JPMSA

JOURNAL OF THE PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

SPRING 2026 | IN THIS ISSUE:

pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

Become a PMSA Member

GAIN ACCESS TO EXCLUSIVE CONTENT, GET INSPIRED BY LEADERS IN THE INDUSTRY, AND EXPAND YOUR PROFESSIONAL NETWORK.

Joining the Pharmaceutical Management Science Association (PMSA) offers several valuable benefits. Firstly, you'll be granted access to a comprehensive online archive containing a wealth of presentations from past events such as Conferences, Summits, and Symposia.

Additionally, membership includes access to the *PMSA Journal*, a respected publication that features research articles, case studies, and reviews relevant to the industry.

Moreover, PMSA is excited to announce the upcoming launch of a new content repository. This digital resource will further enrich your knowledge base, offering an even broader spectrum of information and tools tailored to the needs of pharmaceutical management professionals.

SCAN TO JOIN

pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

LEARN MORE AT
www.pmsa.org

Table of Contents

Table of Contents, continued

READ THE **DIGITAL VERSION** OF THE JOURNAL

Scan the QR code or visit pmsa.org/resources/pmsa-journal.

Official Publication of the Pharmaceutical Management Science Association (PMSA)

The mission of the Pharmaceutical Management Science Association not-for-profit organization is to efficiently meet society's pharmaceutical needs through the use of management science.

The key points in achieving this mission are:

i. Raise awareness and promote use of Management Science in the pharmaceutical industry

ii. Foster sharing of ideas, challenges, and learning to increase overall level of knowledge and skill in this area

iii. Provide training opportunities to ensure continual growth within Pharmaceutical Management Science

iv. Encourage interaction and networking among peers in this area

Please submit correspondence to:

Pharmaceutical Management Science Association
1024 Capital Center Drive, Suite 205
Frankfort, KY 40601
info@pmsa.org
(877) 279-3422

Executive Director:
Stephanie Czuhajewski, MPH
sczuhajewski@pmsa.org

PMSA Board Of Directors

The *Journal of the Pharmaceutical Management Science Association (ISSN 2473-9685 (print) / ISSN 2473-9693(online))* is published annually by the Pharmaceutical Management Science Association.

PMSA Journal: Spotlighting Analytics Research

PMSA is pleased to announce the 2026 Journal of the Pharmaceutical Management Science Association (PMSA), the official research publication of PMSA

The Journal publishes manuscripts that advance knowledge across a wide range of practical issues in the application of analytic techniques to solve Pharmaceutical Management Science problems, and that support the professional growth of PMSA members. Articles cover a wide range of peer-reviewed practice papers, research articles and professional briefings written by industry experts and academics. Articles focus on issues of key importance to pharmaceutical management science practitioners.

If you are interested in submitting content for future issues of the Journal, please send your submissions to info@pmsa.org.

GUIDELINES FOR AUTHORS
Summary of manuscript structure:
An abstract should be included, comprising approximately 150 words. Six key words are also required. All articles and papers should be accompanied by a short description of the author(s) (approx. 100 words).

Industry submissions: For practitioners working in the pharmaceutical industry, and the consultants and other supporting professionals working with them, the Journal offers the opportunity to publish leading-edge thinking to a targeted and relevant audience.

Industry submissions should represent the work of the practical application of management science methods or techniques to solving a specific pharmaceutical marketing analytic problem. Preference will be given to papers presenting original data (qualitative or quantitative), case studies and examples. Submissions that are overtly promotional are discouraged and will not be accepted.

Industry submissions should aim for a length of 3000-5000 words and should be written in a 3rd person, objective style. They should be referenced to reflect the prior work on which the paper is based. References should be presented in Vancouver format.

Academic submissions: For academics studying the domains of management science in the pharmaceutical industry, the Journal offers an opportunity for early publication of research that is unlikely to conflict with later publication in higher-rated academic journals.

Academic submissions should represent original empirical research or critical reviews of prior work that are relevant to the pharmaceutical management science industry. Academic papers are expected to balance theoretical foundations and rigor with relevance to a non-academic readership. Submissions that are not original or that are not relevant to the industry are discouraged and will not be accepted.

Academic submissions should aim for a length of 3000-5000 words and should be written in a third person, objective style. They should be referenced to reflect the prior work on which the paper is based. References should be presented in Vancouver format.

Expert Opinion Submissions: For experts working in the Pharmaceutical Management Science area, the Journal offers the opportunity to publish expert opinions to a relevant audience.

Expert opinion submissions should represent original thinking in the areas of marketing and strategic management as it relates to the pharmaceutical industry. Expert opinions could constitute a review of different methods or data sources, or a discussion of relevant advances in the industry.

Expert opinion submissions should aim for a length of 2000-3000 words and should be written in a third person, objective style. While references are not essential for expert opinion submissions, they are encouraged and should be presented in Vancouver format.

Industry, academic and expert opinion authors are invited to contact the editor directly if they wish to clarify the relevance of their submission to the Journal or seek guidance regarding content before submission. In addition, academic or industry authors who wish to cooperate with other authors are welcome to contact the editor who may be able to facilitate useful introductions.

Thank you to the following reviewers for their assistance with this issue of the *PMSA Journal:*

Nathan Corder, Eli Lilly and Company

Shihan He, Novo Nordisk

Ewa Kleczyk, Prolaio

J.P. Tsang, Bayser

James Young, UCB

Appala Venkata Subhadra Raju Dantuluri, Incyte Corporation

Editor: Srihari Jaganathan, UCB

Reframing Valuation and Capital Structure in Life Sciences: A Complexity-Informed AI Decision Framework

Samuel Renwick, CFA – RNA Advisors
Jerry A. Rosenblatt, PhD – Rosenblatt Life Science Consultants

Abstract: Pharmaceutical and biotechnology decisions are routinely supported by linear tools such as discounted cash flow, risk-adjusted net present value (rNPV), and static scenario analyses. While useful, these tools can under-represent how value emerges in life sciences, where clinical evidence evolves, physician adoption follows network patterns, payers adjust access rules in response to outcomes, and competitors react strategically. This article proposes a complexity-informed, data-adaptive framework that treats valuation and capital structure as dynamic systems. The approach combines agent-based stakeholder behavior models, probabilistic forecasting, Bayesian updating using real-world data consistent with evolving regulatory expectations, and decision analytics grounded in real options and value-of-information theory. Practical modeling patterns, governance considerations, and implementation steps are provided for forecasting, market access, valuation, and financing decisions.

Keywords: complexity science, life-science valuation, rNPV, capital structure, agent-based modeling, Bayesian updating, real-world evidence, real options, pharmaceutical forecasting

Introduction

The pharmaceutical and biotechnology industries operate within environments defined by deep uncertainty, nonlinear dynamics, and dense interdependencies among clinical, regulatory, commercial, and financial stakeholders.[1,2] Drug development timelines are notoriously unpredictable; clinical data mature over years and across heterogeneous patient populations; payer and physician behavior responds unevenly to new evidence; and capital markets move in cycles that materially shape the feasibility of financing and exit strategies.[3,4,5] Despite this complexity, many of the tools used to value assets and model ownership in the life sciences still assume a relatively stable and predictable world.

Traditional discounted cash flow (DCF) models, risk-adjusted net present value (rNPV) calculations, static scenario analyses, and spreadsheet-based capitalization tables have long served as the workhorses of life-science financial planning.[6] These tools have substantial merits: they impose analytical discipline, offer a common vocabulary across functional groups, and satisfy stakeholder demands for quantitative rigor. However, they share an implicit assumption: that outcomes are usefully approximated by average expectations, that risks can be captured through discount rates or success probabilities, and that the relationships among key variables are linear and independent.[7]

This article proposes an alternative: a complexity-informed, data-adaptive framework that treats valuation and capital structure as components of a dynamic system. Rather than replacing established tools (DCF, rNPV, scenarios, cap tables), the framework extends them by incorporating explicit heterogenous stakeholder behavior models, probabilistic outcome distributions and correlated risks, Bayesian updating with real-world data/evidence (RWD/RWE), and decision analytics grounded in real options theory and value-of-information analysis. The goal is not analytical sophistication for its own sake but practical improvement in how life-science organizations allocate capital, structure transactions, and sequence strategic decisions under genuine uncertainty.

For clarity, “AI” in this context should be interpreted broadly as analytical automation and decision-support: simulation, probabilistic inference, machine-learning-assisted parameter estimation, and pipeline-driven updating. The framework does not require generative AI to function, although large language models may be useful in specific, governed workflows (e.g., monitoring regulatory/competitive text signals and converting them into structured features).

The importance of these improvements is underscored by recent empirical work. Rosenblatt et al.[8] demonstrated that rare and ultra-rare disease products follow distinct ‘R-shaped’ uptake curves, reaching 40% of peak patient volume within one year and peak penetration within five years, significantly faster than non-rare products (see Figure 1). Their analysis also showed that drivers such as efficacy, company size, and unmet need interact in complex, nonlinear ways that traditional static models do not capture. These findings illustrate precisely the kind of dynamic, stakeholder-driven behavior that a complexity-informed framework is designed to accommodate.

Practitioner takeaways

For pharmaceutical practitioners, the practical output of a complexity-informed framework is not a “more complex spreadsheet.” It is a decision workflow that (a) quantifies uncertainty transparently, (b) links evidence-generation plans to economic value, and (c) makes the incentive effects of financing terms explicit. In practice, the framework supports:

- Forecasts that reflect network diffusion and payer feedback, not only smooth penetration curves.
- Valuations expressed as distributions (with tail behavior) rather than single numbers.
- Adaptive updates that incorporate clinical readouts, payer decisions, and early launch signals as formal Bayesian evidence updates.
- Explicit valuation of managerial flexibility (expand, defer, partner, abandon) and prioritization of the highest-value uncertainty-reduction activities (VOI).

- Cap-table and term-sheet analysis that forecasts incentive alignment and “preference overhang” dynamics under realistic exit distributions.

Limitations of Linear Valuation Models

The standard toolkit for life-science valuation centers on rNPV, a methodology that assigns probability-weighted values to future cash flows, discounting them to the present.[6,7] While rNPV provides a disciplined starting point, its limitations become consequential when applied to assets operating in environments characterized by feedback loops, path dependence, and emergent behavior.

Point Estimates and the “Illusion of Precision”

Traditional rNPV models typically rely on single “most likely” projections for peak sales, market share, pricing, and development timelines. These point estimates, while convenient, can create an illusion of analytical precision that obscures the true range of possible outcomes.[9,10] In the real world, Life-science assets exhibit distributions of outcomes that are frequently skewed and heavy-tailed: a small number of programs generate outsized returns while the majority produce modest or negative value.[11] Smoothing these realities into averages can simultaneously understate both downside risk and upside opportunity, leading to systematically biased capital allocation decisions.

Static Assumptions in a Dynamic Environment

Conventional models also tend to treat key parameters (market size, penetration rates, pricing, competition) as scenario-capable fixed inputs rather than dynamic variables that co-evolve with the system.[12] In practice:

- Patient response varies across subpopulations and over time.
- Physician adoption follows network-mediated diffusion patterns rather than the smooth S-curves typically assumed.[13]
- Payers adjust coverage and formulary positions in response to accumulating evidence and competitive dynamics.
- Competitors react strategically to both successes and failures in adjacent programs.
- Regulatory and HTA expectations can shift with class experience, safety signals, and evolving endpoints.

These interactions create feedback loops and path dependence that can amplify or dampen value trajectories in ways that static spreadsheet models are structurally unable to represent.

The result is not merely forecast error; it is decision error (e.g., mistimed financing, mis-scoped launches, suboptimal evidence plans, and avoidable misalignment in transaction economics).

One practical signal that static assumptions are breaking down is the shape and speed of adoption. In some rare and ultra-rare disease settings, uptake can be front-loaded as concentrated specialist networks converge rapidly on effective therapies, while broader primary-care markets may diffuse more slowly. Figure 1 provides a conceptual schematic of these contrasting patterns.

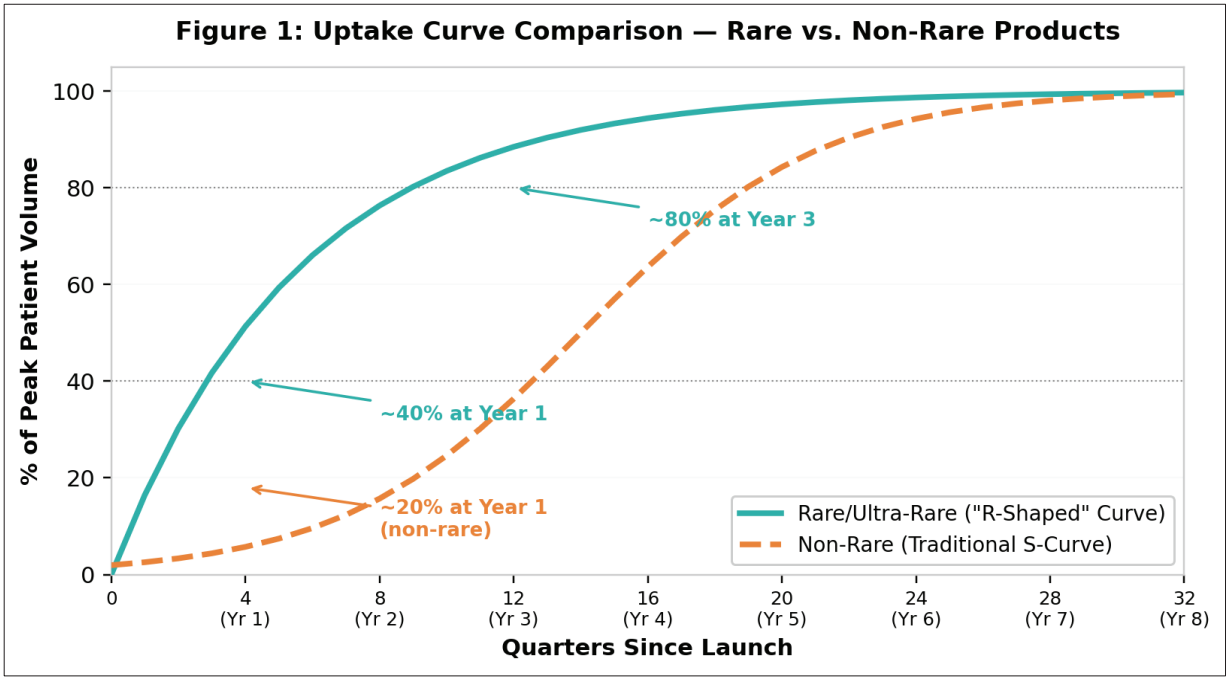


***Figure 1**: Rare/ultra-rare disease products follow steep "R-shaped" uptake curves, reaching peak penetration significantly faster than non-rare products (adapted from Rosenblatt[8] et al.)*

Capital Structure as an Incentive System - Not a Snapshot

Similarly, capitalization tables are often treated as static ownership snapshots. However, in reality, liquidation preferences, conversion terms, anti-dilution provisions, option pools, milestone tranches, and debt covenants create an interconnected incentive landscape. When exit outcomes are uncertain and heavy-tailed, the shape of the distribution interacts with terms in non-obvious ways: the same cap table can be motivating under high-tail outcomes but demotivating under moderate exits, or vice versa. Treating capital structure as a dynamic system helps teams anticipate these incentive effects before they become constraints.

