RegenMed

WHITE PAPER

CIRCLE DATASETS AS GROUND TRUTH FOR HEALTHCARE AI

October 2025

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Problem

Artificial intelligence (AI) has entered medicine with extraordinary promise—and deep fragility. From diagnostic imaging to population health analytics, thousands of algorithms have been published, yet only a fraction deliver reliable results in clinical practice. The reason is fundamental: *AI in healthcare has outgrown its data foundation.*

Most medical AI models are trained on unstructured or opportunistic data extracted from electronic health records, claims systems, or research silos. These sources were never designed for scientific reproducibility, regulatory compliance, or continuous learning. They are riddled with bias, missingness, and unverifiable provenance. The result is a widening gap between technological potential and clinical trust.

The Insight

If healthcare AI is to achieve clinical reliability, its foundation must change—from uncurated data to **structured, validated, longitudinal, and interoperable real-world evidence**. That is the core insight behind **RegenMed's Circle Datasets**.

Each Circle Dataset is built through a predefined **Observational Protocol (OP)**—a ruleset that governs what data are collected, when, and under what clinical context. Captured through RegenMed's **inCytes™** (clinician-facing) and **Benchmarc™** (patient-facing) systems -- and supported by burden-minimizing **Processes** -- these data form a complete longitudinal record linking interventions, outcomes, and follow-up observations.

Every data element carries full **provenance metadata**—who entered it, when, under which protocol version—creating a dataset that is both scientifically reproducible and ethically auditable.

The Solution: Circle Datasets

Circle Datasets transform raw clinical data into what AI has always needed but rarely had: **ground truth**. They are built from the ground up to be:

- **Structured:** protocol-driven and mapped to standard terminologies (ICD, CPT, LOINC, SNOMED, FHIR).
- **Validated:** checked through clinician review and automated logic.

- **Longitudinal:** tracking patients and conditions over time for causal understanding.
- **Interoperable:** harmonized across institutions and systems.
- **Federated:** locally controlled yet globally comparable, preserving privacy and institutional autonomy.

Together, these qualities enable AI models to learn from *verified reality* rather than opportunistic noise.

The Scientific Foundation

Circle Datasets are derived from structured Observational Protocols – which may be IRB- or MEC-approved -- and aligned with FDA, EMA, and other standards for real-world evidence and trustworthy AI. They embody the same methodological rigor as clinical trials but operate continuously in clinical practice. This produces statistically significant real-world datasets which are:

- Auditable for regulators.
- Replicable for researchers.
- Clinically meaningful for practitioners.

Circle Datasets support the principles of *reproducibility, transparency, and accountability*—the core criteria for trustworthy medical AI.

The Ecosystem

Each Circle contributes to a federated network of participating institutions and independent providers —a **Circle Ecosystem** where data quality improves as participation grows.

- Clinicians receive feedback through benchmarking dashboards.
- Researchers gain access to de-identified longitudinal cohorts.
- Regulators obtain continuous, auditable real-world evidence.
- Patients benefit from safer, explainable, and continuously improving AI tools.

This creates a self-reinforcing feedback loop: better data → better models → better care → broader participation → still better data.

The Future

Healthcare is entering a phase where *data integrity* will matter more than *model novelty*. RegenMed's Circles Platform establishes the infrastructure for that era—a verifiable, federated, and continuously learning evidence system that bridges human expertise and machine intelligence. **Circle Datasets** are not another data repository; they are the **scientific substrate** for trustworthy AI in medicine. They turn artificial intelligence into accountable intelligence.

THE COMING RECKONING FOR AI HEALTHCARE MODELS

Artificial intelligence is poised to reshape healthcare at a scale and speed unlike prior technology. Advances in machine learning, data science and algorithmic reasoning promise to unlock new insights, automate complex decision-processes, and deliver more personalized, efficient, and cost-effective care.

Yet this promise belies a fundamental tension: while the tools have become more capable, the underlying data and infrastructure that power them remain deeply inadequate for clinical use.

