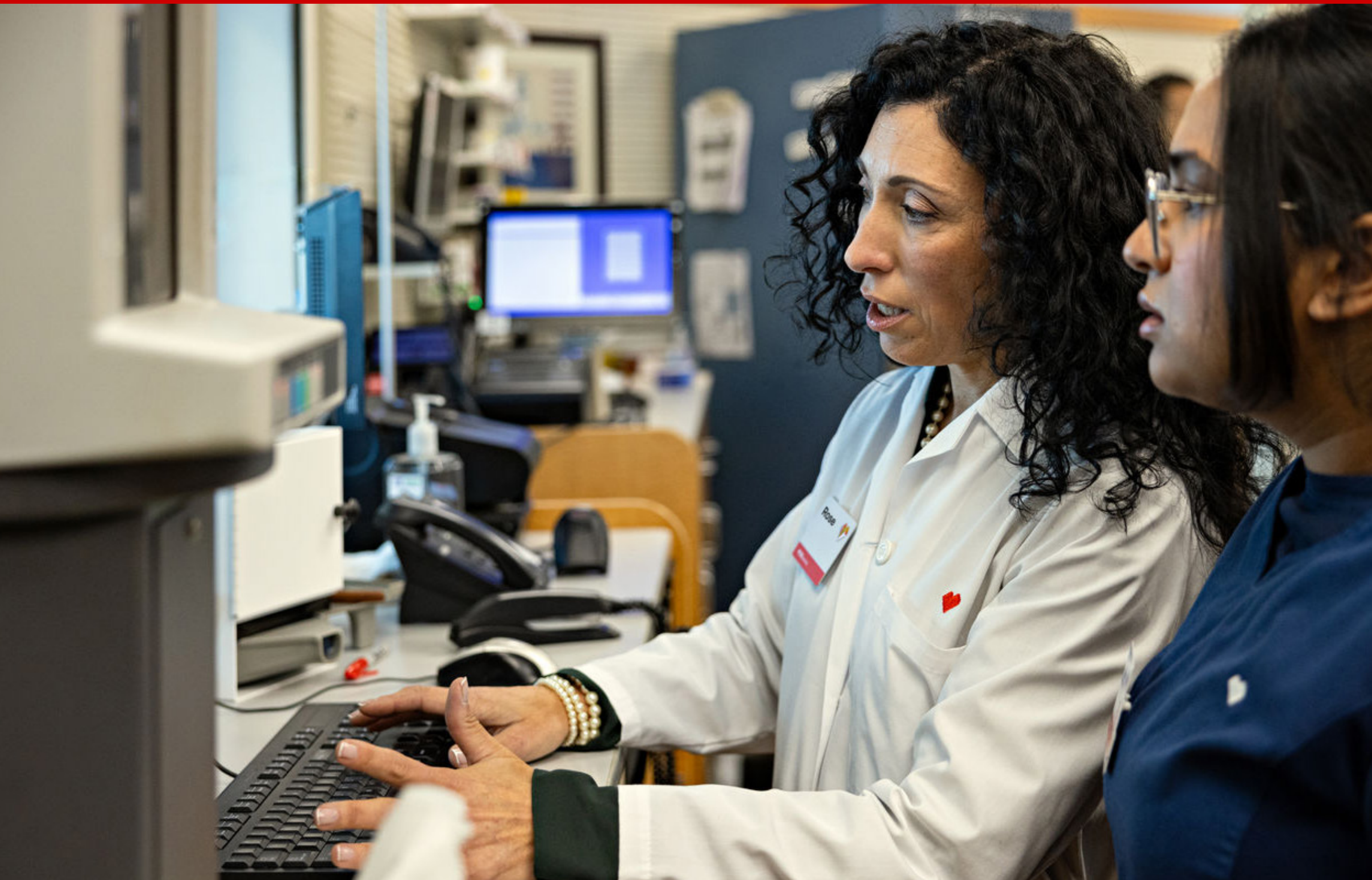
The Flight Simulator for Management:

Generative-Agent Simulations for Faster, Safer Decisions



Introduction

Markets don't run on autopilot; they're built on thousands of deliberate choices and underlying assumptions.