Complexity Science as a Lens for Life-Science Valuation

Complexity science provides a rigorous conceptual foundation for understanding how value emerges in systems characterized by many interacting agents, feedback loops, nonlinear dynamics, and adaptation.[2,14] Life-science valuation is a natural domain for this lens because the underlying subject matter, human biology itself, is the prototypical complex adaptive system (see Table 1).

Table 1: Comparison of Traditional and Complexity-Informed Valuation Approaches

Dimension	Traditional Approach	Complexity-Informed Approach
Revenue Forecast	Single peak-sales estimate with market-share curve	Distribution of trajectories via agent-based adoption models
Risk Quantification	Discount rate or binary stage-gate probabilities	Monte Carlo simulations with correlated risk factors
Competitive Dynamics	Static competitive landscape assumed at model date	Game-theoretic and agent-based competitor response models
Data Integration	Periodic manual updates to spreadsheet inputs	Bayesian updating with streaming RWD/RWE
Strategic Optionality	NPV of single “expected” pathway	Real options valuation of strategic flexibility
Stakeholder Behavior	Aggregated assumptions (e.g., average physician)	Heterogeneous agent behavior (KOLs, payers, patients)
Capital Structure	Static cap table snapshot	Dynamic system of incentives, triggers, and feedback
Output Format	Single NPV or scenario-based valuation range	Probabilistic value distributions with confidence intervals

Human Biology as the Native Complex System

The biology metaphor maps directly commercialization and valuation. A practical way to ground the complexity lens in life sciences is to treat human biology as the “native” complex adaptive system that drug developers are attempting to influence.[15] Biological systems exhibit hallmark features of complexity: feedback loops that regulate homeostasis, nonlinear dose-response relationships where small agitations/modifications can produce large effects (or vice versa), compensatory pathways that buffer against intended therapeutic interventions, and multiple stable states that can shift as therapies are introduced, adjusted, or combined.[16]

Across the continuum from discovery through commercialization, value creation can be understood as the process of learning to reliably shake up, and then stabilize, an adaptive system under uncertainty. Early-stage signals are inherently noisy and context-dependent. Effects can be delayed, emergent, or mediated by interactions among mechanism of action, patient heterogeneity, clinical protocol design, and real-world prescribing behaviors. In biopharma, where clinicians and scientists already reason in systems terms, a complexity-

informed valuation framework is not an abstraction imposed from outside the domain; rather, it reflects how therapeutic impact actually unfolds.

From Biological Complexity to Financial Modeling

This biological metaphor maps directly to forecasting and valuation challenges. Just as a therapeutic molecule must navigate a complex biological system to produce clinical benefit, a life-science asset must navigate a complex commercial and financial ecosystem to generate economic value.[17] The parallels are instructive. For example: biological heterogeneity maps to market segmentation; compensatory mechanisms correspond to competitive response; dose-response nonlinearities mirror the relationship between evidence quality and payer access decisions; and, homeostatic buffering parallels the resistance to change inherent in established treatment paradigms (see Table 2).

Table 2: Mapping Biological Complexity to Valuation and Commercial Dynamics

Biological Feature	Biological Example	Commercial Analog	Valuation Implication
Feedback Loops	Cytokine signaling cascades	KOL adoption driving peer prescribing	Non-linear uptake; tipping-point dynamics
Nonlinear Dose-Response	Therapeutic window vs. toxicity	Evidence quality vs. payer access	Discontinuous value jumps at milestones
Compensatory Pathways	Resistance mutations in oncology	Competitive entry and formulary switching	Dynamic market share erosion modeling
Multiple Stable States	Remission vs. relapse states	Market equilibria pre/post-disruption	Distinct equilibrium scenario paths
Patient Heterogeneity	Pharmacogenomic variation	Segment-level adoption patterns	Sub-population value disaggregation

Value Inflection Points and Phase Transitions

The life sciences are punctuated by value inflection points: threshold events that can reconfigure both perceived value and the appropriate modeling approach.[18] Passing from target validation to a credible preclinical package, from IND-enabling work to first-in-human dosing, from early proof-of-concept to registrational clarity, or from approval to evidence-supported market access can resemble a phase transition: the same underlying asset suddenly becomes legible to a broader market of capital providers and strategic partners.

At early stages, a smaller number of willing investors and limited liquidity often forces valuation to rely more heavily on option-like logic, sparse comparables, and high-uncertainty priors. As programs cross thresholds that reduce key uncertainties (scientific, regulatory, commercial, or reimbursement), the pool of potential financiers and acquirers expands, liquidity conditions improve, and valuation methods may appropriately shift toward comparables, partnership benchmarks, or cash-flow-anchored frameworks.[19] This discontinuity helps explain why value can change dramatically without a proportional change in “expected” outcomes, because the

market's capacity to fund, price, and realize those outcomes has changed (see Figure 2 and Table 3).

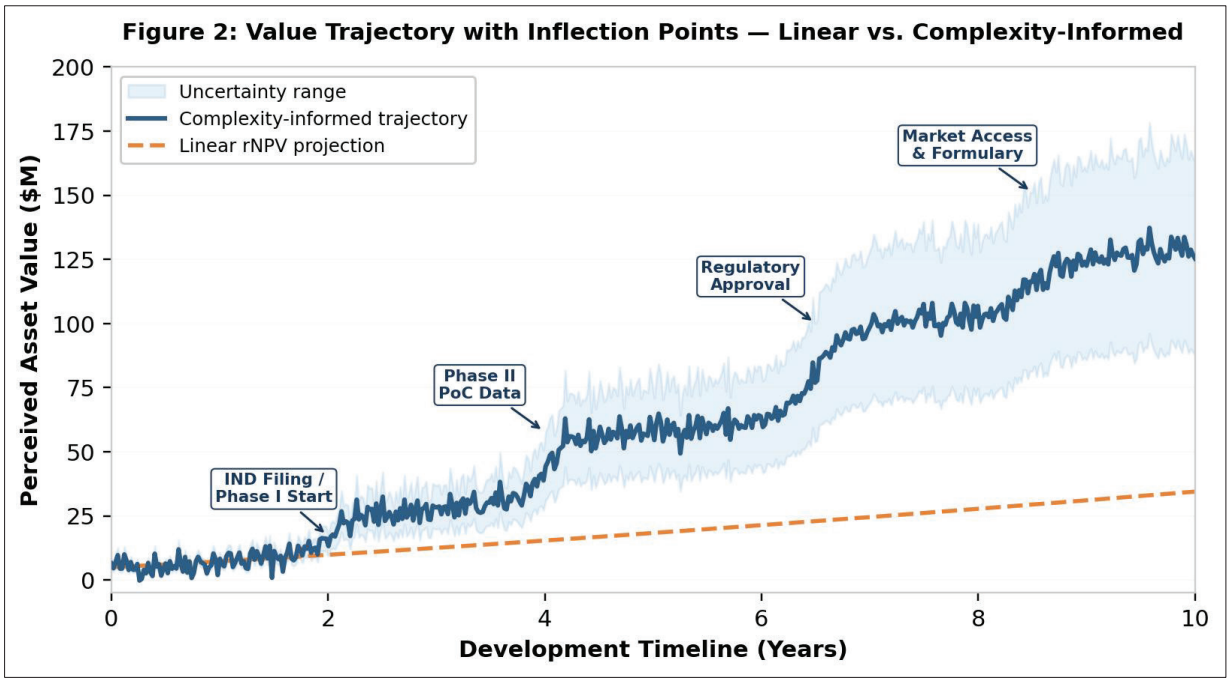


Figure 2: A complexity-informed trajectory captures discontinuous value jumps at inflection points, whereas a traditional linear rNPV projection smooths over these phase transitions.

Table 3: Value Inflection Points Across the Development Continuum

Inflection Point	Uncertainty Resolved	Investor Pool Shift	Valuation Method Shift	Early Warning Signals
Target Validation → Preclinical	Biological plausibility	Angels/grants → Seed VCs	Scorecard → Option-pricing	In-vitro reproducibility; IP clarity
IND → Phase I	Regulatory pathway feasibility	Seed → Series A/B	Option-pricing → rDCF	Tox findings; CMC scalability
Phase II PoC → Registrational	Clinical efficacy signal	Series B → Growth/crossover	rNPV with refined tree	Effect size variability; dose-response
Approval → Market Access	Regulatory/reimbursement risk	Growth → Public/strategic	rNPV → Comparables/DCF	ICER thresholds; label breadth
Launch → Established Brand	Commercial execution	Public → Institutional	DCF with RWE inputs	Early Rx trends; formulary wins

Early Warning Signals and Critical Transitions

An important advantage of a complexity-informed approach is its capacity to recognize early signs that a development program or commercial opportunity may be approaching an inflection point.[20] In practice, programs rarely fail or succeed gradually. Instead, they often experience sudden acceleration or unexpected fragility. Complexity science has identified several generic early warning signals that precede critical transitions: increasing variance in outcome measures, critical slowing down, and growing correlation among previously independent variables.[21]

Figure 3 illustrates these early warning signals in two panels. In panel (a), the blue line tracks a clinical or market metric over time. Notice that the average value of this metric does not change dramatically as it approaches the red dashed "Critical Transition" line. If a manager were watching only the average, nothing would appear alarming. However, the light blue shaded area surrounding the line, which represents the range of variation around the average, tells a different story. Starting well before the critical transition (around quarter 20 in the figure), this band begins to widen progressively (the Warning Zone). The system is still functioning, still oscillating around roughly the same average, but it is taking larger swings and recovering less cleanly. By the time the red dashed line is reached at approximately quarter 37, the system has lost enough stability that it tips into a fundamentally different state. Critically, the widening variance in the Warning Zone is not the critical transition itself. It is the *advance warning* that a transition is approaching. The gap between when the warning becomes visible and when the transition actually occurs is where the practical value lies for decision-makers. In the figure, this gap spans roughly fifteen quarters of lead time.

Panel (b) shows the same phenomenon through a different lens: autocorrelation, which measures how strongly each quarter's value resembles the previous quarter's value. When autocorrelation is low (around 0.2), the system recovers quickly from disturbances, indicating resilience. As the metric climbs toward 0.7 and beyond, the system takes progressively longer to bounce back from vacillations, a phenomenon known as "critical slowing down." The system is becoming sluggish in its ability to self-correct, and once autocorrelation crosses the warning threshold, the critical transition is likely imminent.

Two pharmaceutical examples illustrate the practical relevance of these signals.

Pharmaceutical Example 1: Consider a specialty product whose quarterly sales have historically fluctuated within a narrow and predictable band. Over several quarters, the average remains roughly on forecast, but the quarter-to-quarter swings become noticeably wider: one quarter comes in well above plan, the next falls well below, then overshoots again. A traditional review might dismiss each quarter individually as a one-time anomaly. A complexity-informed analysis would recognize the widening variance as a warning signal. The underlying commercial system, encompassing payer coverage decisions, physician prescribing patterns, competitive positioning, and patient adherence, may be losing its stability. Within a few quarters, the product

could experience a sudden and sustained shift in trajectory, either a breakout driven by a favorable formulary decision or a sharp decline triggered by new competition. The widening variance provides lead time to investigate root causes and prepare contingency plans before the shift materializes.

Pharmaceutical Example 2: Consider a Phase III clinical program monitoring its primary efficacy endpoint across interim analyses. Early in the trial, site-to-site variation in the endpoint is modest and consistent. As enrollment expands into a broader patient population and external conditions evolve (new standard-of-care guidelines, shifting referral patterns), the variation across sites begins to increase, even though the pooled average endpoint remains within the expected range. A conventional data monitoring approach focused on the mean would see a program tracking to plan. A complexity-informed approach would flag the increasing site-level variance as an early warning: the trial system may be approaching a threshold beyond which the pooled result could shift materially, potentially requiring protocol amendments, sample size adjustments, or accelerated enrollment at high-performing sites.

In both cases, the critical insight is the same: by the time the average itself moves decisively, the window for proactive intervention has often closed. Monitoring variance and autocorrelation provides earlier, actionable intelligence (see Figure 3).

Figure 3: Early Warning Signals Preceding Critical Transitions

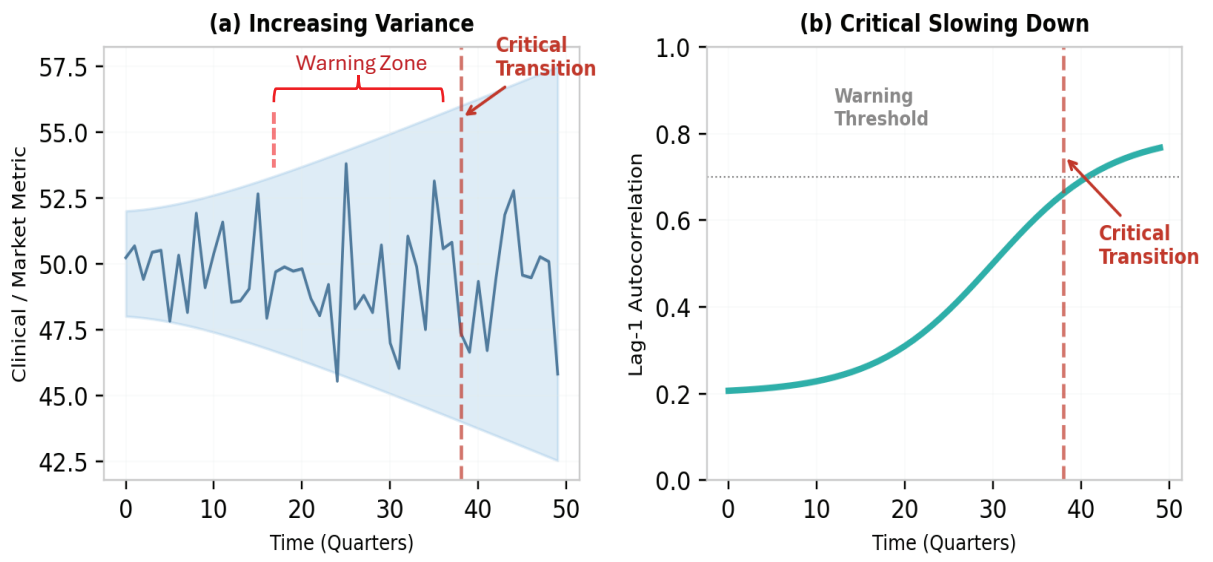


Figure 3: *Early warning signals preceding critical transitions. In panel (a), the blue line represents a clinical or market metric whose average remains roughly stable, while the light blue shaded band shows the expanding range of variation that signals an approaching transition. The red dashed line marks the critical transition point itself. In panel (b), rising autocorrelation (the system's tendency to retain its previous state rather than self-correcting) crosses a warning threshold as the system loses resilience. In both panels, the warning signals emerge well before the transition occurs, providing decision-makers with lead time for proactive action.*

Dynamic Capital Structure as a Complex System

In most life-science organizations, capitalization tables are treated as static snapshots that show ownership and economics under a limited set of financing and exit scenarios.[23] A complexity-informed perspective recognizes that capital structure functions as a dynamic system in which decisions, incentives, and constraints are interconnected. Liquidation preferences, conversion terms, anti-dilution provisions, option pools, vesting schedules, and milestone-based payments all shape the incentive landscape for founders, employees, investors, and strategic partners (see Table 4).