A reckoning is underway in medicine and healthcare more broadly — one in which the mismatch between AI's aspirations and healthcare's realities will define which tools succeed, which fail, and how patients ultimately benefit.

The False Dawn

Healthcare systems globally are under pressure: aging populations, chronic disease prevalence, clinician shortages, rising costs and value-based care mandates are forcing innovation. In this context, AI emerges as a seductive strategic lever. The growth in computing power, availability of large datasets, and increased regulatory openness have fueled AI's expansion into healthcare. [1] For many stakeholders—clinicians, executives, researchers—the question is no longer *if* AI will be applied in healthcare, but *when* and *in what form.*

A recent U.S. survey across 43 health systems found that while 90 % reported at least partial deployment of AI in imaging/radiology. However, only 38 % judged that AI for risk stratification achieved high success. [2] These findings reflect that adoption alone is insufficient; meaningful clinical impact remains elusive. This discrepancy arises because of the persistent gap between techno-promise and trustworthy, clinically-effective implementation.

High Expectations v. Fundamental Weaknesses

With AI advancing faster than the data ecosystem, there is already evident a convergence of expectations and risks. Multiple reviews highlight that AI tools in healthcare raise concerns of bias, inequity, lack of transparency, patient safety, and governance. [3] One study identifies that stakeholder fears about AI include "faulty and manipulated data", "inadequate machine learning", and resultant "bias and discriminatory services." [4]

At the same time, regulatory agencies, legislatures, payers and institutional leaders are applying more scrutiny to claims of AI efficacy, reproducibility and explainability. The gap between initial optimism and real-world penetration is at an inflection point.

If AI is to become mainstream in healthcare rather than a niche novelty, then the underlying inputs and workflows must shift away from opportunistic, ad-hoc data and algorithms, and towards structured, longitudinal, auditable systems. Without that shift, many AI initiatives will be shelved, repurposed, or exposed for low impact — and the negative consequences for patients, providers and investors will be profound.

More Data ≠ Better Care

The Illusory Value Of "Big Data" RWE

The dominant narrative has been that *more* data equates to *better* AI in medicine. Indeed, large-scale EHR (electronic health record) exports, claims data sets, imaging archives and device logs have been assembled with that expectation. However, decades of attempts to leverage high-volume data for clinical decision-support have repeatedly stumbled.

Observers note that data quality, context, annotation, provenance and clinical relevance often fail to meet necessary levels. The oft-quoted "garbage in, garbage out" maxim remains a central barrier to AI's promise. [5] A systematic review described data-management and ethical/regulatory hurdles as key obstacles to AI adoption. [6]

Put differently, the quantity of data has increased, but the *signal-to-noise ratio* remains poor. Clinical data is often opportunistic (entered for operations or billing), unstructured (free text, imaging, logs), inconsistently coded, undocumented for research commondenominators, and may lack longitudinal context. This mismatch diminishes the value of AI outputs and undermines the intuitive "big-data" narrative in medicine.

Why Quantity Alone Is Insufficient

While the narrative of "big data wins" has persisted, multiple reviews have exposed its limitations in healthcare AI. A rapid umbrella review of AI quality standards emphasized that issues such as missingness, bias risk, inconsistent data fusion, and lack of metadata documentation undermine model trust. [7] A dedicated analysis of dataset documentation ("Datasheets for Healthcare AI") argued that even when large datasets exist, insufficient metadata, incomplete provenance, and weak documentation restrict their safe use in clinical AI. [8]

If the input data lacks defined provenance, structure, completeness and longitudinal context, then downstream AI models—even those with advanced architectures—are liable to produce unpredictable or unsafe outputs.

Key AI Failure Modes Driven By Poor Data

Primary failure modes of AI in healthcare comprise several interlinked issues:

Opportunistic Versus Protocol-Driven Capture

Much healthcare data is captured for operational, billing or documentation purposes—not explicitly for research or AI training. That means the data lacks the consistency, completeness or context needed for rigorous modelling. For example, data entries may be missing standardized codes, unstructured text dominates, or longitudinal follow-up is incomplete.

