓ FDA RWE Framework (2018)	✓ FDA Sentinel System
✓ Data Quality Dashboard (DQD)	✓ FDA Considerations Guidance (2023)	✓ EHDEN (European)
✓ PCORnet Curation Reports	✓ EMA Data Quality Framework (2023)	✓ DARWIN EU (EMA)
	✓ ISPOR-ISPE Joint Task Force	
	✓ GRACE Principles	

All of this is valuable. None of it is wrong. But they share one characteristic.

The shared limitation across every existing approach:

Principles-based, not scoring-based.

They tell you what dimensions of data quality should be considered.
They do not compute it. They do not score it. They do not compare datasets.

Field-level checks run after ingestion — not before commitment.
That is the operational gap this framework is designed to fill.

Construction-Layer Transparency — Where This Framework Sits



Digital health precedent

Regulators already expect construction-layer transparency for sensor data and digital endpoints — how raw signals become analytical artefacts.

The gap for RWD

Real-world data undergoes equally complex transformation: coding updates, workflow changes, population drift, vendor processing. No equivalent transparency framework exists.

What this framework provides

A structured, reproducible method for evaluating dataset health at the construction layer — before ingestion, before analytical commitment.

RWD Signal Integrity Modelling — Framework Architecture

C1	C2	C3	C4	C5
Pre-Analysis Integrity	Temporal Shifts	Missingness Patterns	Bias Signatures	Source Reliability

Each scored 1–5

Signal Integrity Score = (Sum of five component scores ÷ 25) × 100 → 0 to 100

80–100	60–79	40–59	< 40
High Integrity	**Moderate**	**Low Integrity**	**High Risk**
Primary regulatory submission	Supportive analyses	Substantial remediation required	Reject or replace
✓ Accept	~ Proceed with care	⚠ Remediate	✗ Reject

Component 1: Pre-Analysis Integrity Score



C1

PREDICTS

Overall analytical reliability — the headline indicator of dataset health.

"Analogous to a credit score — the headline indicator before deeper diligence."

The "first impression" check. Aggregates six dimensions:

Completeness	Are required fields populated across the study population?
Consistency	Do values agree across linked tables and time periods?
Timeliness	Is data delivery cadence aligned with study timeline?
Stability	Do dataset characteristics hold across refreshes?
Documentation	Is metadata complete, current, and accurate?
Usability	Can the dataset be ingested without extensive remediation?

Component 2: Temporal Shifts Analysis



C2

PREDICTS

Longitudinal validity and AI/ML model stability across time.

"Think of the ICD-9 to ICD-10 transition — or COVID-era diagnoses appearing in 2020, altering the disease mix of every dataset spanning that period."

The most common and least detected failure mode in RWD. Evaluates:

Coding shifts	ICD-9→ICD-10 transitions, SNOMED updates, CPT revisions
Workflow shifts	Changes in how clinical concepts are documented over time
Population drift	Changes in patient demographics or disease mix over time
System signal shifts	Changes in how source systems generate the underlying data

Component 3: Missingness Patterns



C3

PREDICTS

Risk of bias, imputation fragility, and reduced statistical reliability.

"Missingness concentrated in patients over 65, in a study where that is your target population, is not a nuisance — it is a fundamental threat to the analysis."

Missingness in RWD is not noise — it is a signal about how data was generated.

Random missingness	Unpatterned absence — least dangerous, most manageable
Structured missingness	Template-driven gaps — specific fields systematically empty
Systematic missingness	Demographic- or provider-specific absence — most dangerous

Component 4: Bias Signatures



C4

PREDICTS

Regulatory rejection risk, fairness concerns, generalisability limits.

Evaluates six structural bias patterns that affect representativeness:

Demographic bias	Systematic over- or under-representation by age, sex, race, ethnicity
Geographic bias	Concentration in specific regions, excluding rural or non-US populations
Socioeconomic bias	Commercially insured datasets under-representing Medicaid, uninsured
Provider-level bias	Specialist-capture datasets missing primary care patient journeys
Sampling bias	Non-random selection creating a patient mix unrepresentative of the target
Survivorship bias	Exclusion of patients who left the system before the observation window

Component 5: Source Reliability Profile



C5

PREDICTS

Reproducibility, audit readiness, and long-term dataset trustworthiness.

"A dataset is only as reliable as the organisation producing it. This component captures what field-level QC cannot — the trustworthiness of the source."

Evaluates the data provider — not just the data itself:

Operational stability	Is the vendor organisation stable and well-resourced?
Documentation quality	Does the vendor produce clear, current data documentation?
Historical consistency	Do past deliveries match what the vendor committed to deliver?
Vendor transparency	Does the provider disclose its data-generating processes?
Update frequency	Are dataset refreshes timely and predictable?

