ZOOMRX

Synthetic Data Is Not Enough

A position paper from ZoomRx

RELEASE DATE

May 2026

READING TIME

10-15 minutes

AUTHOR

Ty Harkness

Head, Sagan Agents



ABOUT THE AUTHOR



Ty Harkness, Ph.D.

HEAD OF SAGAN AGENTS

Ty Harkness, Ph.D. is the Head of Sagan Agents at ZoomRx, where he leads the company's AI platform focused on transforming life sciences insights and analytics. He previously served as Chief of Staff to ZoomRx's Product Development organization, helping shape the company's innovation strategy and platform roadmap.

Ty brings experience across the pharmaceutical value chain, with prior roles at Pfizer, Mayo Clinic, and Charles River Associates. Today, he partners directly with pharma and biotech leaders to reimagine how commercial functions operate in an increasingly AI-driven environment.

Synthetic data has become a catch-all term representing a necessary transformation in how pharmaceutical insights functions operate: AI augmenting or replacing traditional workflows, resulting in dramatically faster insights and dramatically lower costs. In reality, generating synthetic data is only one part of the transformation that needs to take place, and a smaller part than many may suggest. While synthetic methodologies can and should be a valuable tool for insights teams, organizations hoping to drive AI transformation by simply replacing human respondents with synthetic personas are likely to be disappointed in the results.

This paper is ZoomRx's view on what this broader transformation looks like and where synthetic data fits within it, based on fifteen years of deploying both primary and synthetic methods against our own data corpus in the service of our biopharma clients. It is not a survey of the field, but rather an opinionated stake in the ground regarding the current discourse around synthetic data.

SECTION 1

Everyone is asking the wrong question about synthetic data

Every pharma company is evaluating synthetic data right now. Most are asking: how do we use synthetic data to make our research cheaper and faster?

We think that is the wrong question.

The cost savings from replacing human respondents with synthetic ones are real but modest. In a typical study, sample costs are 15–30% of the total project budget. The rest is professional services: design, analysis, synthesis, deliverables. Replace the respondents and you save the sample portion. Everything else is unchanged, meaning that leveraging synthetic data alone is not a dramatic cost play.

The speed case is stronger, but often overstated. A traditional MR cycle takes six to twelve weeks, while synthetic respondents can answer questions immediately. But synthetic data is not the only path to speed. AI moderation can now conduct deep qualitative interviews with human respondents at the scale and pace of traditional quantitative research. Combined with AI-driven analysis and reporting, research

cycles can be compressed from weeks to days without opting for synthetic respondents. Speed is very valuable, but synthetic responses are not required to dramatically compress research cycles.

If cost and speed savings using synthetic data are marginal rather than transformational, then the right question is not "can we use synthetic data to transform our insights function?" Rather, the right question is: "what kind of intelligence system does an insights function need to make better decisions faster, and where do synthetic methodologies fit within that system?"

The rest of this paper answers these questions. Section 2 provides a framework for when synthetic methods belong and when they do not. Section 3 establishes what makes synthetic outputs credible. Section 4 argues that the most valuable application of synthetic is not faster surveys but projected business outcomes. And Section 5 describes the closed-loop intelligence system that ties it all together, with synthetic methodologies as one component among several.

SECTION 2

Wet lab and dry lab: the right model for when to use synthetic

Biomedical research and development (R&D) has a mature operating model for this kind of question. Wet lab is physical experimentation on real biological systems, in a laboratory setting or in clinical trials. Dry lab is computational modeling of those same systems. The rule is simple: use dry lab by default, and do wet lab only when it is necessary. Neither replaces the other, but competent organizations know when to use which.

The same framing applies cleanly to MR:

- Wet lab = primary research with real respondents. Irreplaceable, but slower and more expensive.
- Dry lab = a portfolio of methods: data archive query over prior research, synthetic response generation, and market simulation. Fast, cheap, and where most questions can actually be answered.

The operational question is not “should we use synthetic?” It is “where does this specific question require wet lab?” Two frontiers answer that:

THE KNOWLEDGE FRONTIER

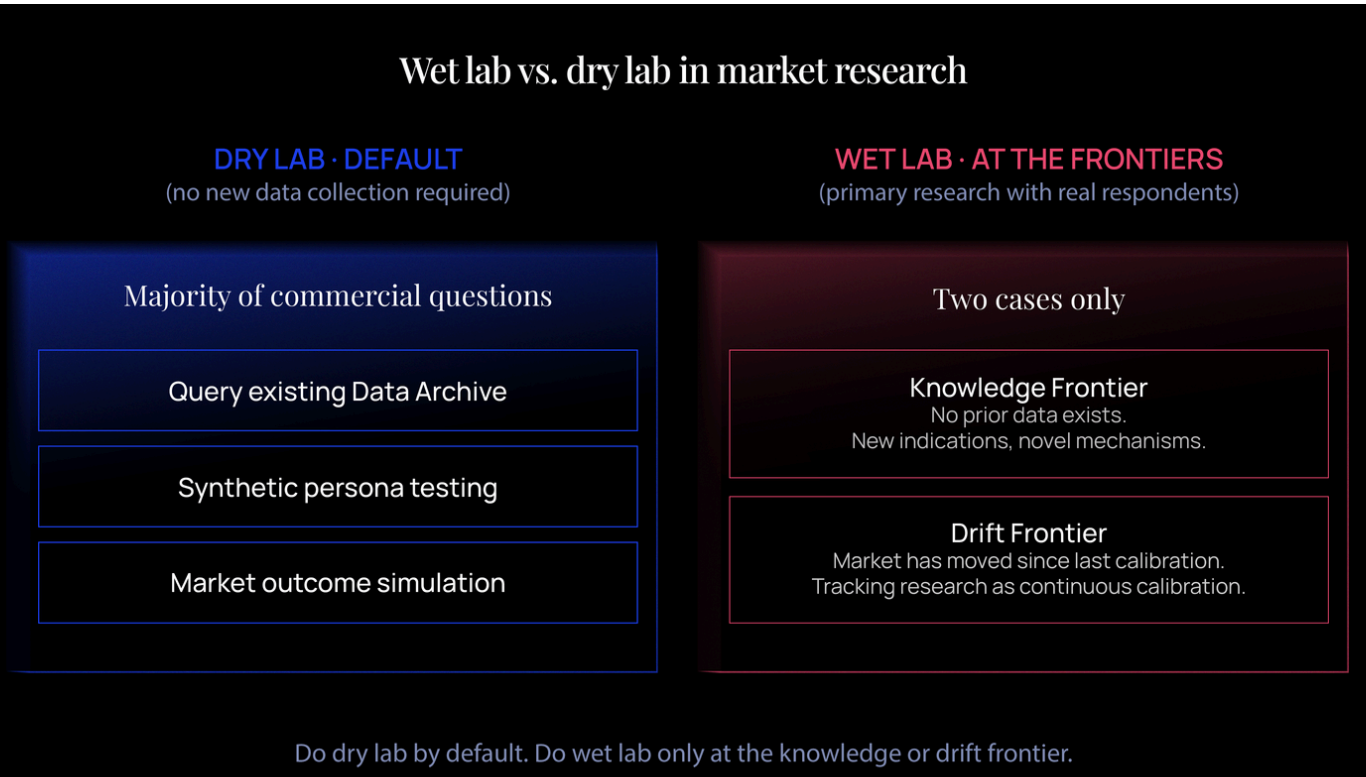
Genuine gaps where no prior data exists. A new indication, a novel competitive mechanism, a question that has never been asked. Wet lab is the only way to generate new signal.

THE DRIFT FRONTIER

The market has moved since the last calibration. A label update, a guideline change, a competitive entry, a payer shift. Tracking research earns its cost here, reframed as continuous calibration rather than quarterly reporting.

Everything between the two frontiers is answerable in the dry lab — dramatically faster than wet lab, and, when the archive already holds the answer, at near-zero marginal cost. That is the majority of commercial questions a brand team will ask in a given year.

Synthetic data specifically is one tool in the dry-lab toolkit — the one that generates new responses from a calibrated model when the archive does not have the answer. Data archive query (recovering what is already known) and simulation (modeling how the market plays out under new conditions) are equally important and often more useful. “Synthetic data” is not the whole dry lab; it is one method within it.



SECTION 3

The data foundation, not the algorithm, is where value lies

Any team with an LLM API key can generate confident-sounding synthetic responses. What they cannot do is ground those responses in the actual behavior of actual physicians, patients, and payers. The quality of dry-lab output is bounded by the primary data underneath it. Foundation is a long accumulation game, not a technology choice.

ZoomRx's foundation is fifteen years of primary MR: ATUs, tracking studies, message tests, HCP-patient conversations, digital behavior, and a 60,000+ HCP panel. That data set was built one study at a time, exclusively within life sciences, and represents a depth of behavioral ground truth that no synthetic-only vendor can assemble retroactively.

More importantly, it is structured into a continuously queryable data archive. That structural move, turning fifteen years of research from filed artifacts into a live intelligence layer, is what makes calibration and retrieval possible at all. Without it, you have a corpus. With it, you have a foundation.

Foundation is necessary but not sufficient. Markets drift, and a model calibrated on last year's data describes a world that no longer exists. Continuous calibration requires a live primary research capability: the same panel, the same fielding infrastructure, running in real time to refresh the model whenever the drift frontier is triggered.