Lack of Provenance And Auditability

When the origin of a data point cannot be traced (who captured it, when, under what protocol) then the chain of custody is broken. This makes it difficult to assess data reliability, correct for bias, or establish audit trails—risks that are increasingly material in regulated environments.

Staleness, Incompleteness And Fragmentation

Healthcare data often lacks standard refresh cadence, may skip key time-points (visits dropped, patients lost to follow-up), and is distributed across fragmented systems. A recent review found that although interest in AI has grown, "research in AI and quality in health care has failed to keep up" because of dataset limitations, fragmented access, and structural bias.

Bias, Generalizability And Lack Of Longitudinal Context

AI models rely on training data that reflect both the target population and the temporal dynamics of disease and treatment. Without longitudinal data following the same patient over time, models may mis-predict progression or outcomes. Moreover, when datasets under-represent specific populations or therapeutic contexts, bias becomes embedded—e.g., a widely used algorithm under-estimated health needs of Black patients due to reliance on cost as a proxy.

Interoperability And Domain-Semantics Issues

When data sources use inconsistent terminologies (ICD vs local codes, CPT mapping gaps, LOINC/test result variations) or lack FHIR alignment, it becomes challenging to merge data streams, define structured features, or maintain semantic consistency across sites or over time.

The place to focus is not simply "collect more data" but to ensure that the data is:

- captured under defined protocols,
- longitudinally consistent,
- auditable and traceable,
- coded to standard terminologies, and
- refreshed and maintained over time.

Regulatory, Ethical and Operational Implications

The foregoing data issues translate into practical risks and barriers:

- **Patient safety risk**: AI tools built on weak data foundations may propagate erroneous or biased outputs—potentially harming patients or worsening disparities. A recent review highlighted "unpredictable errors, inadequate policy and regulatory regime" in the use of AI tools in healthcare. [9]
- **Transparency and trust**: Clinician and patient acceptance of AI is strongly tied to transparency, explainability and trustworthiness. If the dataset underlying a model cannot be audited or rationalized, adoption stalls.

- **Regulatory liability**: As lawmakers and regulators push for AI standards in healthcare, data provenance, traceability and audit trails become mandatory. Models developed with opaque datasets are exposed to higher regulatory friction or failure.
- **Operational inefficiency**: When data is missing, inconsistent or poorly structured, AI development cycles lengthen, require heavy cleaning, and cost more—diminishing ROI and delaying impact.

The Impending Inflection Point: Quality Data Will Replace Big Data

Healthcare AI initiatives are at a pivot point. They must move beyond acquisition of large volumes of generic, low-quality data toward deployment of **high-integrity datasets** -- those with structure, longitudinal follow-up, provenance and semantics built-in. Without this pivot, many AI projects will continue to underperform, face regulatory pushback, and fail to generate trusted clinical impact.

The next section will describe how the RegenMed Circles platform confronts this gap by delivering datasets constructed under well-structured Observational Protocols, validated for auditability, mapped for interoperability, and designed for longitudinal fidelity.

THE POWER OF CIRCLE DATASETS

From Opportunistic Data To Protocol-Driven Evidence

Most clinical data available for artificial intelligence (AI) development originate from electronic health record (EHR) systems, claims databases, or device logs created for operational rather than research purposes. These sources, while vast, are opportunistic and inconsistent. They lack uniform data capture rules, defined follow-up intervals, or standardized outcome measures—features necessary for trustworthy AI training and validation. Their quantity is high; their quality is low.

Circle Datasets, by contrast, are constructed under **Observational Protocols** that specify precisely what data to collect, when, and how. Each OP represents a structured observational study design aligned with clinical intent (e.g., rehabilitation after knee arthroplasty, progression of diabetic neuropathy). This converts passive data collection into an active, rule-governed process similar in rigor to a clinical trial, but implemented within real-world practice.

