RegenMed

WHITE PAPER

THE VERACITY CRISIS

Why "Big Data" Broke Healthcare
And How "Smart Data" Will Fix It

December 2025

TABLE OF CONTENTS

EXECUTIVE SUMMARY: THE "PSEUDO-PRODUCTIVITY" PARADOX

For twenty years, the global healthcare ecosystem has been seduced by a single, expensive promise: digitization would yield salvation. The theory was simple -- if we forced physicians to digitize patient records, aggregated those records into massive "data lakes," and applied machine learning, we would unlock a golden age of precision medicine, lower costs, and better outcomes.

The industry delivered the volume; we now generate zettabytes of health data annually. Yet, the promise remains spectacularly unfulfilled.

Despite spending over $4.5 trillion annually on healthcare in the United States -- nearly 20% of GDP -- life expectancy has stagnated, chronic disease management has plateaued, and the cost of developing a single new drug has risen to approximately $2.6 billion [1]. This phenomenon is known as the "Pseudo-Productivity Paradox": as healthcare became more digitized, it became less efficient, more expensive, and more detached from the patient [2].

The failure lies not in the *amount* of data, but in its *architecture.*

The current standard -- retrospective data mining -- relies on "data exhaust": administrative byproducts created for billing, not science. This paper argues that the era of "big data" (volume) must yield to the era of "ground truth" (veracity). It outlines the architectural failure of current "data swamps" and presents RegenMed's hypothesis-driven Circle Datasets as the necessary correction: a federated system where data is not "cleaned" after the fact, but born clinically relevant and verified.

THE ARCHITECTURE OF FAILURE: ANATOMY OF A "DATA SWAMP"

To understand why current Real-World Evidence (RWE) fails to cure disease or lower costs, we must examine the "Data Lake" architecture used by big data industry incumbents.

The "Schema-on-Read" Trap

Modern data lakes utilize a "Schema-on-Read" approach. They ingest massive volumes of raw, heterogeneous data -- PDF notes, JSON logs, CSV billing files -- in their native format, without predefined structure. The operating logic is: *"Store everything now, figure it out later."*

While this model is cheap for storage, it is catastrophic for usable retrieval. Without strict governance at the point of entry, these lakes degrade into "data swamps" -- chaotic repositories where data quality is unknown. As IBM developers noted in a seminal critique:

> *"A data swamp is a badly designed, inadequately documented . . . data lake. Even though the data exists, the swamp cannot retrieve it without contextual metadata . . .The data is like mystery meat."* [3]

The consequence is that researchers spend 80% of their time "cleaning" data and only 20% analyzing it. They are wading through a swamp, not querying a library.

The "Inference Gap"

The vast majority of data in these lakes is administrative (claims and billing). It lacks clinical intent or context. To answer scientific questions, vendors use proprietary "black box" algorithms to *infer* reality. A data lake might see a billing code for "knee pain" and a pharmacy claim for an opioid. The algorithm *guesses* the patient had surgery. However, the patient might have refused surgery, and the medication could be for a separate issue. The lake does not know.

This results in "probabilistic data". It is a statistical approximation of the truth, not the truth itself. As the National Institutes of Health (NIH) has warned:

> *"Results based on RWE may be overestimated . . . The results do not indicate the effect of the drug itself, but rather the actual practice of adding elements of patient compliance and the environment."* [4]

When FDA approvals or AI models are based on "guesses," patient safety is compromised.

THE PHYSICIAN AS "DATA SERF"

The current data economy is built on the backs of physicians, yet it is structurally designed to exclude them from the value they create.

The "Data Entry Clerk" Phenomenon

The "Productivity Paradox" hits physicians hardest. The digitization of healthcare has transformed highly trained healers into high-paid data entry clerks. Dr. Steven Stack, former AMA President, famously described the crisis:

> *"Imagine, in a world where a 2-year-old can operate an iPhone, you have graduate-educated physicians brought to their knees by electronic health records."* [5]

Studies indicate that following EHR implementation, physician productivity drops by roughly 30% and rarely fully recovers [6]. Doctors now spend "pajama time" -- hours of unpaid work at night -- completing charts to satisfy billing requirements or medical society registries, not clinical needs.

The Incentive Disconnect

Physicians see no revenue from the data they create. Aggregators monetize the data, selling it back to pharma and other licensees. This destroys the incentive for accuracy. If a doctor knows the data is just for billing, they enter the code that ensures payment, not necessarily the code that captures the subtle clinical reality.

In contrast, Circle Datasets allocate 85% of net license revenues to the real-world data contributing physician. The physician becomes a partner in supporting research and advancing healthcare, not a serf of the system.

THE ECONOMIC COST: THE $2.6 BILLION DRUG PROBLEM

If "Big Data" were working, drug development costs should be falling. Instead, they are skyrocketing.

The R&D Efficiency Collapse

The pharmaceutical industry is facing an existential crisis of efficiency:

- **The Cost:** Developing a new prescription drug now costs approximately $2.6 billion, a figure that includes the cost of the thousands of failures that precede every success.