These incentive structures, in turn, influence decision-making around hiring pace, fundraising timing, partnership negotiations, and even development strategy. For example, a high liquidation preference stack may create misalignment between early investors seeking downside protection and founders who require upside motivation. Milestone-triggered tranches can accelerate development decisions in ways that may not optimize for long-term clinical evidence quality. Anti-dilution provisions can discourage down-rounds that might otherwise recapitalize a promising but cash-constrained program.[24] When examined over time, capital structure functions less like a spreadsheet and more like a connected system in which changes in one area propagate through the entire incentive network (see Figure 4).

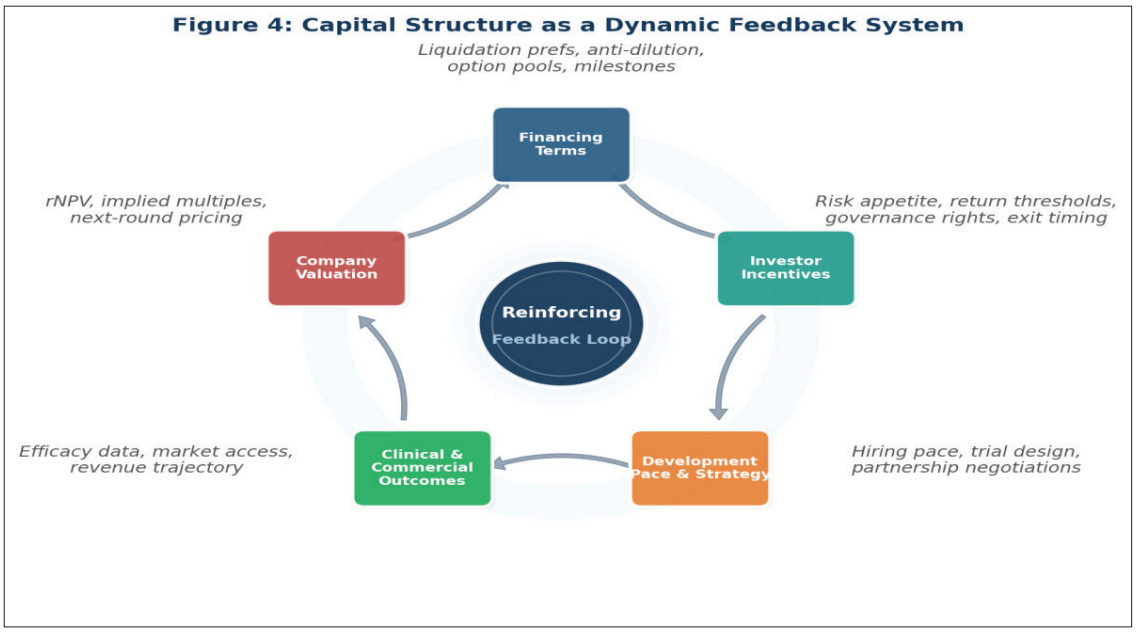


***Figure 4**: Capital structure elements form a reinforcing feedback loop in which financing terms, investor incentives, development strategy, clinical outcomes, and company valuation continuously influence one another.*

Table 4: Capital Structure Elements as Dynamic System Components

Cap Table Element	Incentive Effect	Behavioral Consequence	System-Level Feedback
Liquidation Preferences	Investor downside protection	Risk aversion in exit timing	Misalignment at moderate exit values
Anti-Dilution Provisions	Price protection for early investors	Reluctance to pursue down-rounds	Capital constraints at critical junctures
Milestone Tranches	Funding aligned with de-risking	Accelerated timelines; quality trade-offs	Financial vs. scientific development logic
Option Pool / Vesting	Talent retention and motivation	Cliff effects on key-person risk	Capability shaped by equity economics
Convertible Instruments	Bridge flexibility; deferred pricing	Dilution uncertainty	Structural complexity; governance ambiguity

A Complexity-Informed AI Decision Framework

Drawing together the preceding analysis, this section presents a unified decision framework that integrates four complementary analytical components. The framework is designed to be modular: organizations can adopt individual components incrementally, while recognizing that the greatest value arises from their integration (see Figure 5).

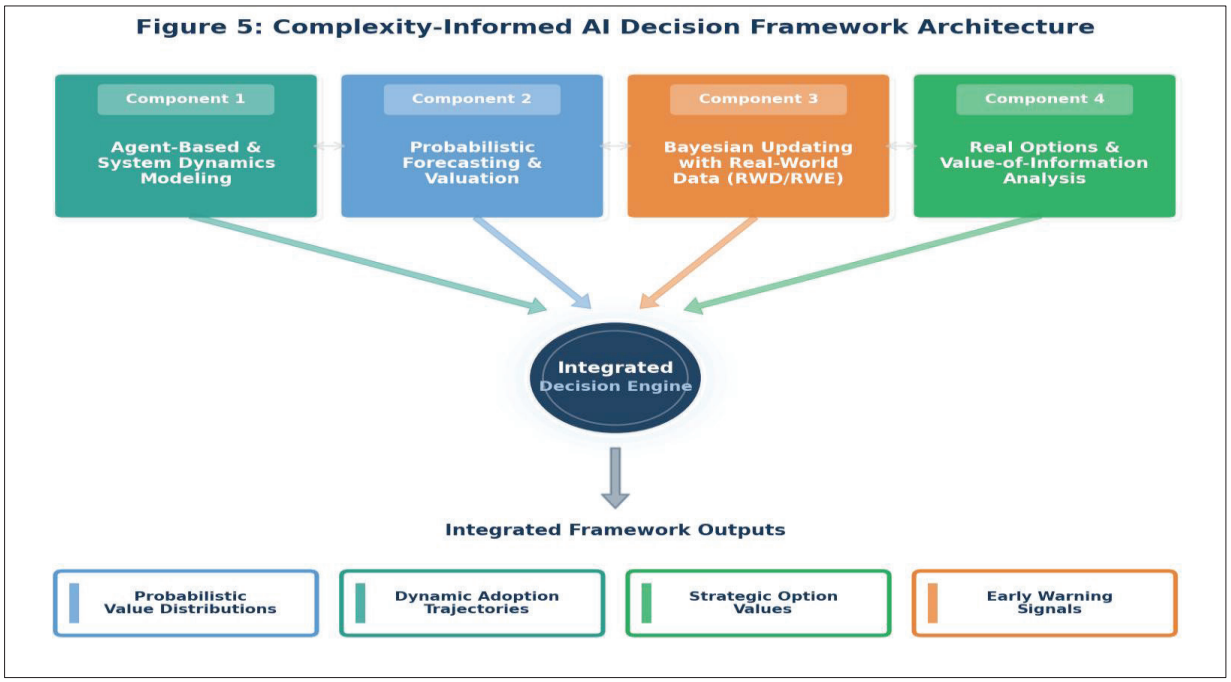


***Figure 5**: The four-component architecture of the complexity-informed AI decision framework. Each component feeds into an integrated decision engine that produces probabilistic outputs, dynamic trajectories, option values, and early warning signals.*

Component 1: Agent-Based and System Dynamics Modeling

The first component replaces aggregate assumptions with explicit behavioral models of key stakeholder groups. Agent-based models (ABMs) simulate physician prescribing decisions as functions of evidence exposure, peer influence, formulary access, and patient outcomes. System dynamics components capture feedback loops between market adoption, evidence generation, payer response, and competitive behavior.[25,26] Together, these models generate adoption trajectories that reflect the network-mediated, nonlinear patterns observed in actual pharmaceutical markets rather than the smooth penetration curves assumed by conventional forecasts.

The relevance of this approach is supported by empirical observations. As Rosenblatt et al.[8] demonstrated, physician adoption of rare disease products followed "R-shaped" uptake curves driven by efficacy, unmet need, company size, and order of entry. This is precisely the type of multi-agent, attribute-dependent behavior that ABMs are designed to model. Products marketed by larger companies reached peak patient penetration significantly earlier, likely reflecting promotional investment, geographic reach, and prior experience. These are dynamics that linear models would treat as fixed multipliers rather than emergent outcomes of agent interactions.

Component 2: Probabilistic Forecasting and Valuation

The second component replaces point estimates with explicit probability distributions for key value drivers. Monte Carlo simulation generates thousands of valuation scenarios by sampling from distributions of clinical endpoints, pricing outcomes, market timing, competitive entry, and regulatory decisions. The output is not a single NPV but a full distribution of asset values, enabling decision-makers to understand the probability of achieving specific return thresholds and the sensitivity of value to key uncertainties (see Figures 6(a) and 6(b)).[27]

Figure 6(a) illustrates this approach. The light blue histogram shows the probability density of asset valuations generated by a Monte Carlo simulation of 10,000 scenarios for a hypothetical asset. Each scenario samples from distributions of clinical success probability, peak revenue potential, pricing and reimbursement outcomes, time to market, and competitive dynamics. The resulting shape is markedly right skewed: most simulated outcomes cluster at lower valuations, but a meaningful tail extends toward much higher values. The dark blue cumulative probability curve (right axis) rises steeply through the lower valuations and then flattens, confirming that the bulk of probability mass lies well below the arithmetic mean.

Three red dashed lines mark the P10, P50 (median), and P90 percentiles of the distribution. The P50 of approximately $100M represents the outcome that is equally likely to be exceeded or undershot. However, the orange vertical line marks the traditional rNPV, which corresponds to the mean of the distribution at approximately $218M. The gap between the median ($100M)

and the mean ($218M) is not an error; it reflects the heavy right tail pulling the average upward. A decision-maker relying solely on the rNPV would anchor on $218M as the “expected” value, yet the median outcome is less than half that figure. This divergence is precisely the “illusion of precision” that probabilistic methods are designed to reveal.

The distributional view changes how decision-makers should evaluate the asset. Rather than asking “Is this program worth $218M?” the relevant questions become: What is the probability that this asset exceeds our minimum return threshold? What conditions drive the right-tail outcomes versus the modal cluster? And which uncertainties, if resolved, would most shift the shape of the distribution? These are fundamentally different, and more actionable, questions than those prompted by a single-number valuation.

Figure 6(b): A Monte Carlo simulation example: A second pharmaceutical example makes the practical value concrete. Consider a mid-stage rare disease gene therapy program. The traditional rNPV analysis produces a single estimate of $320M, derived from a base-case peak revenue of $600M, a 55% probability of technical and regulatory success, standard discount rates, and a projected launch in Year 5. This number enters the portfolio review alongside competing programs, each represented by its own point estimate, and capital allocation proceeds on the basis of rank-ordered NPVs.

A Monte Carlo simulation of the same program tells a richer story. When the model samples from distributions rather than point estimates, the output reveals a bimodal structure: there is a large cluster of scenarios near zero (reflecting the roughly 45% probability of clinical failure or regulatory rejection), and a separate, right-skewed cluster of positive outcomes ranging from $150M to over $1.2B. The positive cluster itself is wide because peak revenue depends on pricing negotiations with a small number of payers, the durability of therapeutic effect (which determines re-treatment economics), and the speed of diagnosis and referral in a disease with low clinical awareness. The $320M rNPV is the probability-weighted average of these two fundamentally different clusters, yet it corresponds to almost no individual scenario. No realistic outcome path actually produces a $320M asset.

This distributional insight directly changes the strategic conversation. The gene therapy team can now see that the dominant source of value uncertainty is not clinical efficacy (which Phase II data have partially de-risked) but rather payer access and diagnosis rate in the positive-outcome cluster. Value-of-information analysis (Component 4) would therefore prioritize a health economics study and a disease awareness initiative over additional dose-finding work. The portfolio committee, in turn, can evaluate whether the program’s bimodal profile complements or concentrates risk relative to other assets in the pipeline. None of these insights are available from the $320M point estimate alone.