The result is a dataset that is **structured**, **validated**, **longitudinal**, and **interoperable**—in effect, a real-world evidence (RWE) substrate suitable for both clinical research and AI validation.

The Circles Platform Architecture

Circle Datasets are generated through the patented **RegenMed Circles Platform**, which integrates two complementary modules:

- inCytes™ – a clinician-facing platform for standardized, protocol-based data capture at the point of care. It guides physicians through structured input forms aligned with observational protocol parameters, ensuring consistency across sites and practitioners.
- Benchmarc™ – a patient-facing system that collects longitudinal outcomes, patient-reported measures, and follow-up data through secure, consented workflows.

These are supported by Processes designed to minimize burden and maximize ethical compliance incentives for clinicians and patients alike. The Circles Platform maintains linkage between physician-entered clinical data and patient-entered outcomes, producing a verified longitudinal chain. Each data record carries metadata defining capture context, time-stamp, and validation status.

This architecture ensures **end-to-end provenance**—a traceable lineage from initial data entry through every transformation, audit, and export.

Core Attributes Of Circle Datasets

Attribute	Definition	Impact
Structured	Data elements defined by observational protocol schema, mapped to standard terminologies (ICD-10, CPT, LOINC, SNOMED CT).	Enables consistency, interoperability, and analytic precision.
Validated	Multi-layer verification: automated logic checks, clinician sign-off, and cross-site audits.	Reduces noise, mis-labeling, and drift; supports regulatory confidence.

Attribute	Definition	Impact
Longitudinal	Repeated measures over defined time horizons, correlated to specific pathologies, treatment protocols, and patient cohorts.	Captures disease progression, treatment response, and causal context.
Interoperable	FHIR-compatible data objects; API-accessible via standardized schemas.	Facilitates multi-institutional studies, AI model training and benchmarking.
Provenance-enabled	Every element carries metadata for origin, transformation, and audit history.	Provides traceability essential for regulatory and ethical compliance.

Together, these characteristics elevate Circle Datasets from mere data collections to *regulatory-grade, real-world evidence assets.*

Design Philosophy: Reproducibility, Auditability, And Transparency

The Circles approach is shaped by lessons from translational research and AI model evaluation frameworks -- e.g., FDA's Good Machine Learning Practice guidelines [10], NIST AI Risk Management Framework [11], and EMA RWE methodology guidance [12]. Across these standards, three qualities recur: **reproducibility**, **auditability**, and **transparency**.

- **Reproducibility** demands that results derived from data can be independently verified by others using the same dataset and definitions. Circle Datasets achieve this through locked data dictionaries and protocol-version control.
- **Auditability** ensures every transformation—from ingestion through analysis—is logged and attributable. This is enforced through immutable metadata and periodic third-party audits.
- **Transparency** involves explainable provenance: users can trace outcomes back to the exact data points and capture contexts from which they were derived.

By embedding these design principles natively, the Circles Platform generates datasets which can serve as *reference-grade materials* for validating and benchmarking AI models—particularly those claiming clinical utility.

Clinical And Technical Alignment

The structure of Circle Datasets aligns closely with major interoperability and RWE standards:

- **FHIR (R4) Resources:** Observation, Condition, Procedure, Medication Request, Encounter, and Patient. (See Appendix A.)
- **Terminologies:** ICD-10 for diagnosis, CPT for procedures, LOINC for lab tests, SNOMED CT for clinical findings, PROMIS and other standardized outcomes assessments.
- **Data quality metrics:** Completeness, conformance, plausibility, timeliness (consistent with OHDSI and FDA Sentinel standards [13].
- **Security and compliance:** Data encrypted at rest and in transit, governed under HIPAA and GDPR principles; access managed via federated identity.

These alignments ensure that Circle Datasets can integrate with external research environments, regulatory submissions, and machine-learning pipelines without conversion loss.