- **The Failure Rate:** The success rate for drugs entering Phase 1 clinical trials has plummeted to just 6.7% in 2024, down from 10% a decade ago.

- **The Cause:** Relying on "probabilistic data" leads to false signals. Pharma companies invest billions chasing targets that looked promising in “data swamp” analytics, but fail when tested in the rigorous reality of a clinical trial.

The "Synthetic Control" Myth

The industry hopes to use synthetic control arms (using historical data instead of a placebo group) to save money. However, using “swamp data” for this is dangerous. If the historical data is inferred rather than observed, the comparison is invalid.

- **The Reality:** You cannot build a control arm from billing codes. You need ground truth -- verifiable, longitudinal outcomes -- to safely replace a placebo group.

THE AI HAZARD: HALLUCINATION AS A SERVICE

The rapid adoption of generative AI and large language models (LLMs) in healthcare has exposed the dangers of "messy" data like never before. AI is a magnifying glass; if you point it at a swamp, you get a clearer view of the mud.

"Garbage In, Fiction Out"

AI Hallucinations occur when a model *"perceives patterns or objects that are nonexistent"* [7] or invents facts to fill gaps in its training data. An AI trained on "probabilistic data" learns the *guesses* of the algorithms that cleaned the data. It does not learn medicine; it learns the biases of billing codes. As Vimal Kansal notes:

> *"If their training data is missing correct information . . . or contains errors or biases, the model has no choice but to generate an answer from whatever related patterns it did learn — which could be wrong."* [8]

In healthcare, a "wrong guess" is a malpractice suit or a patient injury.

The Need for Grounding

To be safe, medical AI must be grounded in verifiable facts. It requires data with a chain of custody. Current data lakes strip away provenance. The AI doesn't know *who* said the patient had cancer, only that the code exists.

In Circle Datasets, every datapoint is linked to a specific physician, a specific time, and a specific clinical hypothesis. This provides the ground truth necessary for Retrieval Augmented Generation (RAG), the only safe architecture for medical AI.

THE SOLUTION: HYPOTHESIS-DRIVEN ARCHITECTURE

RegenMed is not building a better "data swamp." It is building a “Mint for Evidence”. Its patented architecture represents a fundamental shift from retrospective inference to prospective evidence generation.

Schema-on-Capture: The End of Inference

RegenMed replaces the "Schema-on-Read" model with "Schema-on-Capture." Data is only accepted into the Circle Datasets lake if it conforms to a rigorous pre-structured Observational Protocol at the moment of entry at the point of care [4].

Circle Datasets do not just record *what* happened; they record *why*. Every datapoint is an answer to a specific clinical question. This creates longitudinal deterministic (rather than probabilistic) evidence. There is no algorithm guessing what happened; the physician and patient explicitly stated it.

Federated Sovereignty

Instead of the "vacuum cleaner" model (sucking data out of hospitals into a central cloud), RegenMed’s Circles platform enables federated data architecture. Patient data can remain resident at a local data node. Researchers, payers, AI models and others in the healthcare ecosystem send questions *to* the data, and only the *answer* comes back.

This "Zero-Copy" approach solves the massive security crisis of centralized breaches. In 2024 alone, over 725 large healthcare breaches exposed 190 million records [9]. By enabling the local maintenance of data, Circles can remove the "honey pot” target for hackers.

CONCLUSION: THE VALUE SHIFT

The market is shifting. Pharma, payers, providers, researchers, AI developers, regulators, and patients are realizing that the "big data" experiment has failed. The promise of volume has been broken by the reality of noise.

- **Yesterday:** Value was defined by volume ("We have 100 million records").
- **Tomorrow:** Value is defined by veracity ("We have 10,000 records that are demonstrably true").

RegenMed stands alone as the provider of this veracity. By respecting the physician, protecting the patient, and validating the clinical/scientific hypothesis, it is moving the industry out of the swamp and onto solid ground.

REFERENCES

[1] PatentPC. "The Cost of Drug Development: How Much Does It Take to Bring a Drug to Market?" (2025).

[2] National Institutes of Health (PMC). "Measuring What Matters: The Pseudo-Productivity Paradox in Health Care." (2024).

[3] IBM Developer. "Data lakes and data swamps." (2018).

[4] NIH PMC. "Proceed with Caution When Using Real World Data and Real World Evidence." (2019).

[5] Becker's Hospital Review. "25 quotes that show just how fed up physicians are with EHRs." (2015).

[6] Health Catalyst. "Physician Burnout and the EHR: Addressing Five Common Burdens."

[7] IBM. "What Are AI Hallucinations? Risks & Mitigation."

[8] Kansal, Vimal. "Understanding and Mitigating AI Hallucinations." Medium (2025).

[9] Sprinto. "Healthcare Data Breach Statistics: HIPAA Violation Cases and Preventive Measures in 2025." (2025).