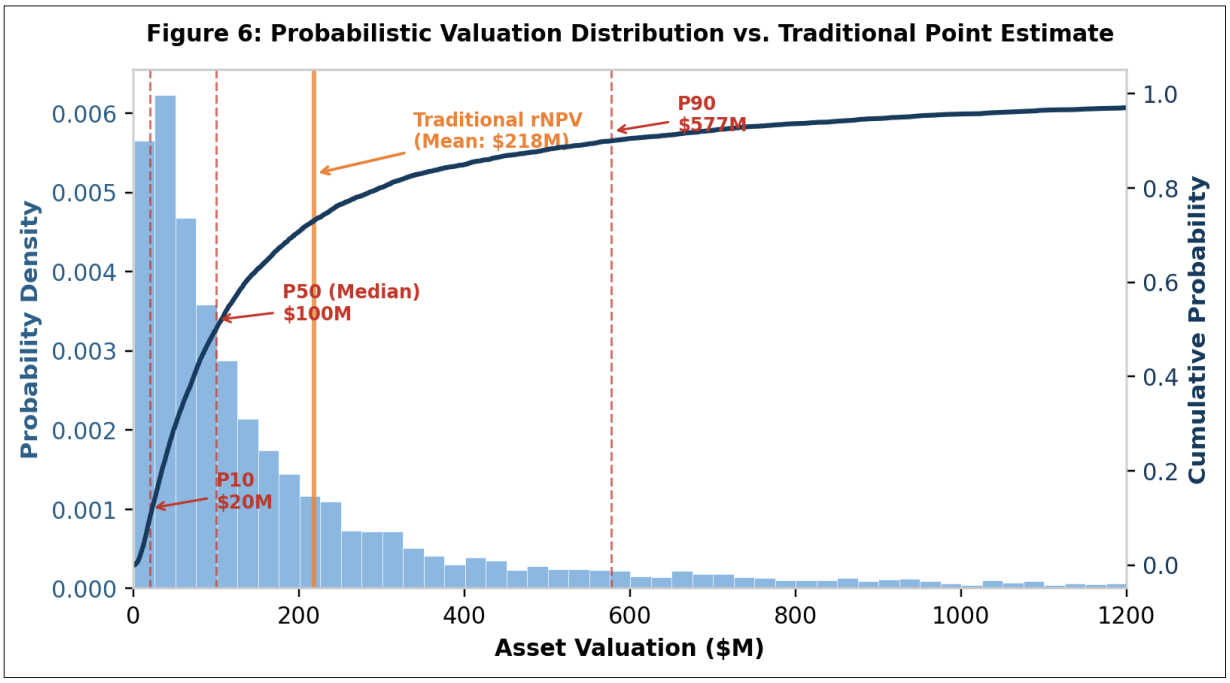


Figure 6(a): *Probabilistic valuation distribution for a life-science asset generated by Monte Carlo simulation. The light blue histogram shows the right-skewed density of possible valuations; the dark blue curve traces cumulative probability. Red dashed lines mark the P10, P50 (median), and P90 percentiles. The orange line indicates the traditional rNPV (mean), which at $218M is more than double the median of $100M due to heavy right-tail outcomes. This divergence illustrates why single-number valuations can systematically misrepresent the most likely outcome and obscure the distributional structure that drives strategic decisions.*

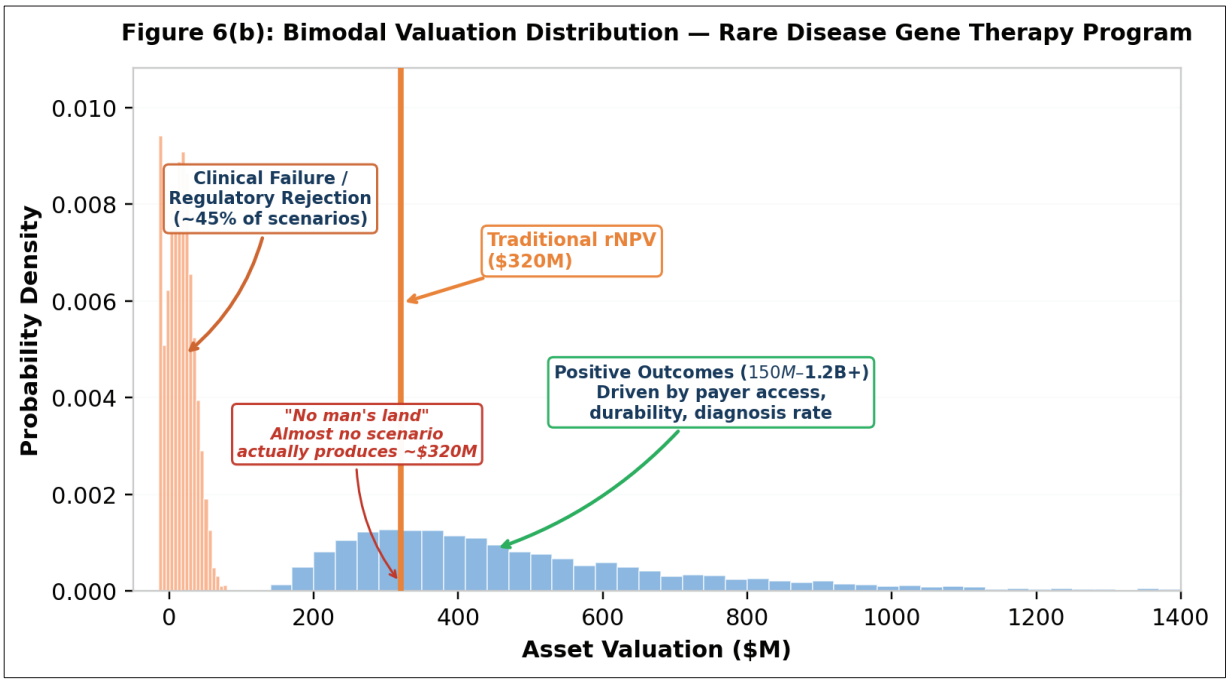


Figure 6(b): *Bimodal valuation distribution for a rare disease gene therapy program. The orange bars (left) represent the ~45% of Monte Carlo scenarios in which the program fails clinically or is rejected by regulators, clustering near zero. The blue bars (right) show the ~55% of positive-outcome scenarios, ranging from $150M to over $1.2B depending on payer access, therapeutic durability, and diagnosis rate. The orange vertical line marks the traditional rNPV of $320M, which falls in the gap between the two clusters, illustrating why the point estimate corresponds to almost no realistic outcome path.*

Component 3: Bayesian Updating with Real-World Data

The third component introduces systematic Bayesian updating of model assumptions as new data become available. Real-world data (RWD) from electronic health records (EHRs), claims databases, registries, and post-marketing surveillance[28] provide continuous signals about treatment response, adoption patterns, persistence, and safety. Rather than treating RWD solely as post-launch confirmation, the framework incorporates real-world evidence (RWE) as a mechanism for updating prior assumptions throughout the development lifecycle, consistent with evolving regulatory expectations from the FDA and EMA.[29]

From a governance standpoint, Bayesian updating is attractive because it makes the “why” of model changes explicit: assumptions do not drift informally, they are revised because new evidence provides input for additional assumptions. The Bayesian approach formally handles the sequential nature of evidence accumulation. Each new data release (interim analysis, advisory committee outcome, early launch prescription data, health technology assessment decision) updates the posterior distribution of asset value[8], enabling decision-makers to distinguish between genuine signals and noise. The value of this approach has been demonstrated in analogous pharmaceutical applications: for example, the use of longitudinal claims data to track rare disease product uptake in real time, enabling dynamic recalibration of forecasting assumptions as market behavior unfolds (see Figure 7).

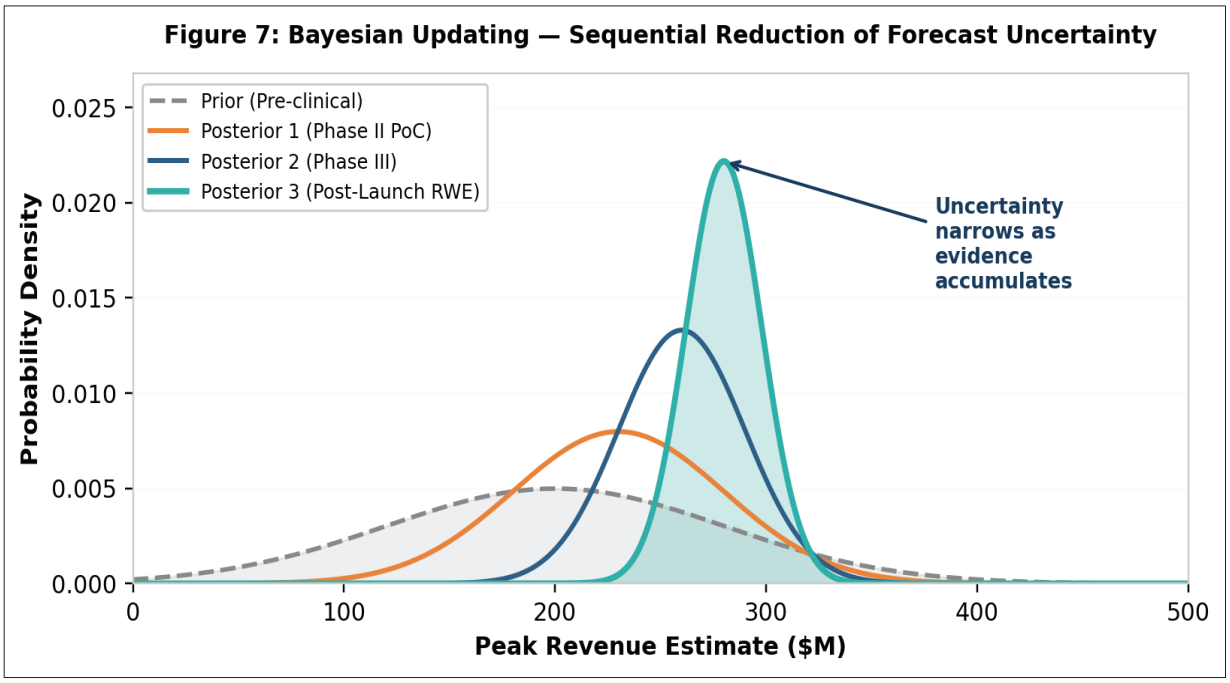


***Figure 7**: Bayesian updating progressively narrows forecast uncertainty as evidence accumulates from preclinical priors through Phase II, Phase III, and post-launch real-world evidence.*

Component 4: Real Options and Value-of-Information Analysis

The fourth component addresses strategic optionality, an area where the gap between traditional and complexity-informed approaches is perhaps most consequential. Traditional rNPV treats development as a committed pathway, discounting future cash flows as though the organization will proceed regardless of intermediate outcomes. Again, in reality, life-science development is characterized by a series of decision gates at which managers can expand, contract, defer, abandon, or pivot based on accumulating evidence.[30,31]

Real options analysis (ROA) explicitly values this managerial flexibility. Drawing on financial options theory, ROA recognizes that an early-stage program is not simply a discounted stream of expected cash flows; it is a portfolio of options whose value depends on the volatility of outcomes, the cost of exercising the next stage, and the time available before decisions must be made. In a life-science context, the option to expand might correspond to pursuing additional indications after a successful proof-of-concept; the option to defer could represent waiting for a competitor's readout before committing to a registrational trial design; and the option to abandon reflects the value of limiting losses when early clinical signals are unfavorable.

Value-of-information (VOI) analysis complements ROA by answering a distinct but related question: what is the expected economic value of reducing a specific uncertainty before making

an irreversible commitment?[32] Where ROA values the flexibility to change course, VOI quantifies how much that flexibility is worth for each source of uncertainty, and therefore which evidence-generation activities should be prioritized. For example, if the dominant source of valuation uncertainty is payer access rather than clinical efficacy, VOI analysis may indicate that investing in a health economic outcomes study generates more decision-relevant information than conducting an additional dose-finding cohort.

Together, ROA and VOI provide a principled basis for sequencing strategic decisions and prioritizing evidence generation. The logic operates as follows: ROA identifies the set of strategic options available at each decision gate and quantifies their combined value. VOI then identifies which uncertainties, if resolved, would most increase the value of those options, thereby directing evidence-generation resources toward the highest-value questions. This integration is particularly powerful in portfolio contexts, where organizations must allocate limited capital across multiple programs, each with different option structures and information gaps.

Consider a practical illustration. A mid-stage oncology company has three programs approaching decision gates: Program A requires a go/no-go on a Phase III investment; Program B faces a partnership-versus-independent-development decision; and Program C must choose between two indication strategies. ROA reveals that Program B has the highest option value due to the availability of a partnership structure that limits downside exposure. VOI analysis then shows that the key uncertainty for Program B is not clinical efficacy (already de-risked by Phase II data) but rather commercial differentiation versus an emerging competitor. The optimal evidence-generation investment is therefore a focused competitive positioning study, an insight that would not emerge from a traditional rNPV comparison of the three programs (see Table 5).

Table 5: *How Real Options Analysis and Value-of-Information Work Together to Sequence Decisions*

Decision Gate	ROA Question	Real Option Types	VOI Question	Priority Evidence Activity	Integrated Decision Output
Preclinical → IND	What is the option value of entering the clinic vs. deferring or redirecting?	**Defer** (wait for data); **Abandon** (terminate program); **Switch** (backup molecule, salt form, modality); **Expand** (additional disease models to create indication optionality); **Stage** (IND-enabling work in tranches)	Which uncertainty most limits option value: target biology / mechanism translatability, toxicology, CMC scalability, competitive / IP landscape, or breadth of mechanism across indications?	Targeted tox study; mechanistic biomarker validation; additional disease-model screening (if mechanism-VOI is dominant); FTO / IP landscape analysis; CMC feasibility assessment	Proceed to IND if highest-VOI uncertainty is resolved. Invest in indication breadth if mechanism-VOI dominates. Switch to backup molecule if tox-VOI is high and backup exists. Defer if key uncertainties remain unresolved.
Phase I → Phase IIa / PoC	What is the option value of advancing to proof-of-concept vs. pausing for additional translational data?	**Stage** (small open-label PoC before larger study); **Defer** (wait for competitor readouts or biomarker data); **Abandon** (unacceptable FIH safety/PK); **Switch** (dose, schedule, formulation, patient selection strategy); **Expand** (explore additional patient subpopulations)	Is the dominant uncertainty translational (does the mechanism engage the target in humans?), safety/tolerability at therapeutic doses, PK/PD relationship, or patient selection / enrichment strategy?	Biomarker-driven PoC design; PK/PD modeling and dose-response characterization; patient enrichment / selection biomarker study; translational bridging study from animal to human	Advance to PoC if translational-VOI is low (FIH data confirm target engagement). Redesign dose/schedule if PK-VOI dominates. Invest in biomarker enrichment strategy if patient-selection-VOI is highest. Abandon if safety signals preclude therapeutic window.
Phase IIb → Phase III	What is the option value of a registrational investment vs. partnering or licensing at this stage?	**Expand** (additional arms, indications, geographies); **Contract** (narrow to most responsive subpopulation); **Partner / License** (transfer partial rights, retain residual optionality); **Stage** (adaptive trial design with interim analyses); **Defer** (await competitor or HTA signals); **Compound option** (Phase III success creates launch,	Is residual uncertainty concentrated in efficacy magnitude (effect size durability), safety / tolerability profile, commercial differentiation vs. competitive set, or payer access / reimbursement likelihood?	Dose-ranging confirmation study; competitive landscape and differentiation analysis; payer advisory board or HTA simulation; focused bridging study in target subpopulation; regulatory pre-submission meeting	Invest in Phase III if efficacy-VOI is low (already de-risked by IIb). Partner if commercial-VOI is dominant and partner brings market access capability. Contract to enriched population if effect-size-VOI is high in broad population. Stage via adaptive design if multiple uncertainties are balanced.