Comparative Advantage

Compared with uncurated EHR exports or secondary data aggregations, Circle Datasets offer several measurable advantages:

- **Signal-to-noise ratio:** structured capture under protocol reduces uninformative entries.
- **Temporal continuity:** longitudinal follow-up captures trajectories rather than static snapshots.
- **Causal interpretability:** consistent data across time points supports causal inference methods.
- **Regulatory confidence:** provenance and audit logs simplify evidence submission.
- **Cross-institutional reproducibility:** standardized schema ensures replicable results across sites.
- **Far lower costs, and greater coverage across all health and wellness domains.**

These advantages collectively define Circle Datasets as *ground-truth* materials for healthcare AI—data that can verify or falsify model predictions against real, reproducible evidence.

The Scientific Integrity of Circle Datasets

Clinical-Grade Evidence In Real-World Form

Scientific integrity in healthcare data means more than accuracy—it encompasses *rigor, reproducibility, and clinical relevance*. Traditional randomized controlled trials (RCTs) achieve this through protocolization, blinding, and controlled environments. But they are limited in scope, cost, and timeliness.

Circle Datasets extend that rigor into the *real world* by embedding structured observational protocols directly into clinical workflows. Each dataset is derived from collaboration with one or more physicians expert in their field – overseen if desired by an Institutional Review Board (IRB) – resulting in a structured **Observational Protocol (OP)**. That OP is then implemented across multiple clinical sites anywhere in the world through RegenMed's **inCytes™** and **Benchmarc™** platforms.

This approach operationalizes what regulatory science increasingly refers to as **"regulatory-grade Real-World Evidence (RWE)"**—data suitable for clinical, regulatory, and post-market decision support. Both the U.S. FDA and EMA have issued detailed guidance endorsing structured, provenance-rich RWE for complementing clinical trials. [14]

Embedded Rigor Through Observational Protocol Design

Each Circle Dataset begins with an OP that defines:

- **Anatomical region, pathology/diagnosis, treatment protocol, and relevant medical codes,** all aligned with clinical relevance.
- **Longitudinal, well-correlated outcomes data-capture schedule** (e.g., baseline, 3 m, 6 m, 12 m follow-ups).
- **Standardized instruments** for outcomes and diagnostics.
- **Required data elements** mapped to common data models (CDMs).
- **Quality-assurance and completeness thresholds**.

This turns real-world data (RWD) into *pre-specified observational evidence* rather than ad-hoc chart extractions. The methodological logic mirrors frameworks such as the **FDA's Real-World Evidence Program Guidance** [15] and the **ISPOR–ISPE Good Practices for RWE.** [16]

Longitudinal Structure And Statistical Power

Unlike static EHR snapshots, Circle Datasets follow patients over defined time intervals, supporting **within-subject comparisons**, **survival analyses**, and **causal inference modeling**. Longitudinal follow-up allows:

- **Trajectory modeling** of disease progression.
- **Assessment of treatment durability**.
- **Identification of delayed adverse outcomes**.
- **Improved generalizability** across diverse practice settings.

This design is consistent with recommendations from **PCORI** and **OHDSI** emphasizing longitudinal RWD for comparative effectiveness and predictive modeling. [17]

Validation And Quality Assurance Pipeline

Scientific integrity also depends on verifiable data quality. Each Circle Dataset is being designed to undergo multi-layer validation:

- **Automated syntactic validation** – real-time logic checks at point of entry, ensuring conformance to OP schema.
- **Cross-field and temporal consistency** – detection of implausible values or inconsistent time stamps.
- **Clinician review loop** – required physician sign-off for each patient record, adding human verification.
- **Inter-site audits** – periodic cross-comparison to identify outlier capture patterns.
- **External quality metrics** – completeness, plausibility, and timeliness scored per Kahn et al., 2016 Data Quality Metrics for EHR Data.