Consider a streamlined checkout flow that reduces a purchase to a single tap. It's built on a core assumption: that consumers value speed over comparison shopping. Or think of standardized credit card fee disclosure tables that assume customers take the time to read and interpret information simply because it's presented clearly. Recommendation systems across digital platforms similarly rest on the belief that past behavior reliably predicts future preferences. And auction-style marketplaces presume bidders will act rationally, even though last-second "sniping" is now part of the culture.

When these assumptions drift even slightly from reality, systems can behave in unexpected ways—risk is mispriced, the wrong users show up, or the right ones leave. Classic economic examples such as the "market for lemons" and the mechanics of bank runs demonstrate how small misreads of human behavior can escalate into system-wide issues.

The same dynamic plays out inside companies. A product team's onboarding flow, a retailer's promotional calendar, or a lender's approval criteria are all constructed around implicit beliefs about real customers. When those beliefs are even modestly off, flawless execution still fails to deliver the expected results.

Leaders know this, which is why inside every modern enterprise, teams in insights, product, strategy, and finance are tasked with turning uncertainty into decision-quality evidence. Much of this work relies on historic data, such as operational metrics, demographics, and purchase histories, stitched together by analysts to infer customer behavior and market dynamics.

When possible, teams complement these views with A/B tests or natural experiments to isolate behavioral effects. However, these methods carry persistent constraints: access to the right participants is slow and expensive; niche populations can be essentially unreachable; and many questions, especially those involving second-order effects or network dynamics, are simply not testable in the wild without unacceptable cost or risk. Experiments often take months, face operational or ethical limits, or require extrapolation from contexts that only imperfectly resemble the decision at hand. As a result, even the best operators are forced to make pivotal calls with partial views of their customers and markets.

A new capability is changing that calculus: simulation with generative agents.

To address these gaps, CVS Health has built high-fidelity synthetic populations, “agentic twins” that behave consistently with the real people they are modeled on using Simile’s generative-agent technology. Using these agentic twins, leaders can run “what-if” exercises, forecasting the behavior of customers, colleagues, analysts, and competitors. Given descriptors such as demographics, preferences, and prior behavior, these agents can participate in surveys, navigate product experiences, debate tradeoffs, and interact with one another over time. The result is not a crystal ball, but something closer to a flight simulator for management: a way to pressure-test designs, messaging, pricing, and policies against a library of plausible futures before those choices meet the market.

Over the past year, CVS Health has applied generative-agent simulations as a decision-support capability, grounded on 2.9 million consented responses from more than 400,000 participants across 200-plus behavioral scenarios.

The program is now used to: (1) uncover friction across end-to-end journeys–digital and physical; (2) access and query difficult-to-reach populations; (3) conduct digital product testing before rollout; and (4) run multi-market simulations to benchmark competitive perception. These simulations enable CVS Health to explore customer journeys, market responses, and operational trade-offs safely—expanding the range of questions that can be tested quickly and responsibly.

This article offers a pragmatic playbook for making simulation a strategic capability in 2026 and beyond, distilling what CVS Health has learned about building a durable program using Simile’s technology. This paper begins with a clear, nontechnical overview of the underlying science–why and how generative-agent simulations work–and summarizes the latest validation results to anchor expectations about accuracy, bias, and generalizability. The heart of the piece is an operating guide: how to start with individual-level studies (surveys and experiments), scale to multi-agent and ecosystem simulations (market and network effects), and integrate simulation with your existing research stack, instrumentation, and decision rites. The CVS Health program serves as a running example, highlighting the rollout steps, guardrails, and governance practices that keep the technology responsible and useful.

Simulation is becoming a first-class instrument of managerial judgment. For organizations that need to learn faster than the market changes, it shifts both the cost curve and the cadence of evidence. As an early adopter, CVS Health is already compounding an advantage: more hypothesis tested, more edge cases explored, fewer surprises post-launch. The question for leaders is no longer whether to use simulation, but where to start–and how to build a capability that lasts.

Simulating People with Generative Agents

There is a certain inevitability to simulation. Engineers already design and verify complex systems virtually; operators rehearse supply-chain and manufacturing decisions with agentic twins; risk teams price portfolios against thousands of weather and catastrophe scenarios. However, social simulation–despite its enormous economic value–has been the final frontier because human behavior is complex. People converse, reason, coordinate, and adapt; groups form norms; networks amplify shocks. For decades, scholars and practitioners relied on highly stylized agent-based models: a handful of fixed parameters standing in for cognition, motivation, and social context. These models yielded elegant theory and useful intuition, but their simplicity made generalization across markets difficult and made it easy to miss the contingency and diversity that characterize real human behavior.