Decision Gate	ROA Question	Real Option Types	VOI Question	Priority Evidence Activity	Integrated Decision Output
		lifecycle, and platform options)			
Approval → Launch Strategy	What is the option value of a broad launch vs. a focused or staged market entry?	**Expand** (additional indications, line extensions, geographies); **Contract** (launch in narrow specialist subpopulation first); **Stage** (phased market entry to generate access evidence); **Switch** (reformulation, combination, delivery route); **Grow / Platform** (label expansion creates lifecycle and next-generation options)	Is the binding constraint payer access / formulary positioning, physician adoption and prescribing behavior, patient identification and diagnosis rate, or competitive timing (first-mover vs. fast-follower dynamics)?	RWE pilot program or early-access data collection; KOL engagement and medical education study; health economics / budget impact analysis for payers; disease awareness and diagnosis pathway initiative; competitive positioning research	Stage launch to generate access evidence if payer-VOI is high. Broad launch if physician-adoption-VOI dominates and KOL network is favorable. Invest in disease awareness if diagnosis-rate-VOI is the binding constraint. Prepare lifecycle strategy if competitive-timing-VOI signals first-mover advantage is time-limited.

Note: Real option types at each gate are not mutually exclusive. Multiple options may be active simultaneously, and exercising one option (e.g., staging) may create or extinguish others (compound option logic). VOI priorities may shift as evidence accumulates through Bayesian updating (Component 3), and this adaptive recalibration is a feature of the framework, not a limitation.

Integration: Turning Model Outputs into Decisions

Integration is where the framework becomes operational. A practical workflow is:

- Define the decision to be supported (e.g., Phase III commit vs. partner; launch scope; financing timing) and the decision-maker risk/return criteria.
- Define the state variables and uncertainties that matter for that decision (clinical, regulatory, access, competitive, financing).
- Run ABM/system dynamics to generate behaviorally realistic trajectories and feed them into probabilistic valuation.
- Use Bayesian updating as evidence becomes available to refresh and maintain an auditable evolution of assumptions.
- Apply ROA/VOI at each decision gate to value flexibility and rank evidence-generation activities by expected decision value.
- Translate outputs into clear governance artifacts: distributions, decision boundaries, and scenario narratives linked to observable signals.

The practical value of this integrated approach lies in its ability to convert abstract strategic discussions into quantifiable, evidence-linked decisions. Rather than debating whether to “be aggressive” or “de-risk” a program in qualitative terms, decision-makers can evaluate the

economic trade-offs of specific evidence-generation investments against specific strategic options. This is particularly important in environments where capital is constrained and the opportunity cost of misallocated R&D spending is high[29,30,31] (see Table 6).

Table 6: Integrated Framework Components and Decision Applications

Component	Primary Methodology	Decision Domain	Key Output
Stakeholder Behavior	Agent-based; system dynamics	Forecasting; market access strategy	Network-aware adoption trajectories
Probabilistic Valuation	Monte Carlo; fat-tailed distributions	Asset valuation; portfolio optimization	Value distributions with CIs
Adaptive Updating	Bayesian inference; RWD/RWE	Forecast refinement; regulatory strategy	Posterior distributions; narrowing bounds
Strategic Optionality	Real options; value of information	Go/no-go; partnering; financing	Option values; decision boundaries

Fat-Tailed Distributions and Portfolio Implications

The framework's emphasis on probabilistic methods is particularly important given the well-documented skewness of life-science returns. Empirical data consistently show that a small number of programs generate outsized returns while the majority [33]produce limited or negative value. Traditional valuation models that rely on expected values (means) systematically misrepresent this landscape. The "average" outcome for a development-stage program is a statistical artifact that rarely corresponds to any actually realized outcome.

A complexity-informed framework explicitly recognizes the heavy-tailed nature of life-science returns[34], allowing decision-makers to evaluate the probability of achieving specific return thresholds rather than relying on point estimates that blend fundamentally different outcome scenarios into a single number. For portfolio management, this perspective suggests that diversification strategies should be evaluated not only by their effect on average expected returns but by their impact on the probability of capturing tail outcomes.

Implementation Considerations and Governance

Transitioning from traditional linear models to a complexity-informed framework involves practical challenges that must be addressed for successful adoption. Three areas merit particular attention: data infrastructure, organizational capability, and model governance.

Data Infrastructure and AI Integration

The framework's reliance on probabilistic methods and Bayesian updating places significant demands on data infrastructure. Organizations need access to structured clinical trial data, real-

world claims and EHR data, competitive intelligence feeds, and capital markets information.[35] Modern AI and machine learning tools can assist in processing, structuring, and extracting signals from these heterogeneous data sources. Natural language processing can monitor regulatory and competitive developments; machine learning models can identify patterns in adoption data; and automated pipelines can feed updated parameters into the Bayesian framework. However, AI integration should be viewed as an augmentation of analytical capability rather than a replacement for expert judgment.

Organizational Capability and Cross-Functional Integration

A complexity-informed framework inherently requires collaboration across functional boundaries. Forecasting, valuation, and capital structure analysis cannot be siloed in finance or commercial analytics when the models explicitly depend on clinical probability assessments, regulatory strategy inputs, and market access assumptions.[36] Successful implementation requires governance structures that bring together clinical development, regulatory affairs, commercial strategy, finance, and business development in a shared analytical environment.

Model Governance and Transparency

The increased sophistication of complexity-informed models carries a risk: the potential for "black box" outputs that stakeholders cannot interrogate or trust. Effective model governance requires explicit documentation of assumptions, sensitivity analyses that reveal key value drivers, and transparent Bayesian audit trails[37] that show how and why assumptions have been updated. Scenario narratives - qualitative descriptions of the conditions under which specific outcomes are more or less likely - can bridge the gap between quantitative outputs and decision-maker intuition.

The Importance of Phased Implementation

While the full complexity-informed framework represents a substantial analytical advance, it is neither necessary nor advisable for organizations to attempt a wholesale transition from traditional methods. The shift from linear to complexity-informed valuation is best understood as an evolutionary process that builds institutional capability, stakeholder confidence, and analytical infrastructure incrementally. A phased approach reduces adoption risk, allows organizations to demonstrate early value, and creates the feedback loops necessary for the framework itself to improve over time.

Phased implementation is critical for several reasons. First, the analytical sophistication required increases substantially across the framework's four components. Monte Carlo simulation represents a comparatively modest extension of existing rNPV capabilities, whereas agent-based modeling and real options analysis require skills and governance structures that most pharmaceutical organizations do not yet possess.[35] Attempting to deploy all four components

simultaneously risks overwhelming both the analytics team and the decision-makers who must interpret the outputs.

Second, organizational trust in new methods must be earned through demonstrated accuracy. Decision-makers who have relied on deterministic scenarios for years will not immediately embrace probabilistic distributions, regardless of their theoretical superiority. By beginning with probabilistic extensions of familiar methods, organizations allow stakeholders to experience practical benefits before introducing more conceptually demanding components. Third, phased implementation creates natural checkpoints at which the organization can assess whether additional complexity is generating proportional decision value. Not every organization will need full deployment of all four components; the appropriate stopping point depends on portfolio complexity, capital constraints, and organizational readiness.

Table 7 presents a structured roadmap for organizations seeking to adopt the framework, with each phase building upon the capabilities established in the preceding phase.

Table 7: Phased Implementation Roadmap

Phase	Activities	Requirements	Expected Outcomes
Phase 1 (Months 1–6)	Replace point estimates with distributions; Monte Carlo valuation; cross-functional review	Analytics training; executive sponsorship; data audit	Probabilistic ranges; improved risk communication
Phase 2 (Months 6–12)	Agent-based adoption models; Bayesian pipelines; RWD integration	Data infrastructure; governance committee	Dynamic forecasts; early-warning dashboards
Phase 3 (Months 12–18)	Real options/VOI analysis; dynamic cap table; AI competitive intelligence	Decision science expertise; board literacy	Optionality management; optimized allocation

Phase 1: Foundation. The first phase focuses on replacing point estimates with probability distributions. This means augmenting existing rNPV models with Monte Carlo simulation layers that sample from distributions of clinical success probabilities, peak revenue assumptions, time-to-market, and pricing scenarios. This phase also establishes the cross-functional model review process essential for later phases. The organizational requirements are comparatively modest: analytics team training in probabilistic methods, executive sponsorship, and a data audit to identify existing information assets and gaps. When decision-makers begin receiving valuation ranges with explicit confidence intervals rather than single numbers, the quality of strategic dialogue improves immediately.

Phase 2: Integration. The second phase introduces agent-based adoption models that replace smooth penetration curves with network-mediated diffusion trajectories, and Bayesian updating pipelines that continuously refine assumptions as new clinical, commercial, and competitive data become available. This phase requires more substantial investment in data infrastructure,

including connections to real-world data sources and the establishment of a cross-functional governance committee. The expected outcomes include dynamic forecasts that adapt to evolving market conditions and early-warning dashboards that flag programs approaching inflection points. Phase 2 is where the framework begins generating insights that are qualitatively different from, rather than simply more precise than, traditional approaches.

Phase 3: Advanced. The third phase deploys the most analytically demanding components: real options and value-of-information analysis for strategic decision sequencing, dynamic capitalization table modeling, and AI-augmented competitive intelligence. This phase requires specialized decision science expertise and board-level literacy in probabilistic methods, without which the framework's outputs risk being misunderstood at the governance level. Organizations that reach Phase 3 will have the capability to manage strategic optionality as a core competency rather than an ad hoc response to external events.

It is worth emphasizing that the value of phased implementation is not merely logistical; it is epistemological. Each phase generates organizational learning that informs the design and calibration of the next. The probability distributions produced in Phase 1 become the priors for Bayesian updating in Phase 2. The dynamic forecasts generated in Phase 2 provide the volatility estimates and scenario structures needed for real options analysis in Phase 3. In this sense, the implementation roadmap itself embodies the complexity-informed principle of adaptive, evidence-driven decision-making.[2,14,36]

Discussion

For professionals involved in pharmaceutical management, forecasting, and strategic planning, the complexity-informed framework offers several practical benefits that extend beyond improved numerical precision.

- First, it shifts the focus from producing precise but fragile point estimates to understanding ranges of outcomes and the conditions under which those outcomes are more or less likely.[38] This reframing improves communication with investors, boards, and strategic partners.
- Second, risk is reconceptualized from a number added to a discount rate to a function of how development, market, and financing elements are structured and connected. This structural view of risk enables more nuanced discussions about how to mitigate, transfer, or accept specific uncertainties through clinical design, partnership structures, or financing instruments.
- Third, the framework encourages closer integration between clinical strategy, commercial planning, and financial decision-making. By making the interdependencies among these domains explicit and quantifiable, it provides a shared analytical platform

that can break down functional silos that frequently limit strategic insight in pharmaceutical organizations.

- Fourth, the incorporation of real options and value-of-information analysis provides a principled basis for sequencing decisions under uncertainty. Rather than committing prematurely to a single development path, organizations can identify which uncertainties are most consequential, which evidence would be most valuable in resolving them, and how to structure development programs to preserve strategic flexibility.

These benefits are particularly salient in the rare disease context. As demonstrated by Rosenblatt et al.,[8] the speed and shape of rare disease product uptake depend on complex interactions among efficacy, unmet need, company capabilities, and order of entry. In markets where patient populations are small and highly concentrated, the interaction effects among these variables are amplified, making the case for complexity-informed modeling even more compelling. Understanding the appropriate rate of adoption, as opposed to only the final peak share, has significant implications for commercialization, business development, and M&A decisions.

Conclusion

The life sciences represent one of the clearest examples of an industry where traditional linear models struggle to keep pace with reality. The complexity of drug development, grounded in the biological reality of adaptive systems, demands analytical tools that can accommodate feedback, emergence, nonlinearity, and genuine uncertainty. The framework proposed in this article does not seek to replace established methods but to extend them into a more adaptive, transparent, and strategically useful architecture.

By treating valuation and capital structure as components of a dynamic system, informed by agent-based stakeholder models, probabilistic distributions, Bayesian updating with real-world evidence, and decision analytics grounded in real options theory[2,14,39], pharmaceutical organizations can develop more flexible, transparent, and resilient decision frameworks. While no model can eliminate uncertainty, approaches that better reflect how value actually emerges over time can meaningfully improve strategic choices in an increasingly uncertain and competitive environment.

The authors propose that the pharmaceutical industry is well positioned to advance the practical application of these methods. The interdisciplinary expertise represented within the organization, spanning forecasting, analytics, market research, and strategic planning, provides a natural foundation for developing, validating, and refining complexity-informed decision tools that can measurably improve outcomes for patients, investors, and the broader life-science ecosystem. The proposed framework does not seek to replace established methods but to

extend them into a more adaptive, transparent, and strategically useful architecture. By integrating stakeholder behavior models, probabilistic valuation, Bayesian updating with real-world evidence, and decision analytics grounded in real options and value-of-information theory, organizations can improve capital allocation, transaction structure, and the sequencing of high-stakes decisions.

While no model eliminates uncertainty, a model that better reflects how value emerges over time can materially improve decision quality.

Next Steps

Future research should explore empirical validation of complexity-informed valuation models against conventional rNPV approaches using retrospective portfolio data, the development of standardized early-warning signal metrics applicable across therapeutic areas and development stages, and the integration of large language models and generative AI into the Bayesian updating pipeline. The authors intend to pursue this work in a disciplined sequence, with subsequent publications specifically addressing: (1) the conditions under which the proposed framework's complexity adds genuine strategic and predictive value over traditional rNPV with scenario analysis; (2) how the agent-based modeling components would be empirically validated, including requisite data requirements and their practical achievability; and (3) detailed case studies demonstrating application of the framework in both rare and non-rare disease contexts.

A legitimate concern is that introducing a complexity-informed AI decision framework may create confusion and decision paralysis within pharmaceutical organizations. However, this "complexity versus simplicity" critique focuses largely on point prediction accuracy, while the methods we propose — agent-based models, Monte Carlo simulation, Bayesian updating, and real options analysis — are fundamentally decision-support tools designed to characterize uncertainty and value managerial flexibility, not to generate single-number forecasts or valuations. Whether this critique applies hinges on the distinction between traditional forecasting and valuation grounded in decision analysis.

The strongest takeaway for pharmaceutical valuation is not to avoid complexity altogether but to keep valuation and forecast model inputs straightforward while utilizing appropriately complex decision frameworks. For example, rNPV or simple heuristics such as exponential smoothing remain suitable for baseline revenue projections, but real options thinking should be layered onto go/no-go decisions at phase gates, Monte Carlo applied for honest uncertainty communication to stakeholders, Bayesian updating employed for sequential trial design, and ABM leveraged for AI-assisted insight — domains where structured complexity is more valuable than forecast precision. The risk to guard against is not complexity per se, but false precision.

About the Authors

Samuel Renwick is the Founder and CEO of RNA Advisors, a strategic advisory firm specializing in life-science financing, capital structure optimization, and transaction strategy. Sam advises pharmaceutical and biotechnology companies on complex financing architectures, partnering decisions, and portfolio strategy, bringing a systems-oriented perspective to capital allocation under uncertainty. He has been worked with over 1,000 venture capital-backed biotechnology companies on finance and valuation related matters in his career. Prior to founding RNA, Sam ran the life science team at SVB Analytics. He holds multiple graduate degrees in psychology and business/finance. He was the recipient of the J. Fred Weston award from UCLA Anderson for excellence in finance.