Each validation event can be recorded in the dataset's provenance log, forming an **immutable audit trail**—a critical element of both scientific defensibility and regulatory readiness.

Provenance As A Scientific Instrument

Provenance is not merely administrative—it is analytical. Being able to trace every observation to its origin improves the interpretability of model performance and enables transparent peer review.

The **W3C PROV model** and its adaptation for healthcare data management [18] underpins the Circles metadata layer, ensuring that transformations, derivations, and linkages are machine-readable. Such "explainable lineage" supports emerging *model-card* and *data-card* documentation practices [19] -- a key expectation in NIST's **AI Risk Management Framework 1.0**. [20] By embedding these standards, Circle Datasets allow independent reviewers to understand not only *what* the data shows but *how* it was generated.

IRB, MEC, And Ethical Oversight

Any Observational Protocol contributing to a Circle Dataset can be reviewed and approved by an **Institutional Review Board (IRB)** or **Medical Ethics Committee (MEC)** at participating institutions. RegenMed assists with that process.

Consent frameworks are available to meet or exceed **HIPAA**, **GDPR**, and **Common Rule** requirements for secondary data use. [21] Ethical oversight possibilities extend to:

- Secure patient de-identification via the **Safe Harbor** or **Expert Determination** methods.
- Controlled re-identification under documented governance for longitudinal matching.
- Periodic ethics re-review as protocols expand to new sites or indications.

This helps ensure that Circle Datasets remain compliant as living, longitudinal entities rather than one-time data extractions.

Network Effects And Continuous Learning

Because each Circle Dataset is linked to an active clinical protocol, data capture, feedback, and publication form a **closed learning loop**. Participating clinicians receive benchmark dashboards comparing their patient outcomes to aggregate, anonymized network results—encouraging data completeness and accuracy.

This creates a *self-reinforcing ecosystem*: Better data → more credible analyses → more clinician participation → broader coverage → richer models. Such feedback loops are essential for sustaining data quality and represent the "learning health system" ideal described by the **National Academy of Medicine.** [22]

Publication Pipeline And Reproducibility

RegenMed's Circles Network supports peer-reviewed as well as less formal methods of publication of dataset analyses. Each such publication can be based on:

- a persistent dataset DOI,
- version and protocol identifiers, and
- a summary of validation metrics.

This transparent linkage allows external investigators to reproduce findings and regulators to reference the same dataset version. The model parallels reproducibility initiatives such as **ClinicalTrials.gov Results Reporting** and **Vivli Global Data Sharing Platform.** [23]

Summary

By combining observational-protocol design, longitudinal structure, continuous validation, and transparent provenance, Circle Datasets achieve a level of **scientific integrity** typically associated with formal clinical trials -- but at real-world scale and cadence. They operationalize the principles of *trustworthy AI* and *regulatory-grade RWE* as defined by leading authorities, creating a reproducible foundation for model development, evaluation, and approval.

APPENDIX A — DATA MODEL AND FHIR SCHEMA ALIGNMENT

Overview

Interoperability is the foundation of trustworthy real-world evidence (RWE). Without consistent schema and terminological mapping, even validated datasets cannot be safely merged, benchmarked, or reused. The **RegenMed Circles Platform** therefore is being built to adhere to an explicit data-model architecture aligned with:

- **HL7 FHIR R4** resources for structural interoperability,
- **Standard clinical terminologies** (ICD-10-CM, CPT, LOINC, SNOMED CT, RxNorm), and
- **Common Data Models (CDMs)** used by FDA Sentinel, OHDSI OMOP, and PCORnet.

This alignment allows Circle Datasets to operate simultaneously as *clinically meaningful records*, *computational resources*, and *regulatory-grade evidence sources.*

Conceptual Data Architecture

Each Circle Dataset will reflect a **modular schema** derived from FHIR R4. Data elements will be organized under five principal resource groups:

Resource Group	Core FHIR Resources	Typical Content	Alignment Notes
Patient Identity & Demographics	Patient, RelatedPerson, Organization	Age, sex, ethnicity, site of care	De-identified UID mapping; FHIR tags mark anonymization status.
Clinical Encounters & Observations	Encounter, Observation, Condition	Diagnoses, vitals, PROMs, labs	LOINC and SNOMED codes; temporal indexing for longitudinal follow-up.