That constraint has changed. Modern generative AI systems give us a new substrate for social simulation. Trained on a vast corpora of language and interaction, these models produce plausible, coherent behavior in natural language, the medium where much of business actually happens (shopping, support, negotiation, persuasion, collaboration). Used naively, though, they drift toward an "average internet persona," reproducing stereotypes and majority viewpoints. That is precisely where enterprise questions are most demanding: leaders rarely ask, "What would the average person do?" They ask, "How would *this* subpopulation behave in *this* context?"–for example, newly enrolled Medicare members evaluating a mail-order pharmacy workflow, or first-time small-business owners reacting to a price change in a SaaS product.

The answer is to model individuals, not averages. The simulation capability described relies on generative agents: AI agents seeded with consented, person-level data (e.g., interviews, past choices, and longitudinal signals) that serve as faithful proxies for real people. These agents are not the people themselves; they are privacy-preserving digital counterparts designed to help teams test hypotheses quickly and safely.

Individuals as the Quantum Unit

Whether you are fielding a survey, running a behavioral experiment, or exploring market ripple effects, individuals are the quantum unit from which all group and ecosystem dynamics emerge. Working at the individual level gives teams three practical advantages:

1. Targetability.

You can sample the exact population you care about (e.g., Spanish-speaking caregivers in Texas who have churned from a pharmacy loyalty program) rather than hoping a generic model “acts” like them.

2. Composability.

Individual agents can be arranged into panels, segments, and multi-agent markets, enabling both micro-level and system-level analysis from the same building blocks.

3. Auditability.

Because each agent is tied to a documented calibration process and data lineage, you can inspect what drives a result and update it as the world changes.

How Generative Agents are Created and Validated

The validation results outlined below are drawn from published Simile research and academic collaborations with Joon Sung Park, Co-Founder and CEO of Simile and are cited here to ground the approach scientifically, while CVS Health's contributions focus on enterprise application and deployment.

The technical architecture is straightforward in concept. Agents maintain structured memories, retrieve what is relevant, form intentions, and update beliefs as they encounter new information. Simile first demonstrated realistic, longitudinal behavior in 2023 by populating a simulated town ("Smallville") with such agents, who developed routines, shared information, and organized collective activities without being explicitly scripted. In 2024 Simile recruited more than 1,000 U.S. adults–sampled to be representative across age, gender, race, education, income, and state–and conducted two-hour, voice-to-voice semi-structured interviews administered by an AI interviewer. The interview guides, designed independently by sociologists at Stanford and Princeton as part of the American Voices project, elicited respondents' life histories, values, and experiences. Simile transcribed those interviews and used them to seed each participant's agent.

To test fidelity, Simile had the original participants complete a battery of instruments chosen independently of the interview script: modules from the General Social Survey, Big Five personality measures, behavioral-economics games, and a set of randomized controlled trials previously

published in the Proceedings of the National Academy of Sciences. Their corresponding agents completed the same battery.

On survey outcomes, the agents predicted their source individuals' responses at roughly 85% of the reliability with which people reproduce their own answers over time; across the randomized experiments, the effect-size patterns correlated above 0.9. In plain terms: the agents were accurate enough to serve as decision-support partners for many research and design tasks, particularly when speed, reach, longitudinal observation, or scale is required.

Since then, CVS Health has focused on decision support. In CVS Health enterprise deployments today, we further calibrate agents with first-party data–historical survey responses, CRM events, support interactions, or A/B test results–subject to consent and governance. In internal tests, these calibrated simulations have replicated CVS Health known findings with agreement rates up to 95% and anticipated the direction and relative magnitude of new results ahead of fieldwork. Live studies remain the arbiter for high-stakes choices, but simulation expands the frontier of what can be explored quickly and safely. The upshot is pragmatic: by grounding agents in real individuals and validating them against independent measures, social simulation becomes a credible complement to traditional research–one that helps teams ask better questions, prioritize what to test in the wild, and enter the market with fewer surprises.