Jerry A. Rosenblatt is the Founder and CEO of Rosenblatt Life Science Consultants. His professional career has been devoted to advancing Marketing Science, the application of scientific methods to the understanding of marketing and management challenges. He has been working with pharmaceutical companies for more than 30 years applying his expertise to assist clients with Forecasting & Opportunity Assessment, Business Valuation and the creation of innovative marketing education and learning & development programs. He is the former Head of the Global Forecasting & Opportunity Assessment consulting group of IMS Health (IQVIA). Jerry holds an MBA and PhD (Marketing Science) and is a recipient of the PMSA Lifetime Achievement Award.

---

[1] Holland JH. Hidden Order: How Adaptation Builds Complexity. Reading, MA: Addison-Wesley; 1995.
[2] Arthur WB. Foundations of complexity economics. Nat Rev Phys. 2021;3(2):136–145.
[3] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32(1):40–51.
[4] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016;47:20–33.
[5] Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–286.
[6] Stewart JJ, Allison PN, Johnson RS. Putting a price on biotechnology. Nat Biotechnol. 2001;19(9):813–817.
[7] Bogdan B, Villiger R. Valuation in Life Sciences: A Practical Guide. 3rd ed. Berlin: Springer; 2010.
[8] Rosenblatt JA, Ghenadenik A, Rosenberg R, Anand A. Using longitudinal claims data to predict response uptake in rare and ultra-rare diseases. J Pharm Manag Sci Assoc. 2025;Spring:5–14.
[9]Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9(3):203–214.
[10] Schuhmacher A, Hinder M, von Segundo F, Hartl D, Gassmann O. The pharmaceutical productivity gap – Incremental decline in R&D efficiency despite transient improvements. Drug Discov Today. 2024;29(10):104128.
[11]Grabowski H, Vernon J. Returns to R&D on new drug introductions in the 1980s. J Health Econ. 2000;19(6):931–955.
[12]Villiger R, Bogdan B. Getting real about valuations in biotech. Nat Biotechnol. 2005;23(4):423–428.
[13] Iyengar R, Van den Bulte C, Valente TW. Opinion leadership and social contagion in new product diffusion. Mark Sci. 2011;30(2):195–212.

[14] Mitchell M. Complexity: A Guided Tour. Oxford: Oxford University Press; 2009.
[15] Kitano H. Systems biology: A brief overview. Science. 2002;295(5560):1662–1664.
[16] Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–113.
[17] Kauffman S. The Origins of Order: Self-Organization and Selection in Evolution. Oxford: Oxford University Press; 1993.
[18] Hartmann M, Hassan A. Application of real options analysis for pharmaceutical R&D project valuation. Res Policy. 2006;35(3):343–354.
[19] Fernandez JM, Stein RM, Lo AW. Commercializing biomedical research through securitization techniques. Nat Biotechnol. 2012;30(10):964–975.
[20] Scheffer M. Critical Transitions in Nature and Society. Princeton: Princeton University Press; 2009.
[21] Scheffer M, Bascompte J, Brock WA, et al. Early-warning signals for critical transitions. Nature. 2009;461(7260):53–59.
[22] Metrick A, Yasuda A. Venture Capital and the Finance of Innovation. 3rd ed. New York: Wiley; 2021.
[23] Metrick A, Yasuda A. Venture Capital and the Finance of Innovation. 3rd ed. New York: Wiley; 2021.
[24] Kaplan SN, Strömberg P. Financial contracting theory meets the real world. Rev Econ Stud. 2003;70(2):281–315.
[25] Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. Proc Natl Acad Sci USA. 2002;99(suppl 3):7280–7287.
[26] Axtell RL, Farmer JD. Agent-based modeling in economics and finance: past, present, and future. J Econ Lit. 2022;60(4):1–76. Axtell, Robert L., and J. Doyne Farmer. 2025. "Agent-Based Modeling in Economics and Finance: Past, Present, and Future." Journal of Economic Literature 63 (1): 197–287.
[27] Glasserman P. Monte Carlo Methods in Financial Engineering. New York: Springer; 2003.
[28] Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence: What is it and what can it tell us? N Engl J Med. 2016;375(23):2293–2297.
[29] U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. Silver Spring, MD: FDA; 2018.
[30] Trigeorgis L. Real Options: Managerial Flexibility and Strategy in Resource Allocation. Cambridge, MA: MIT Press; 1996.
[31] Lee W, Wong WB, Kowal S, Garrison LP, Veenstra DL, Li M. Modeling the ex ante real option value in an innovative therapeutic area: ALK-positive non-small cell lung cancer. Pharmacoeconomics. 2022;40:623–631.
[32] Claxton K. The irrelevance of inference: A decision-making approach to the stochastic evaluation of health care technologies. J Health Econ. 1999;18(3):341–364.
[33] Tufts Center for the Study of Drug Development. Outlook 2023. Boston, MA: Tufts CSDD; 2023. Tufts Center for the Study of Drug Development, 2023. Clinical Sites Are Optimistic Despite Growing Challenges. Clinical Leader Report, Jan 17, 2023
[34] Taleb NN. The Black Swan: The Impact of the Highly Improbable. New York: Random House; 2007.
[35] Schneeweiss S. Learning from big health care data. N Engl J Med. 2014;370(23):2161–2163.
[36]Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191–200.
[37] Varian HR. Big data: New tricks for econometrics. J Econ Perspect. 2014;28(2):3–28.
[38]Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011;10(6):428–438.
[39] Arthur WB. Complexity and the economy. Science. 1999;284(5411):107–109.

Decoding the Hype of TrumpRx with DTC Distribution: Is Life Science's Commercial Model Ready for the Change?

Anakh Sawhney – Bernards High School, Bernardsville, New Jersey, United States
Devesh Verma, PhD – Principal, Axtria Inc., Berkeley Heights, New Jersey, United States

Abstract: The pharmaceutical industry is undergoing a fundamental shift in distribution strategy as direct-to-consumer (DTC) platforms proliferate and federal initiatives like TrumpRx reshape market dynamics. This research examines whether these emerging channels have increased drug access for patients or created new forms of inequity, which therapeutic areas demonstrate viability for DTC models beyond the GLP-1 category that catalyzed this transformation, and whether DTC distribution genuinely benefits pharmaceutical commercial operations. Employing a mixed-methods approach that combines qualitative policy analysis with quantitative techniques including expert panel assessment of therapeutic area characteristics and market expansion modeling, this study analyzes publicly available data from the Centers for Medicare and Medicaid Services, Kaiser Family Foundation surveys, and manufacturer financial disclosures. The analysis examines five therapeutic areas with active TrumpRx participation: obesity/GLP-1, type 2 diabetes, rheumatoid arthritis, migraine, and hepatitis C. The findings reveal that access benefits distribute asymmetrically across patient populations: approximately 88 million Americans experience high benefit through savings exceeding $500 monthly, 32 million Medicare Standard Part D beneficiaries experience moderate benefit through savings of $125 monthly, while 210 million experience neutral impact because they already have optimal coverage, and 13 million Medicare Low Income Subsidy beneficiaries face negative impact from TrumpRx pricing that exceeds their subsidized copays. The analysis identifies two critical paradoxes: racial and ethnic minorities who would benefit most from DTC pricing face the greatest digital access barriers, and rural communities with the highest need for home delivery models have the lowest broadband infrastructure to support them. Therapeutic area viability proves highly concentrated, with obesity and weight management scoring 90.8 out of 100 on a composite viability index derived from expert panel assessment, while oncology scores 38.5. Two therapeutic areas achieve high viability status, five fall in the medium tier, and three demonstrate low viability. DTC platforms generate strategic value primarily through market expansion rather than margin improvement, with an estimated addressable market of 38.9 million patients for GLP-1 therapies alone, where even a conservative 2 percent capture rate would generate billions in incremental annual revenue. Additional strategic value accrues through adherence value capture and patient lifetime value extension enabled by DTC platform support services.

Keywords: direct-to-consumer, pharmaceutical distribution, TrumpRx, GLP-1, market access, health equity, commercial strategy

INTRODUCTION

The pharmaceutical industry's distribution landscape is experiencing its most significant disruption since the emergence of pharmacy benefit managers in the 1980s. Driven by the unprecedented commercial success of glucagon-like peptide-1 (GLP-1) receptor agonists for diabetes and obesity, manufacturers have increasingly explored direct relationships with patients that bypass traditional intermediaries. Eli Lilly's LillyDirect platform, launched in January 2024, represented a watershed moment in this evolution, offering tirzepatide directly to consumers at prices substantially below list. Novo Nordisk, Pfizer, and other major manufacturers have followed with their own initiatives, while the federal government's TrumpRx marketplace, launched in February 2026, promises to institutionalize this channel shift at national scale.[1,2]

The strategic implications of this transformation extend far beyond the GLP-1 category that catalyzed it. Pharmaceutical commercial leaders must now evaluate whether DTC distribution genuinely enhances profitability or merely captures patients who would have accessed therapy through traditional channels. They must determine which therapeutic areas beyond obesity and diabetes demonstrate genuine viability for direct models, and they must understand how these emerging channels affect different patient populations across the insurance landscape. These questions carry particular urgency as the TrumpRx platform has launched with participation commitments from sixteen major pharmaceutical manufacturers, including Pfizer, AstraZeneca, Merck, Novartis, Sanofi, and Gilead.[3]

This research addresses three fundamental questions that pharmaceutical commercial strategists must answer as they navigate this transition. First, has DTC distribution increased drug access for patients, or does it create new forms of inequity that offset its apparent benefits? Second, is DTC viable across therapeutic areas beyond the GLP-1 drugs that drove its emergence, and if so, which characteristics predict success? Third, does DTC distribution genuinely benefit pharmaceutical sales and profitability, or does it cannibalize existing channels while compressing margins?

BACKGROUND

The Emergence of Pharmaceutical DTC Distribution

The traditional pharmaceutical distribution model has remained largely unchanged for decades. Manufacturers sell to wholesalers, who distribute to pharmacies, which dispense to patients whose costs are mediated by pharmacy benefit managers negotiating on behalf of insurers. This multilayered system produces the well-documented gross-to-net phenomenon, where list prices bear little relationship to the net prices manufacturers ultimately realize after rebates, fees, and discounts that can exceed 50 percent for brand pharmaceuticals.[4]

The GLP-1 class disrupted this equilibrium through a combination of extraordinary clinical efficacy and unprecedented demand that outstripped supply. When Eli Lilly launched Zepbound for obesity in late 2023, the company faced a strategic choice: rely on traditional channels where pharmacy benefit manager coverage decisions and formulary negotiations would determine access or establish direct relationships with the substantial population willing to pay out-of-pocket for a therapy their insurance would not cover. LillyDirect represented the latter choice, offering Zepbound at $550 per month through affiliated telehealth

providers, a substantial reduction from the $1,060 list price, though still requiring patients to pay entirely out-of-pocket.[5]

The platform's structure illustrates the DTC value proposition. By connecting patients directly with independent telehealth providers for prescriptions and partnering with home delivery pharmacies, LillyDirect captures patients who otherwise might be excluded from therapy entirely. The company does not disclose platform-specific revenue, but the strategic logic is clear: there is potential value added due to direct patient engagement, these patients represent incremental volume that traditional channels cannot capture, and the company has the potential to retain a higher percentage of each dollar than it would after gross-to-net reductions in the traditional channel.[6]

The TrumpRx Federal Marketplace

The TrumpRx initiative emerged from Executive Order 14297, signed in May 2025, directing the Department of Health and Human Services to establish most-favored-nation pricing for pharmaceutical products sold in the United States. The executive order leveraged the threat of pharmaceutical tariffs to encourage manufacturer participation, and by late 2025, sixteen major companies had announced agreements that included commitments to offer products through the TrumpRx.gov marketplace at substantial discounts.[7]

The platform operates as a federal portal connecting patients directly to manufacturer DTC programs. Unlike Medicare's negotiated drug prices under the Inflation Reduction Act, which affect only specific high-spend Part D drugs, TrumpRx theoretically extends discounted access to any product manufacturers choose to include. At launch in February 2026, the platform listed 43 brand-name medications with discounts ranging from 33 percent to 93 percent off list prices. GLP-1 medications feature prominently: Eli Lilly's Zepbound is available at $299 monthly (down from $1,087 list), Novo Nordisk's Wegovy pen at $199 monthly (down from $1,349 list), and the oral Wegovy pill at $149 monthly. Pfizer's Xeljanz for rheumatoid arthritis is offered at $1,518 (down from $3,204), migraine treatment Zavzpret is available at $595 (down from $1,190), and Gilead's hepatitis C cure Epclusa is offered at $2,425 for the complete treatment course (down from $24,920 list price, representing a 90 percent discount).[8]

The commercial implications remain uncertain. Experts note that TrumpRx pricing may not benefit most Americans, who would pay less through existing insurance arrangements. As one analysis observed, insured patients using traditional pharmacy benefits often realize effective discounts of 50 to 55 percent after rebates, meaning TrumpRx's discounts may actually cost patients more than their current copayments in some cases.[9] This observation foreshadows the segmented access effects that emerge prominently in the present analysis.

METHODOLOGY

This study employs a mixed-methods approach combining qualitative policy analysis with quantitative modeling of market factors and patient segment characteristics. The research design addresses each of the three research questions through distinct but complementary analytical frameworks.

Data Sources

The analysis draws on publicly available data from multiple authoritative sources. Insurance coverage rates and affordability metrics derive from Kaiser Family Foundation surveys and reports, which provide nationally representative estimates of coverage by payer type and cost-related medication non-adherence across

demographic groups.[10] Disease prevalence data originate from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System, which provides state-level estimates of chronic condition rates.[11] Digital access and broadband availability data come from the American Community Survey, Federal Communications Commission broadband mapping, and reports from the Federal Reserve Bank of Atlanta on telehealth access disparities.[12,13]

Manufacturer financial data derive from quarterly earnings releases and annual reports filed with the Securities and Exchange Commission. Pricing information for DTC platforms and TrumpRx commitments comes from the TrumpRx.gov website (accessed February 2026), company announcements, White House communications, and industry analyses. Gross-to-net dynamics are estimated using Drug Channels Institute data and published financial disclosures indicating that brand pharmaceutical net prices typically reflect 45 to 55 percent reductions from list.[14]

Analytical Framework

The first research question, whether DTC has increased drug access for patients, combines patient segment analysis with examination of digital access disparities by race, ethnicity, and geography. The analysis constructs a benefit matrix that cross-tabulates patient insurance segments against five therapeutic areas with TrumpRx participation: obesity/GLP-1, type 2 diabetes, rheumatoid arthritis, migraine, and hepatitis C. For each segment-therapy combination, dollar savings are calculated by comparing pre-DTC costs (list prices for uninsured, typical copays for insured segments) against TrumpRx pricing. Segments are then classified into four benefit tiers: high benefit (savings exceeding $500 monthly), moderate benefit ($100 to $500 monthly savings), neutral (no material change), or negative benefit (cost increase from switching to TrumpRx). This tiered classification enables systematic comparison of how DTC distribution affects different patient populations across therapeutic areas.