Resource Group	Core FHIR Resources	Typical Content	Alignment Notes
Procedures & Interventions	Procedure, Device, MedicationRequest	Surgical procedures, implants, drug orders	CPT / RxNorm mapping; causal linkage to encounters via basedOn.
Outcomes & Follow-up	Observation, QuestionnaireResponse, Condition	PROMIS scores, adverse events, recovery metrics	Observational Protocol (OP) defines temporal schedule.
Administrative & Provenance	Provenance, AuditEvent, ResearchStudy, DocumentReference	Capture metadata, consent, audit trail records	W3C PROV-DM compatible metadata structure.

Each record will carry a **universal Circle Identifier (CID)** — a UUID linked to site, protocol version, and timestamp — ensuring global uniqueness and audit traceability.

Mapping To Common Terminologies

Circle Datasets will use harmonized vocabulary mappings that maintain fidelity across data-sharing contexts:

Domain	Primary Terminology	Typical Use	Cross-walk Reference
Diagnoses	**ICD-10-CM, SNOMED CT**	Disease classification and phenotyping	WHO ICD-10 2023; SNOMED International 2024
Procedures	**CPT (AMA), SNOMED CT**	Surgical and outpatient interventions	AMA CPT Code Set 2024

Domain	Primary Terminology	Typical Use	Cross-walk Reference
Laboratory & Biometrics	**LOINC**	Quantitative lab and vital measurements	Regenstrief LOINC v2.77 (2024)
Medications	**RxNorm**, **ATC**	Drug identity and therapeutic class	NIH RxNorm API 2024
Outcomes & PROMs	**PROMIS**, **WHODAS 2.0**, custom LOINC extensions	Standardized patient-reported outcomes	PROMIS HealthMeasures 2023
Units & Measures	**UCUM**	Standard measurement units	UCUM Specification 2023

Terminology alignment will be maintained through a **reference mapping service** built on FHIR ConceptMap resources. Each mapping version will be recorded in the provenance ledger, supporting historical reproducibility and regulatory audit.

Longitudinal and Relational Structure

The Circle schema is explicitly **longitudinal**, linking repeated observations, encounters, and interventions to the same patient episode over time. Key design features are:

- **Episode linkage:** Each dataset defines an EpisodeOfCare resource serving as a temporal anchor for all related events.
- **Causal relationships:** FHIR basedOn, partOf, and derivedFrom references establish directional relationships between diagnostic, therapeutic, and outcome events.
- **Version control:** Each data element contains a meta.versionId and meta.lastUpdated field, ensuring temporal traceability for every modification.
- **Temporal integrity:** ISO 8601 timestamps with timezone normalization permit cross-site synchronization within ± 1 second.

This relational structure will support survival analysis, target-trial emulation, and causal inference techniques recommended for regulatory-grade RWE (Hernán & Robins, AJE 2016; Hubbard et al., *NEJM* 2024).

Data Quality Metrics And Validation

Quality assurance will follow the **Kahn et al. Data Quality Framework** (*J Am Med Inform Assoc* 2016) and the **OHDSI Data Quality Dashboard (DQD)** methodology. Four core dimensions will be tracked:

- **Completeness** – Proportion of mandatory fields populated.
- **Conformance** – Compliance with FHIR profile and terminology standards.
- **Plausibility** – Logical consistency across time and clinical context.
- **Timeliness** – Latency between real-world event and data capture.

Each site will be able to generate a monthly data-quality report automatically logged to its federated provenance ledger. Anomalies trigger automated alerts and re-validation routines.