Scoring Model — Simple by Design

Signal Integrity Score = (Sum of five component scores ÷ 25) × 100

Each component scored 1–5 · Maximum raw score: 25 · Composite: 0–100

✓ 80 – 100	**HIGH INTEGRITY**	Suitable for primary regulatory submission, pivotal trials, and high-stakes AI/ML development.
~ 60 – 79	**MODERATE**	Suitable for exploratory or supportive analyses with specific limitations documented.
⚠ 40 – 59	**LOW INTEGRITY**	Use only with substantial remediation. Not suitable for regulatory submission without targeted improvement.
✗ < 40	**HIGH RISK**	Recommend rejection or replacement. If used, all limitations must be explicitly disclosed.

Simplicity is intentional. A score too complex to compute is a score nobody computes.

Worked Example — Two Candidate Datasets, Same Field-Level QC Result

Component	Dataset A	Dataset B	Difference
Pre-Analysis Integrity	4	2	-2
Temporal Shifts	5	3	-2
Missingness Patterns	4	2	-2
Bias Signatures	4	2	-2
Source Reliability Profile	5	3	-2
Signal Integrity Score	**88**	**48**	**-40**

Both datasets passed conventional field-level QC — OMOP Achilles would give both a thumbs-up.

What the Score Changes — From Reactive to Proactive

Dataset A **Score: 88** **HIGH INTEGRITY**	**DECISION: SELECT** Study proceeds on stable, well-documented, representative data. Regulatory rationale is structured and archived.
Dataset B **Score: 48** **LOW INTEGRITY**	**DECISION: REJECT before ingestion** Two months of downstream remediation prevented. Alternative dataset evaluated proactively. No sunk analytical cost.

The score and supporting justification are archived — creating a structured, auditable rationale for dataset selection.

Regulatory Alignment — Built In, Not Bolted On

Regulatory expectation	Framework component
Data provenance and traceability	**Source Reliability Profile**
Population representativeness	**Bias Signatures**
Metadata transparency	**Pre-Analysis Integrity + Missingness Patterns**
AI/ML explainability and stability	**Temporal Shifts + Bias Signatures**
Reproducibility and audit readiness	**Pre-Analysis Integrity + Source Reliability**

A sponsor who archives a Signal Integrity Score alongside their dataset-selection decision has produced a structured, auditable rationale at submission time.

Limitations — Three I Want to Put on the Table

1

Operational definitions are in development

The 1–5 scoring anchors need explicit definitions for what constitutes a 3 versus a 4 on each component. Without those, scoring is structured judgment — not yet a reproducible rubric.

Mitigation: Supplementary scoring rubric with worked examples from common RWD source types is the immediate next deliverable.

2

Applied validation has not been completed

I have not yet scored Optum, MarketScan, or Sentinel with this framework and reported the results. The framework as presented is the conceptual architecture — not yet a validated instrument.

Mitigation: Applied validation against named public RWD sources is the natural next phase of work.

3

Inter-rater reliability has not been established

If two assessors score the same dataset independently, do they arrive at similar composite scores? This is a property the framework needs to demonstrate before operational adoption.

Mitigation: Inter-rater reliability study with at least two independent assessors is on the roadmap.

What's Next — A Four-Stage Research Agenda

Stage 1

Operational definitions

Complete the 1–5 scoring rubric for all five components with worked examples drawn from EHR, claims, registry, and linked datasets.

Immediate

Stage 2

Applied validation

Score named public RWD sources — Optum, MarketScan, CPRD, Sentinel — and report the results. Test whether the composite score predicts downstream analytical failures.

Next 6 months

Stage 3

Inter-rater reliability

At least two independent assessors score the same datasets. Target weighted kappa ≥ 0.70 across all five components before recommending operational adoption.

6–12 months

Stage 4

Multi-stakeholder consensus

Modelled on how CONSORT and SPIRIT developed — a working group process to establish a recognised minimum reporting standard for RWD source documentation.

12–24 months

Collaboration welcome — especially on applied validation and operational definitions. Contact details on the next slide.

One sentence about what this framework is trying to do:

"Real-world data integrity is currently evaluated reactively,
after analytical investment has begun.
This framework evaluates it proactively, before commitment."

The point is not a perfect scoring system.
The point is that a structured, reproducible answer
is better than the answer we work out three months later.

Yogesh Kumar Gupta · bioyog@gmail.com · ORCID: 0009-0001-2635-3619

Published frameworks: zenodo.org (BATPC, PNDI, and more) — all CC-BY 4.0

Happy to take questions — and genuinely interested in collaborators.