This creates a structural asymmetry in the category. A vendor without a primary research arm can build a synthetic model, but cannot keep it honest. They can train on public data and hope. Only a small number of providers can do both: build the foundation and run the calibration loop. Doing so requires being the research firm that became an AI company, not the other way around.

Two questions separate real platforms from AI theater in every vendor evaluation:

> *What primary data was your model built on?*
>
> *How is it kept current?*

Vendors who cannot answer both with specificity are selling a decaying asset.

SECTION 4

Synthetic personas are a means, not an end

Market research has always been a proxy. The survey is a method; the point is the decision it informs: prescribing share, treatment persistence, launch uptake, switching behavior. Research has never been the end in itself.

Synthetic personas inherit that same proxy status. Yet the dominant synthetic persona product on the market today is: make synthetic personas take the same survey humans have always taken, just faster. That is a lack of imagination dressed up as innovation.

If you have built a calibrated model of a physician, patient, or payer — a model that captures their beliefs, attitudes, prescribing context, and decision logic — the least interesting thing you can do with it is hand it an ATU questionnaire. You already know what that exercise returns. The industry has been running it for decades.

The interesting move is to skip the survey and project the outcome directly. Ask the model what you actually wanted to know: How does this launch play out over eight quarters? How does share shift if a competitor enters in Q3? Which message moves prescribing behavior in which segment? Those are business questions. A credible model can answer them, not by generating synthetic survey responses to be analyzed, but by simulating the behaviors themselves.

The mechanism is a population of agents (HCPs, patients, reps, payers, etc.), each with distinct identities, beliefs, and behaviors, interacting with each other rather than responding in isolation to a survey instrument. Each agent's behavior emerges from its attitudes and context. Aggregated across the population, those behaviors produce outcomes at the market level: prescribing share, treatment persistence, launch trajectory. This collapses what has traditionally been two separate disciplines, market research and forecasting, into one simulation. The research becomes the forecast.

This is a category shift, not an incremental improvement. From "synthetic survey respondents" to "simulated market outcomes." Different product, different evaluation criteria, different validation standard.

Survey-response simulation is validated against respondent concordance. Market-outcome simulation is validated against historical trajectories: did the model predict the launch curves that actually played out?

Synthetic personas retain narrow tactical value: message screening, variant pre-testing, scenario interrogation before committing to primary research. These are legitimate uses and we deploy them. But the strategic unlock, the one that reshapes how commercial decisions are made, is at the market-outcome level. That is the part of this category we are most serious about.

SECTION 5

The endpoint is a closed-loop intelligence system

Everything in this paper compounds. A capable, agentic Data Archive is a critical prerequisite to assessing existing knowledge and compounding it over time. The wet-lab / dry-lab framework decides when to use synthetic vs human respondents. Continuous calibration keeps the dry lab honest. Outcome projection skips the survey-as-output trap. But these are not independent capabilities. They form a loop:

1. The Data Archive surfaces knowledge gaps, exposing the distance between what has been measured and what a current decision requires.
2. Gaps trigger knowledge generation, either through primary research with real respondents (wet lab) or through simulation and synthetic methods (dry lab). The framework from Section 2 decides which.
3. New knowledge flows back into the Data Archive, expanding the foundation.
4. Calibrate, reassess, cycle forward. Drift-frontier signals are detected as the archive accumulates; models are refreshed; and the loop begins again.