The simulations described here are not

Strategic Playbook in Four Stages

abstract or purely synthetic. They are grounded representations of real customers, calibrated on observed behavior, history, and context. In the truest sense, they function as an amplification of customer voice, allowing organizations to consult customers even when they are not directly reachable.

CVS Health is already changing how large it learns from customers, colleagues, and partners. For CVS Health, the practical question has shifted from *whether* to experiment with this capability, to *how* to adopt it in a way that compounds advantage quarter after quarter.

The playbook below reflects what CVS Health has seen scale in enterprise settings. The core technology stays constant–model real individuals as calibrated generative agents–and sophistication grows by adding three dimensions: **time** (longitudinal experience), **interaction** (people influencing one another), and **environment** (markets and places).

Throughout this paper, CVS Health uses these simulations as a practical decision-support capability, illustrating how this progression translates into measurable business impact across our customer, patient, and enterprise experiences.

The throughline across all four stages is simple: bringing a more authentic voice of the customer into decision-making earlier, faster, and with greater coverage than traditional methods allow.

Stage 1.

Static, individual-level simulation: an "always-on" synthetic panel

The opening move is to create calibrated generative agents that represent real people and to query them at a point in time. In practice this looks familiar: surveys, concept tests, and message trials, or structured interviews–only faster, cheaper, and always on.

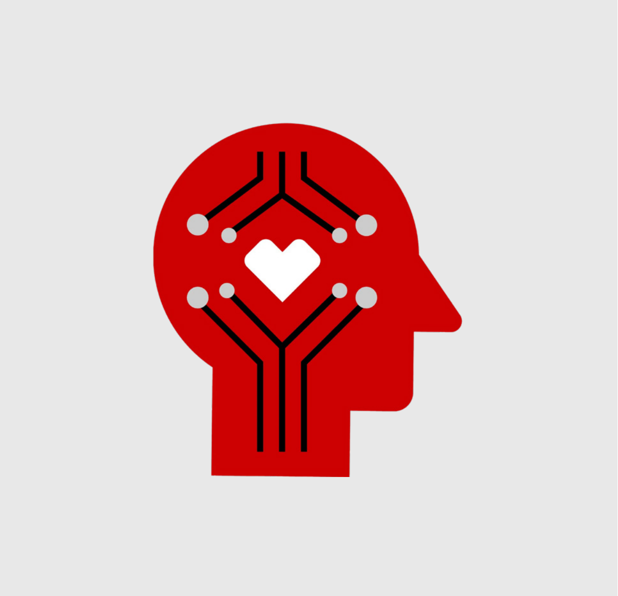
Validation is straightforward: back-test results against recent human data, report accuracy by segment, and use the simulator to prioritize hypotheses rather than to replace decisive field tests.

CVS Health Vignettes

At CVS Health, the "always-on" panel is anchored to priority customer segments across pharmacy, retail, and clinical services. Teams routinely re-create prior research to validate fidelity, then use the panel to triage messaging and offer concepts the same day.

For example, prior CVS Health qualitative research had established a clear link between patient experience and adherence. The challenge was not identifying *that* experience matters, but interpreting *why* – which experience factors matter most, for which patients, and under what conditions. Traditionally, answering those questions required weeks of stakeholder alignment, assembling new datasets, or commissioning additional qualitative studies, slowing leadership's ability to act on time-sensitive growth and health outcomes.

Using Simile's calibrated digital agents, CVS Health compressed this process into hours. Researchers replicated known adherence patterns, segmented maintenance-medication users, and ran a MaxDiff analysis to force trade-offs among competing experience drivers, surfacing clear priorities rather than just broad, undifferentiated importance ratings. The results showed clear separation: trust in the pharmacy and confidence that medications are handled correctly consistently ranked as the strongest drivers, followed by convenience-related factors such as refill ease and pickup experience.



Crucially, the value was not just ranking factors but pairing those quantitative priorities with interpretable “why” evidence. Follow-up questioning of non-adherent CVS patients in Simile’s agent population surfaced concrete barriers – confusion around instructions, refill timing anxiety, and service experiences – and specific interventions to help. Using Simile enabled CVS Health to expand the surface area of traditional research from knowing that experience matters to knowing which experience changes are most likely to change behavior, and why.