Provenance And Metadata Alignment

Provenance metadata will be captured in compliance with **W3C PROV-DM** and FHIR Provenance resource specifications.

- Each event records who, what, when, and where of data creation and modification.
- Cross-references link each FHIR resource to its OP identifier and version number.
- Audit trails use AuditEvent resources with cryptographic hashes for immutability.
- Metadata objects are replicated across federated nodes for redundant preservation.

This design achieves *end-to-end traceability* required under FDA 21 CFR Part 11 and GDPR Article 30 record-keeping standards.

Cross-Model Compatibility

Circle Datasets can be transformed to major research CDMs via automated ETL pipelines:

Target CDM	Conversion Mechanism	Validation Status
OMOP (OHDSI)	FHIR-to-OMOP ETL using FHIR2OMOP adapter (v1.4)	Verified against OHDSI DQD 2024 benchmarks

Target CDM	Conversion Mechanism	Validation Status
Sentinel CDM (FDA)	Direct mapping via FHIR Observation and Encounter resources	Pilot validation complete at partner site
PCORnet CDM	FHIR ResearchStudy ↔ PCORnet Study cross-walk	Under review for multi-site deployment

Bidirectional conversion will preserve semantic equivalence and allow Circle Datasets to integrate into existing federal and academic RWE ecosystems without schema redesign (OHDSI 2023 CDM v6.1).

Interoperability Testing and Certification

The Circles Platform will undergo annual conformance testing to FHIR Implementation Guides (IGs) including:

- **US Core IG (v7.0)** — ONC certified for FHIR R4 server conformance (ONC 2024 Certification Program).
- **International Patient Summary (IPS IG v1.2)** — for cross-border interoperability.
- **SMART on FHIR and FHIR Bulk Data Access (v2.0)** — for secure API-based data export.

Certification testing will follow the IHE Connectathon model for interoperability verification.

Summary

The Circle data model is built to translate clinical reality into machine-interpretable structure without losing context or traceability. By combining FHIR R4 resources, standard terminologies, and federated provenance metadata, RegenMed is also creating a schema that is:

- Clinically rich and regulatorily sound.
- Immediately interoperable across institutions and CDMs.
- Future-proof for continuous AI and RWE integration.

This alignment ensures that every Circle Dataset is not just a collection of observations but a fully auditable, interoperable scientific asset—capable of supporting trustworthy medical AI at scale.

ENDNOTES

[1] https://academic.oup.com/intqhc/article/34/4/mzac080/6749973

[2] Ibid.

[3] https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-024-01414-1

[4] Ibid.

[5] https://www.axios.com/2019/05/30/medical-ai-data-problems

[6] https://www.scirp.org/pdf/ojbm_2024010514595810.pdf

[7] https://www.sciencedirect.com/org/science/article/pii/S1438887124002607

[8] https://arxiv.org/pdf/2501.05617

[9] https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-024-01414-1

[10] U.S. Food and Drug Administration (FDA). Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021.

[11] National Institute of Standards and Technology (NIST). AI Risk Management Framework 1.0. 2023.

[12] European Medicines Agency (EMA). Guideline on Registry-Based Studies and Real-World Evidence. 2024.

[13] OHDSI Collaborative / FDA Sentinel Program. Data Quality Framework for RWD/RWE Generation. 2022.

[14] https://www.fda.gov/media/120060/download; EMA RWE Guideline 2024.

[15] https://www.fda.gov/media/171667/download

[16] ISPOR Task Force Report 2023

[17] https://www.pcori.org/research-related-projects/about-our-research/research-methodology/pcori-methodology-standards. https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:comparative_effectiveness_research_opportunities_using_the_ohdsi_network_v6.0.pdf.

[18] W3C PROV-DM Specification 2013)

[19] Mitchell et al., 2021, arXiv:1810.03993

[20] NIST 2023

[21] HHS 45 CFR 46

[22] NAM Learning Health System Report 2022

[23] Vivli.org