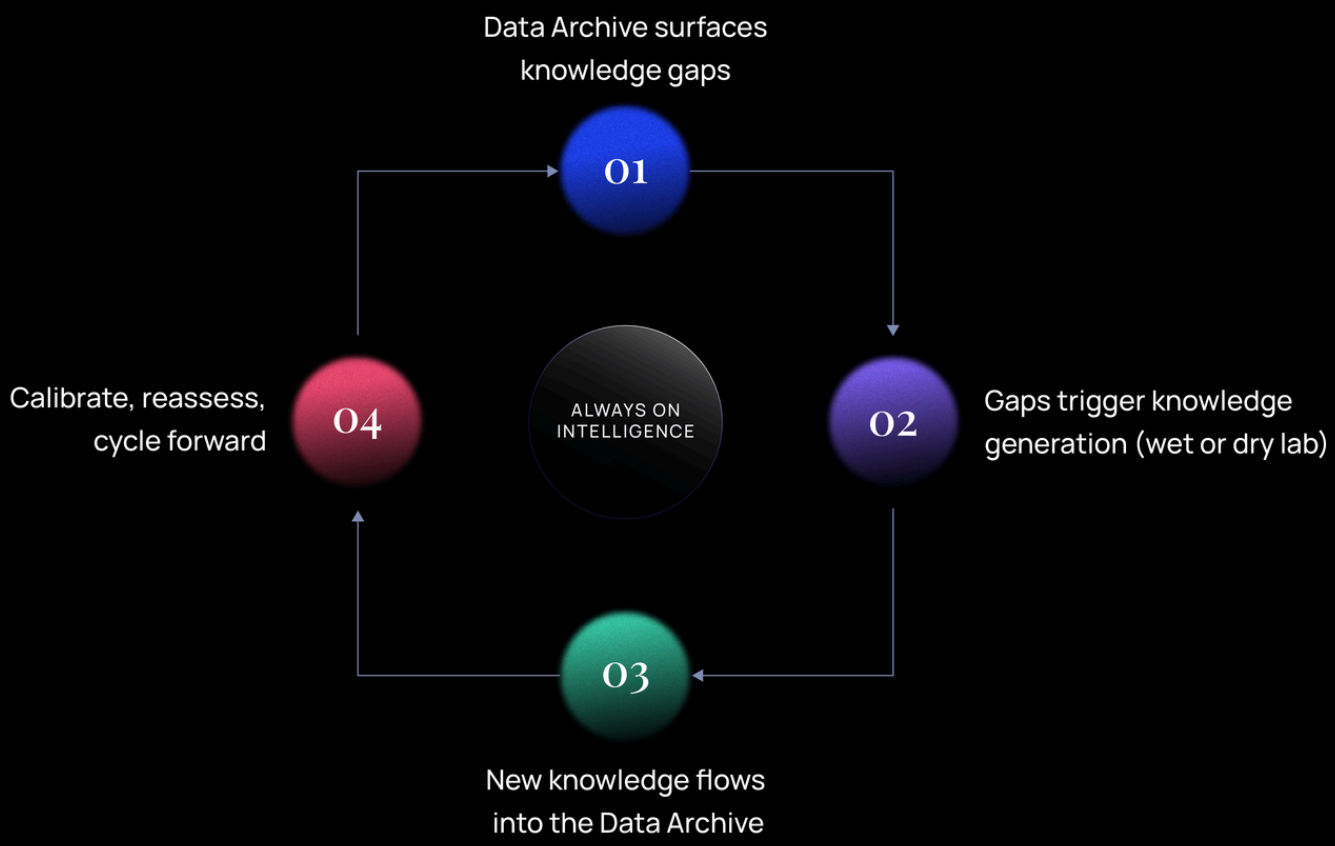
This is the operational shift the category enables. Traditional MR is project-based: commission, deliver, file, forget. The closed-loop intelligence system is cumulative. Every study expands what the organization knows about its market, and the system actively identifies what to go learn next. Knowledge compounds instead of expiring.

For a commercial team, the implication is direct. The gap between what the organization has paid to learn and what it can actually use on its next decision collapses toward zero. Studies completed for one brand become context for every future question across the portfolio. Research stops being a cost center and starts being an asset that appreciates.

Synthetic methods live inside this loop as one tool among several: the method for extrapolating within the archive's reach and for projecting outcomes at the market level. But synthetic methods are not the system. The system is the loop. And the loop is what separates a new generation of continuous commercial intelligence from a faster version of the same old research cycle.

The system we have been building to operationalize this loop is called Sagan Agents. We welcome disagreement, pressure-testing, and pointed questions.

The closed-loop intelligence system



Sagan Agents
BY ZOOMRX

Sagan Agents is ZoomRx's AI-driven platform built to solve a core tension facing pharma insights teams: growing demand for intelligence against flat budgets and research cycles that take 6–12 weeks per study. Sagan Agents operates across two integrated layers - AI-Enhanced Research Streams, which embed AI capabilities across every stage of the research lifecycle, and a Continuous Intelligence Platform, which unifies all studies, data sources, and insights into a single always-on system.

The platform's intelligence layer is powered by three modules that amplify each other:



EXISTING LIBRARY + CUSTOM BUILT

Strategic Insights Agents

Purpose-built agents to accelerate and amplify any customer insights workstream. Segmentation, message testing, patient journey, forecasting, brand health assessment, competitor tracking, synthetic personas, and more.

Built on the Sagan Agents composable skills framework. You describe the workflow, we create the agent.



BASE PLATFORM

Data Archive Intelligence

Your entire research history - quant datasets, qualitative reports, claims, CRM - searchable and synthesizable by AI agents.



BASE PLATFORM

Agentic Market Research

End-to-end AI-powered survey design, fielding, analysis, and reporting. Research question to strategy deck in days.

GET IN TOUCH

ZOOMRX

GLOBAL HEADQUARTERS

UNITED STATES
245 Main Street, FL 2, Cambridge, MA 02142, Boston

INDIA
Floor 2, Block B, Global Infocity Park, Perungudi, Chennai 600 119, Tamil Nadu