The second research question, therapeutic area viability, employs expert panel assessment to classify ten therapeutic categories into viability tiers based on six dimensions scored on a 0 to 100 scale: coverage gap, patient autonomy in treatment decisions, therapy chronicity, affordability sensitivity, market size, and current access barriers. Five industry experts with pharmaceutical commercial strategy experience independently scored each therapeutic area across all dimensions, and the average scores were used to calculate composite viability indices. Therapeutic areas were then classified into three tiers: high viability (composite scores 75 to 100), medium viability (50 to 75), and low viability (below 50). This classification is validated against observed TrumpRx participation patterns to assess predictive accuracy.

The third research question, whether DTC benefits pharmaceutical profitability, is addressed through a market expansion model that quantifies the patient population unable to access therapy through traditional channels. This addressable market is calculated as the product of segment population, disease prevalence, and clinical eligibility rates. Rather than making multiple assumptions about conversion rates across insurance types, the analysis presents the total addressable market and demonstrates that even conservative capture rates yield substantial incremental revenue. Strategic value is further quantified through two additional components: adherence value capture, which estimates revenue retention from improved medication persistence enabled by DTC platform support services, and patient lifetime value extension from increased treatment duration.

FINDINGS

Research Question 1: Has DTC Increased Drug Access for the Patients?

The most significant finding of this research concerns the highly asymmetric distribution of benefits from DTC distribution across patient populations. Far from democratizing access uniformly, DTC platforms create distinct winners and losers whose outcomes diverge based on insurance status, geographic location, and digital access capability.

DTC/TrumpRx Benefit Matrix by Patient Segment and Therapeutic Area

To analyze benefit distribution across patient populations, this study constructs a benefit matrix that cross-tabulates seven patient segments against five therapeutic areas with active TrumpRx participation: obesity/GLP-1 (Zepbound at $299, Wegovy at $199), type 2 diabetes (Ozempic at $199), rheumatoid arthritis (Xeljanz at $1,518), migraine (Zavzpret at $595), and hepatitis C (Epclusa at $2,425 for a curative 12-week course, down from $24,920 list price). The methodology calculates dollar savings for each segment-therapy combination by comparing pre-DTC monthly costs against current TrumpRx pricing. For uninsured patients, pre-DTC cost equals the manufacturer list price. For insured patients, pre-DTC cost reflects typical copay structures based on drug tier placement and plan design. Segments are classified as HIGH benefit (greater than $500 monthly savings), MODERATE ($100 to $500), NEUTRAL (zero change), and NEGATIVE (cost increase).

The benefit matrix reveals consistent patterns across therapeutic areas while highlighting important segment-specific and therapy-specific variations. Uninsured patients experience HIGH benefit across all five therapeutic areas, with monthly savings ranging from $595 for migraine (Zavzpret: $1,190 list to $595 TrumpRx) to $22,495 for the hepatitis C cure (Epclusa: $24,920 list to $2,425 TrumpRx). Similarly, commercially denied and Medicaid non-covering populations consistently achieve HIGH benefit across all therapeutic areas, as these segments face list-price exposure without DTC alternatives. In total, 88 million Americans across these three segments experience high benefit regardless of therapeutic area.

Medicare Standard Part D beneficiaries present a more nuanced picture that varies by therapeutic area. For obesity/GLP-1, type 2 diabetes, and migraine therapies, these 32 million beneficiaries experience MODERATE benefit, saving approximately $125 monthly as TrumpRx pricing combined with the $2,000 annual out-of-pocket cap reduces their effective monthly cost. For hepatitis C, Medicare Part D beneficiaries experience HIGH benefit because Epclusa's TrumpRx price of $2,425 is substantially below the $5,000 or more they would pay out-of-pocket before reaching catastrophic coverage. However, for rheumatoid arthritis (Xeljanz), Medicare Part D beneficiaries experience NEUTRAL impact because TrumpRx's $1,518 monthly price does not meaningfully improve upon their existing specialty tier coverage.

Patient Segment	Population (M)	Obesity/GLP-1	Type 2 Diabetes	RA (Xeljanz)	Migraine	Hepatitis C
Uninsured	28	HIGH	HIGH	HIGH	HIGH	HIGH
Commercial - Denied	20	HIGH	HIGH	HIGH	HIGH	HIGH
Medicaid - Non-Covering	40	HIGH	HIGH	HIGH	HIGH	HIGH
Medicare Standard Part D	32	MODERATE	MODERATE	NEUTRAL	MODERATE	HIGH
Medicare LIS	13	NEGATIVE	NEGATIVE	NEUTRAL	NEUTRAL	NEUTRAL
Commercial - Covered	160	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL
Medicaid - Covering	50	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL

Figure 1. *DTC/TrumpRx Benefit Matrix by Patient Segment and Therapeutic Area*

The 210 million Americans with commercial coverage or Medicaid in covering states experience NEUTRAL impact across all five therapeutic areas, as their existing insurance already provides optimal access with copays typically ranging from $0 to $50 monthly. Most significantly, 13 million Medicare Low Income Subsidy beneficiaries face NEGATIVE outcomes for obesity/GLP-1 and diabetes therapies, where TrumpRx's effective $50 monthly cost exceeds their subsidized copays of $5 to $12 monthly. For rheumatoid arthritis, migraine, and hepatitis C, LIS beneficiaries experience NEUTRAL impact as their subsidized access remains superior to or equivalent to TrumpRx pricing. Figure 1 illustrates the complete benefit matrix across all five therapeutic areas.

Racial and Ethnic Disparities in DTC Platform Accessibility

The analysis identifies two paradoxes that further complicate the access narrative. The first concerns racial and ethnic disparities in digital access. Hispanic Americans experience uninsured rates 11.4 percentage points higher than White Americans (17.9 percent versus 6.5 percent), theoretically positioning them to benefit substantially from DTC pricing. However, Hispanic households are 15 percentage points less likely than White households to have broadband internet access at home (65 percent versus 80 percent), creating a fundamental barrier to DTC platform utilization. Black households face a 9 percentage point broadband gap, while Native American households experience a 20 percentage point deficit.[17] Since 63 percent of the uninsured population consists of people of color, DTC platforms face a structural challenge: the populations most likely to benefit from their pricing cannot access their digital infrastructure. Figures 2 and 3 illustrate these disparities.

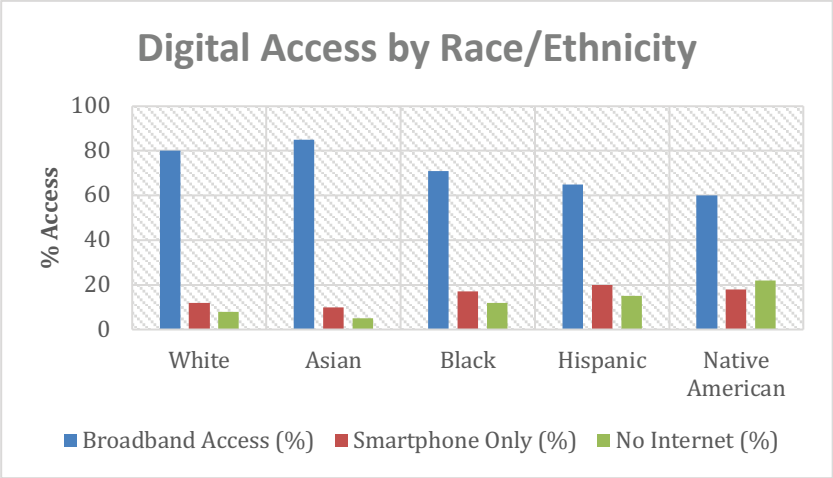


Figure 2. *Racial and Ethnic Disparities – Digital Access*

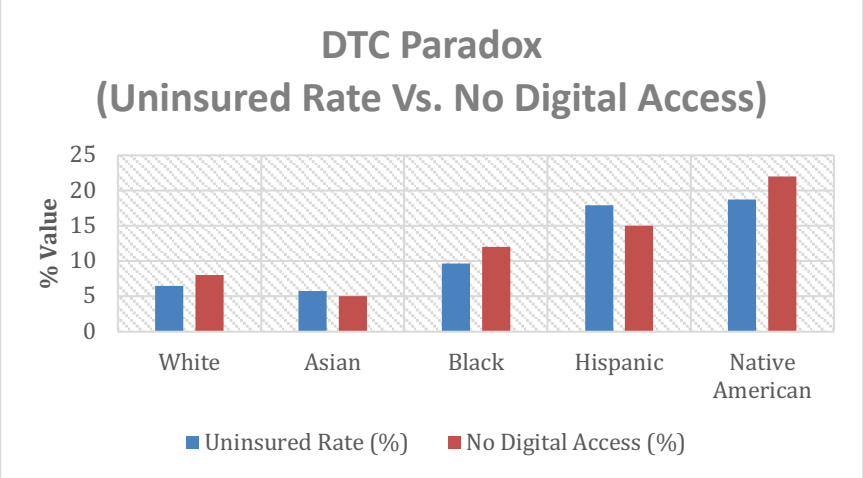


Figure 3. *DTC Paradox – Uninsured Vs. No Digital Access*

Rural vs. Urban Access Metrics: The Dual Disparity Challenge

The second paradox concerns rural and urban disparities. Rural Americans face substantial pharmacy access barriers: only 75 percent live within 10 miles of a pharmacy compared to 98 percent of urban residents, positioning them to benefit significantly from DTC home delivery models. However, rural households are 15 percentage points less likely to have broadband access (72 percent versus 87 percent in urban areas), and rural residents are 42 percent less likely to utilize telehealth services even when available.[18] Meanwhile, 80 percent of federally designated high-needs health professional shortage areas are rural.[19] This creates what researchers have termed as dual disparity: rural populations face barriers in both physical access to healthcare and digital access to telehealth alternatives, and DTC platforms address only the former while depending entirely on the latter. Figure 4 summarizes these geographic disparities.

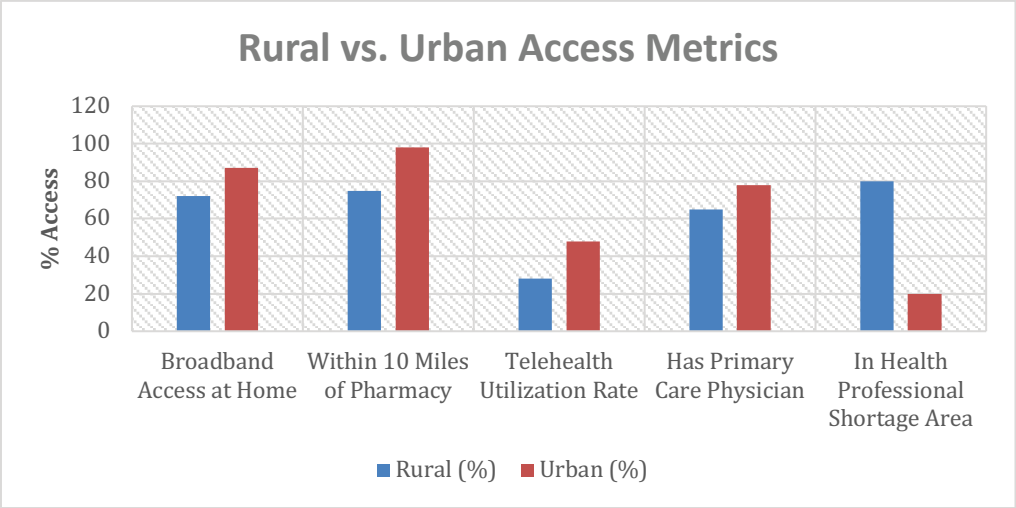
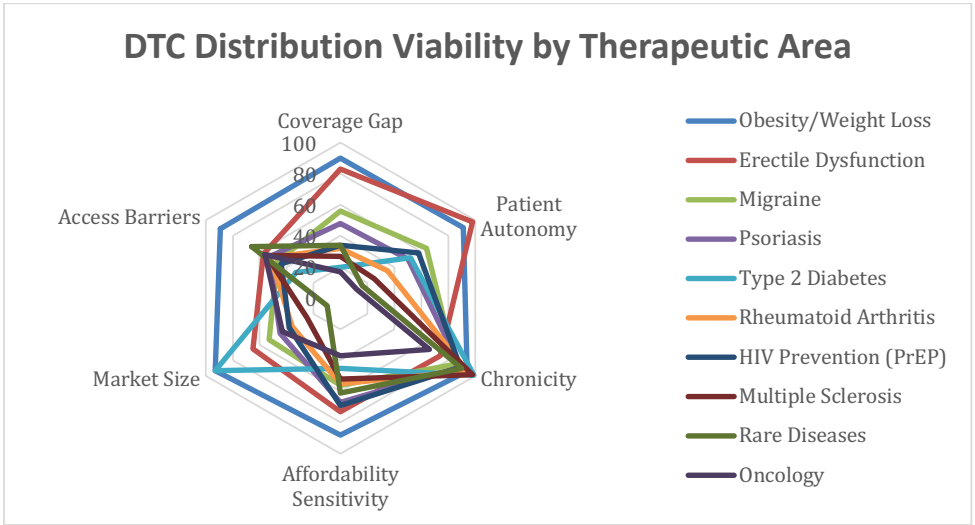


Figure 4. *Rural vs. Urban Access Metrics: The Dual Disparity Challenge*

These findings suggest that DTC distribution alone cannot achieve equitable access improvement without complementary investments in digital infrastructure. The populations with the greatest pharmaceutical access needs are systematically excluded from digital platforms by the same socioeconomic factors that limit their healthcare access more broadly.

Research Question 2: Is DTC Viable Across Therapeutic Areas?

The expert panel assessment reveals that DTC viability is highly concentrated rather than uniformly distributed across therapeutic areas. Five industry experts with pharmaceutical commercial strategy experience independently scored ten therapeutic categories across six dimensions, and the averaged results classified therapeutic areas into three distinct viability tiers. Figure 5 presents a radar chart visualization comparing all ten therapeutic areas across the six scoring dimensions.



Note: Scores represent expert panel assessment integrating multiple data sources, including KFF surveys, CMS coverage databases, CDC prevalence data, and published prior authorization studies. Higher scores indicate greater DTC viability potential.