The same approach has been applied to uncover the “why” behind preferences and motivations of pet owners. During preparation for a national Pet Rx launch, researchers used the panel to uncover 4 core insights related to emotional drivers of pet ownership among pet parents, a population that is heterogeneous and difficult to reach consistently through traditional research. Those insights were translated into 6 concepts tests and immediately tested for resonance, shareability, and brand alignment. What would traditionally require multiple research waves over two to three months was compressed into a three-day cycle, yielding a short list of tested concepts with a higher likelihood of driving prescription fills.

The same approach has been applied to health-critical questions where direct experimentation is slow or sensitive. In many Rx contexts, primary research is constrained by real limitations: asking patients to relive adverse health experiences, probing non-adherence tied to shame or fear, or repeatedly testing speculative concepts that may never reach market can introduce ethical, operational, and bias risks. These constraints often result in fragmented learnings spread across dozens of historical studies, with no safe or efficient way to recombine them into a coherent view of unmet need.

To study Rx growth initiatives, CVS Health used Simile and calibrated agentic twins to synthesize existing research at scale, effectively conducting a meta-analysis of ~50 past studies grounded in real customer and patient voices. By aggregating prior qualitative and quantitative research into calibrated agents, CVS Health researchers could explore how emotional, experiential, and behavioral drivers interact across patient segments – surfacing where unmet needs consistently converged and where uncertainty remained highest. This allowed researchers to identify which questions were worth taking back to the field, and which concept spaces were most likely to warrant more validation.

Across these vignettes, the value is consistent: by surfacing early shifts in understanding, confidence, and intent – especially among populations that are hard to reach or intermittently engaged – teams can anticipate where experience changes are most likely to influence adherence and downstream health outcomes, rather than reacting after outcomes have already diverged.

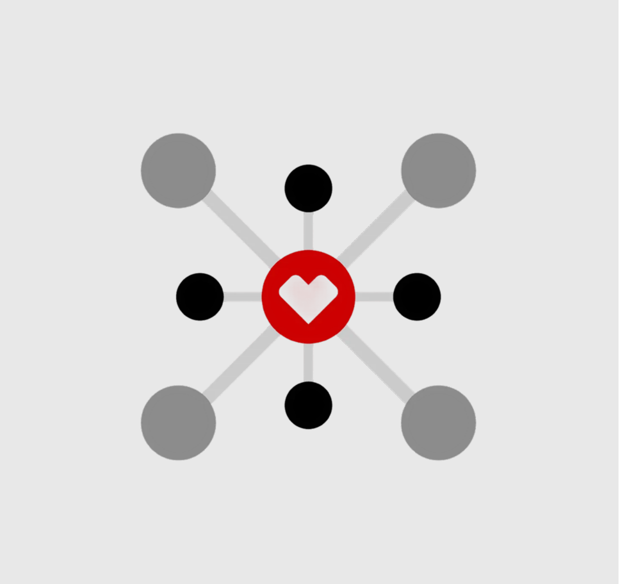
Stage 2.

Dynamic, individual-level simulation: longitudinal experiences

Once CVS Health establishes a stable representation of individuals, the next extension is time. Agents are allowed to experience products and services over time to update beliefs as events unfold. Agents click through real or faithful prototype interfaces, receive messages in sequence, “wait” between steps, and report back on comprehension, friction, and preference changes. The standard of proof rises accordingly. Organizations replay past changes to see whether the simulator reproduces known trends (retrodiction), monitor drift as the world moves, and recalibrate when the underlying population or experience changes.

This opens a new class of questions that are difficult to answer with traditional research alone: *Which part of onboarding drives drop-off? How does attitude shift after a service recovery? When does habit formation begin?* Longitudinal simulation provides a way to reason about these dynamics before they are visible in operational or clinical data.

CVS Health can use calibrated agents to “live” key journeys–pharmacist access, perceived wait times, and message clarity–over days and weeks. The runs reveal where satisfaction bends, when refill and adherence intentions strengthen or weaken, and which reminders, education content, or benefit designs meaningfully change behavior before any in-market rollout.



Stage 3.

Multi-agent simulation: conversations and coordination

Many outcomes in business emerge from interaction: customers questioning associates, pharmacists counseling patients, partners negotiating terms. In this stage, agents engage one another under realistic rules and constraints so that conversations, coordination, and contagion effects can be studied before they are lived. Leadership teams rehearse investor Q&A and pressure-test different disclosure choices; service organizations test escalation paths and objection handling across segments; product teams evaluate how peer influence shapes trial and adoption. Validation shifts from item-level agreement to behavioral signatures: do question distributions, escalation rates, or resolution pathways match what transcripts and logs show in the real world? Governance widens too, with guardrails on tone, fairness, and safety becoming part of the operating model.

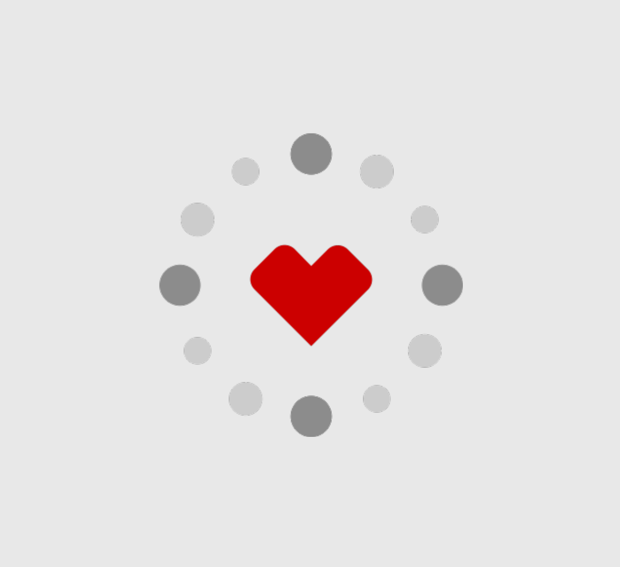
For CVS Health, this stage opens the door to examining frontline interactions – such as pharmacist-patient counseling or support workflows – without assuming that conversational behavior can be fully scripted or optimized. Simulations can be used to stress-test phrasing, sequencing, and escalation policies, highlighting where small interaction design choices may amplify confusion, build trust, or lead to changes in resolution rates by segment.

Stage 4.

World simulation: markets, networks, and place

Finally, agents are embedded in a world (e.g., a market, network, or geography) with rules for resources, incentives, and feedback loops. This is where leaders reason about second-order effects and equilibria: how a benefit change might ripple through competitors' responses; how a new service diffuses across communities; how channel policies alter partner behavior and customer choice. The goal is not to produce a single "answer" but to rank scenarios, expose trade-offs, and stress-test strategies before committing capital. Credibility depends on humility and discipline: calibrate the environment with historical data, insist on retrodictive checks ("would we have predicted last year?"), and pair simulation with targeted fieldwork where small errors would have large consequences.

For CVS Health, this stage builds on validated individual and interactional representations to explore questions such as how alternative service bundles might diffuse across communities, how store footprint or access decisions could reshape utilization patterns, or how competitors might respond to shifts in offering or positioning. This enables investment to flow first to the options with the strongest simulated advantage.



CVS Health as a case study of real-world impact

Over the past year, CVS Health partnered with the Simile team to build a significant simulation capability. That foundation created a persistent, high-fidelity lens on customers and competitors without repeatedly fielding new studies.

CVS Health's early partnership focused on the first two stages of the playbook. Calibrated agents were first used to reproduce findings from prior research, allowing teams to pre-screen ideas and reserve fieldwork for the most promising directions. This "always-on" panel expanded coverage of niche segments and reduced concept-testing cycles from weeks to hours. Dynamic agents were then allowed to "experience" key customer journeys (e.g., access to pharmacists, wait times, and message clarity) and report how perceptions shifted over time. The result was a clearer understanding of what drives satisfaction, adherence, and competitive differentiation.

Together, these early stages delivered four tangible advantages:

Faster validation of known insights.

Simulations have reproduced conclusions from prior research quickly, letting teams pre-screen ideas and hypotheses, and reserve fieldwork for the most promising directions.

Sharper NPS and experience drivers.

By testing end-to-end journeys–including access to pharmacists, wait times, and communication clarity–CVS Health has isolated which levers matter most for satisfaction and where improvements will have the most impact across segments. This proved especially valuable for hard-to-reach populations such as patients with chronic conditions who are slow, expensive, or unevenly represented in traditional surveys and panels, yet disproportionately influence health outcomes.

Adherence and behavior change under real constraints.

Dynamic agents enabled teams to test reminder cadences, education content, and benefit designs to see which combinations increase intent to refill or adopt clinical services–before running costly pilots. Critically, this made it possible to explore questions involving sensitive health behaviors, privacy-constrained data, and second-order effects that are difficult or sensitive to study directly in the wild.

Differentiation and competitive positioning.

By benchmarking perceptions against grocery and mass competitors in simulations, CVS Health identified which experience elements most clearly differentiate, and where investment would actually move the needle. These simulated results were then used to prioritize and validate downstream pilots and experiments, tightening the feedback loop between strategy, experimentation, and real-world outcomes.

CVS Health has begun to explore Stage 3 in select, high-stakes contexts. Multi-agent simulations now inform service interactions and front-line experience design. While early, this work is being used to stress test changes to phrasing, sequencing, and policies related to key messages before they are deployed in the field.

Looking ahead, Stage 4 is on the roadmap for the coming year. With validated agents and conversational dynamics in place, the next step is to embed them in market and geographic models that explore second-order effects–how alternative service bundles, store-footprint choices, or competitor reactions could play out over a planning cycle. Those simulations will not replace fieldwork, but they will rank scenarios and direct investment toward the options most likely to deliver differentiated outcomes.

The throughline is simple: start with individuals, prove fidelity, add time, then interaction, and only then the environment–each step governed and validated on its own terms. That sequence is how simulation matures from a clever experiment into a durable capability.

Towards Responsible Simulations

Simulation of people is a new capability uniquely aligned with business value. It augments existing methods and also enables net-new questions that were previously impractical. That power demands new norms and clear governance–for internal stakeholders (e.g., insights and data-science teams) and external ones (e.g., the panel members whose data seed the agents). It also calls for reporting standards so decision-makers can calibrate trust in simulated results.

Building a reporting standard

Like any model, simulations can err. Continuous back-testing across representative use cases establishes baseline accuracy, but decision-makers need a clear, repeatable way to cite simulated evidence. Think of this as the analogue to the p-value in inferential statistics: a convention that helps leaders interpret confidence. Simulation platforms should expose a calibrated confidence score tied to the likelihood that an output is accurate. Today we compute such scores using a mix of factors (including underlying model log-probabilities and historical back-test performance). For CVS Health, this means reporting a 0–100% reasoning confidence based on supporting evidence, with the threshold tuned to organizational risk tolerance. Over time, the field may converge on a standard.

Simulations as augmentation, not replacement

New capabilities bring change. In the organizations adopting simulation most successfully, the technology is framed as an augmentation of human judgment, not a replacement for existing functions. Where budgets are constrained, simulation can unlock new capacity; but the highest-value use cases keep experts in the loop to pose meaningful questions, stress-test outputs, and design the next experiment. Practically, that means thoughtful change management: start with a few high-impact pilots; instrument rigorous A/B comparisons against incumbent methods; integrate the tool into analysts' flow of work (not as a separate destination); establish a center of excellence for methods, ethics, and vendor selection; and publish transparent, periodic accuracy and bias audits.

Fundamentally, CVS Health's simulations are grounded in real people, not abstract or purely synthetic personas. In the truest sense, CVS Health views them as an amplification of people's voices (often, also of people who are typically harder to reach by traditional research). Using Simile's generative-agent technology allows CVS Health to consult customers even when direct engagement is too slow, sensitive, or costly, extending customer insight into moments where traditional methods fall short.

Used this way, CVS Health treats simulation as a durable organizational capability: a way to reason about customers and markets that is faster, more adaptable, and more comprehensive than any one study, and a way to design products, policies, and marketplaces that are robust to the human realities in which they operate.

CVS Health®

We care.

We show up with compassion and empathy for our customers and our colleagues.

We innovate with purpose.

We listen, adapt and collaborate to develop leading solutions.

We are accountable.

We operate with transparency and integrity to fulfill our commitments.

We prioritize safety and quality.

We set a high bar, with safety and quality at the center of all we do.

Author

Srikant Narasimhan

VP, Enterprise Customer Experience
and Insights CVS Health

The source for data in this document is CVS Health Enterprise Analytics, unless otherwise noted.