Figure 5. *Therapeutic Area Dimension Scores Based on DTC Viability*

The high viability tier (composite scores 75 to 100) contains two therapeutic areas. Obesity and weight management scored highest at 90.8, demonstrating exceptional characteristics across all six dimensions with scores ranging from 88 (affordability sensitivity) to 94 (chronicity). This reflects that only 18-19 percent of large employers cover GLP-1 medications for weight loss, Medicare explicitly excludes obesity drugs, and over 100 million Americans have clinical obesity creating substantial market size. Erectile dysfunction achieved a composite score of 75.0, driven by the highest patient autonomy of any therapeutic area (98 percent) reflecting the highly patient-initiated nature of treatment seeking, combined with significant coverage gaps (83 percent) due to explicit Medicare exclusion since 2006 and Medicaid coverage bans since 2005.

The medium viability tier (composite scores 50 to 75) comprises five therapeutic areas: migraine (60.2), psoriasis (58.7), type 2 diabetes (57.0), HIV prevention (55.3), and rheumatoid arthritis (50.7). These areas share moderate coverage gaps, reasonable patient autonomy, and chronic treatment requirements, but lack the characteristics that position the high-tier areas for DTC success. Type 2 diabetes notably benefits from near-universal insurance coverage (only 20 percent coverage gap) despite having the highest chronicity score of any therapeutic area (99 percent), while rheumatoid arthritis sits at the lower boundary of medium viability due to physician-driven treatment decisions (35 percent patient autonomy) offset by high chronicity (94 percent).

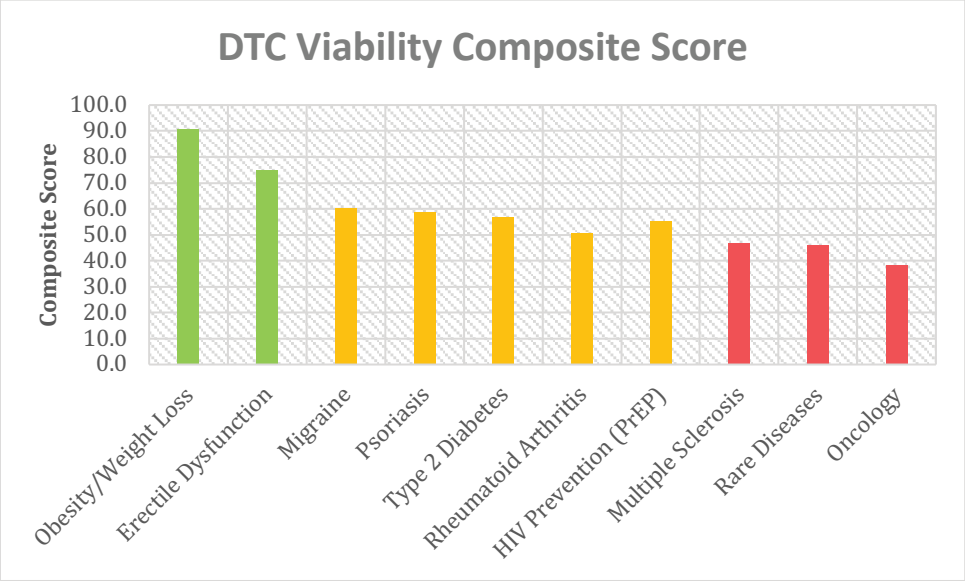


Figure 6. *DTC Viability by Therapeutic Area*

The low viability tier (composite scores below 50) comprises multiple sclerosis (47.0), rare diseases (46.2), and oncology (38.5). These areas share characteristics that fundamentally limit DTC potential: highly physician-driven treatment decisions (patient autonomy scores of 25, 16, and 12 percent respectively), smaller addressable markets, and treatment protocols requiring close clinical supervision. Multiple sclerosis is characterized by complex treatment selection requiring neurologist expertise and 90 percent prior authorization requirements. Rare diseases have the smallest market size (10 percent) and lowest patient autonomy (16 percent) reflecting specialized diagnosis and treatment. Oncology scored lowest overall due to minimal patient autonomy (12 percent), treatment complexity, and protocols incompatible with DTC convenience models. Figure 6 above shows the therapeutic viability chart based on the composite scores.

Validation against observed TrumpRx participation patterns supports the framework's predictive validity. Among the sixteen participating manufacturers, products cluster predominantly in high and medium viability areas. GLP-1 medications from Eli Lilly and Novo Nordisk represent the highest-profile high-tier offerings, while migraine treatments (Zavzpret), diabetes medications (Ozempic, Januvia), and rheumatoid arthritis therapies (Xeljanz) represent medium-tier participation. Notably, manufacturer participation in low-tier therapeutic areas remains limited, with oncology and rare disease products largely absent from TrumpRx despite several participating manufacturers having portfolios in these categories. This pattern suggests manufacturers are implicitly recognizing the same viability factors identified by the expert panel in this analysis.

Research Question 3: Does DTC Distribution Benefit Pharmaceutical Profitability?

The analysis reveals that DTC distribution generates strategic value for pharmaceutical manufacturers primarily through market expansion rather than margin improvement. The strategic value of DTC lies in capturing patients who face access barriers to therapy through traditional channels.

Consider GLP-1 therapies as an example. The analysis identifies four primary segments representing the addressable market: the 28 million uninsured Americans, of whom 7.9 million meet clinical eligibility criteria for GLP-1 treatment based on obesity prevalence; the estimated 20 million commercially insured patients whose plans explicitly exclude obesity medications, representing 5.6 million clinically eligible individuals; the 40 million Medicaid beneficiaries in non-covering states, representing 13.4 million eligible patients; and the 45 million Medicare beneficiaries subject to statutory exclusion of obesity medications, representing 12.1 million eligible individuals.

In total, these segments comprise an addressable market of 38.9 million clinically eligible patients who face barriers to accessing GLP-1 therapy through their current insurance arrangements. This population represents patients who would generate zero revenue through traditional distribution channels. Even a conservative 2 percent capture rate would yield 778,000 new patients. At TrumpRx pricing of $299 monthly for Zepbound, this represents $2.8 billion in annual revenue; at $199 monthly for Wegovy or Ozempic, this represents $1.9 billion in annual revenue. Figure 7 illustrates the addressable market by patient segment.

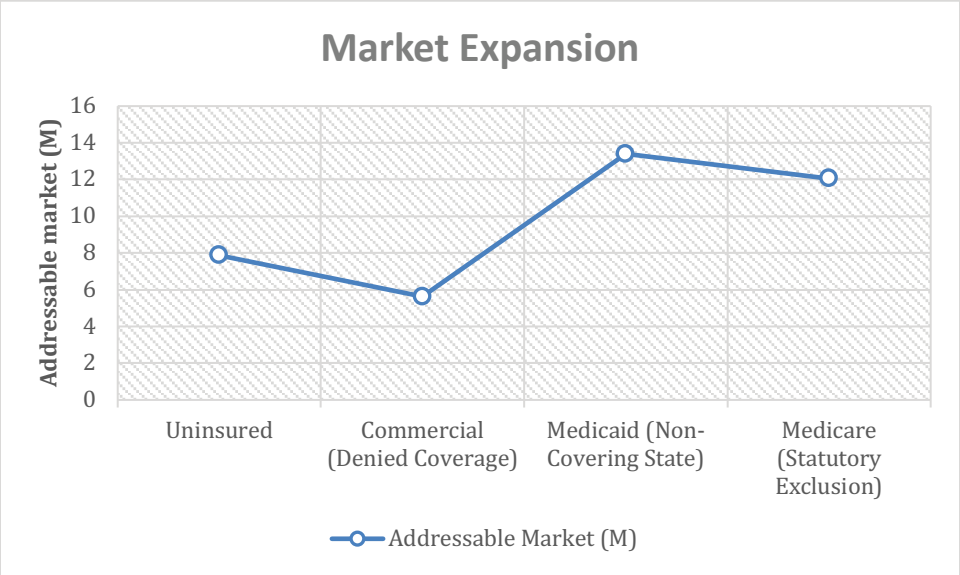


Figure 7. *DTC Market Expansion Opportunity (GLP-1 Analysis)*

Beyond direct revenue capture from market expansion, DTC platforms enable strategic value by improving adherence and extending patient lifetime value. Medication non-adherence costs the United States healthcare system between $100 billion and $528 billion annually, with approximately 50 percent of chronic disease patients failing to take medications as prescribed.[15] Non-adherent patients experience hospital readmission rates of 20 percent compared to 9 percent for adherent patients, and the annual per-patient cost of non-adherence ranges from $949 to $44,190 depending on condition.[16] DTC platforms that incorporate patient support services, including refill reminders, dosing guidance, and direct clinician access, can capture a portion of this value through improved persistence. A conservative 25 percent improvement in adherence among 100,000 DTC patients at a typical monthly prescription cost of $119 would yield $35.7 million in annual revenue retention that would otherwise be lost to discontinuation.

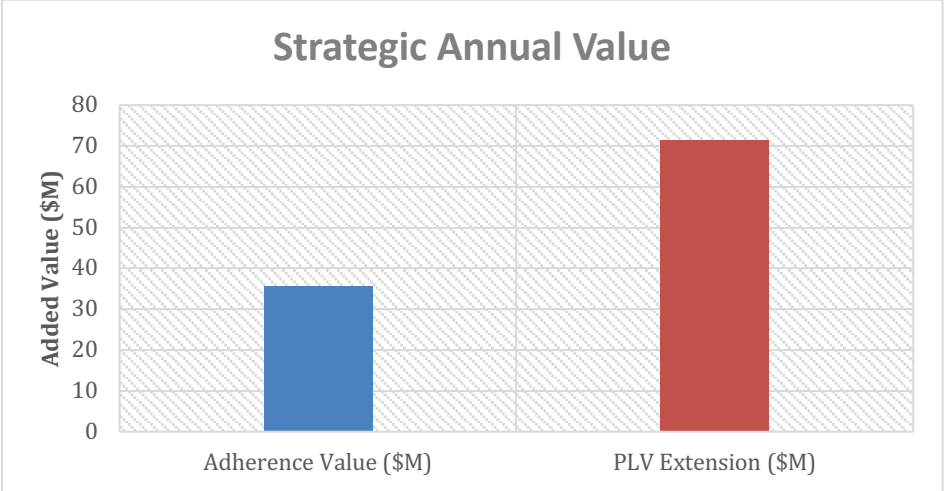


Figure 8. *DTC Strategic Annual Value: Increased Adherence & PLV Extension*

Similarly, DTC platforms may extend Patient Lifetime Value (PLV) by improving persistence with therapy. If the average patient persists for 12 months in the traditional model but for 18 months in the DTC model with support services, an additional 6 months of therapy at $119 per month yields $714 in additional lifetime value per patient. Across 100,000 patients, this represents $71.4 million in incremental lifetime value, a strategic asset that compounds as DTC enrollment grows. Figure 8 illustrates these additional strategic value components.

The significance of these findings is important for DTC platform viability. DTC platforms do not need to compete with traditional channels for existing patients. Their value derives from capturing entirely incremental volume from patients who would otherwise remain untreated, combined with improved retention of those patients through enhanced support services. The market expansion logic and incremental benefits from adherence and lifetime value are the core reasons for a manufacturer to invest in the DTC program despite compressed margins relative to traditional channels.

DISCUSSION

Implications for Pharmaceutical Commercial Strategy

The findings of this research offer several actionable insights for pharmaceutical commercial leaders evaluating DTC distribution strategies. First, patient segment economics should inform pricing and positioning decisions. DTC platforms should not be positioned as alternatives to insurance coverage for patients with good existing access; these patients typically achieve better economics through traditional channels. Rather, DTC should be positioned as an access solution for specific underserved segments, with marketing and pricing designed accordingly.

Second, therapeutic area selection for DTC investment should follow the viability framework dimensions identified by the expert panel in this analysis. Coverage gap, patient autonomy, chronicity, affordability sensitivity, market size, and access barriers collectively predict DTC potential far more accurately than extrapolation from GLP-1 success. Commercial leaders should resist the temptation to pursue DTC strategies in low-viability therapeutic areas simply because the channel has proven successful elsewhere; the characteristics that make obesity and erectile dysfunction suited to DTC are not replicated in oncology, rare diseases, or multiple sclerosis.

Third, DTC investment should be justified based on market expansion rather than margin improvement. Commercial teams should model the addressable population unable to access therapy through traditional channels and recognize that even modest capture of this market yields substantial incremental revenue. The 38.9 million addressable market for GLP-1 therapies demonstrates that DTC platforms can generate billions in annual revenue from patients who would otherwise generate none. Additionally, the strategic value framework presented here, combining market expansion with adherence value capture and patient lifetime value extension, provides a more complete picture of DTC economics than revenue analyses focused solely on direct sales.

Implications for Health Policy

The access analysis carries significant implications for policymakers evaluating federal DTC initiatives like TrumpRx. The finding that Low Income Subsidy Medicare beneficiaries may experience net harm from TrumpRx pricing suggests need for program design modifications that protect vulnerable populations from unintended consequences. More broadly, the digital access paradoxes identified in this research indicate that DTC

distribution cannot substitute for investments in broadband infrastructure and digital literacy programs that address fundamental barriers to platform utilization.

Policymakers should also consider the asymmetric benefit distribution when evaluating DTC programs as access solutions. While 88 million Americans experience high benefit and 32 million experience moderate benefit, over 200 million experience no improvement, and 13 million Medicare LIS beneficiaries experience harm. DTC represents a valuable complement to, rather than replacement for, comprehensive strategies addressing pharmaceutical affordability and access.

LIMITATIONS

This research is subject to several limitations that should inform interpretation of findings. First, pharmaceutical manufacturers do not disclose DTC platform-specific revenue. Second, the therapeutic area viability scoring employs an expert panel assessment framework where five industry experts independently scored each therapeutic area across six dimensions. While this approach provides practitioner-grounded insights and the averaged scores reduce individual bias, the integration of expert judgment into composite scores involves subjectivity that could influence tier classifications. The framework is designed to enable systematic comparison across therapeutic areas while acknowledging the inherent complexity of measuring multidimensional constructs like "patient autonomy" or "access barriers." Third, digital access data may not fully capture variations in digital literacy and willingness to engage with online healthcare platforms independent of infrastructure availability. Fourth, TrumpRx had just launched at the time of this analysis, meaning participation patterns and pricing commitments may evolve as the program becomes fully operational. And finally, the analysis doesn't evaluate any game-theoretic approach that PBMs may deploy once they understand the new DTC pricing reality to continue to establish their value in the pharmaceutical landscape.

CONCLUSION

This research provides evidence that DTC pharmaceutical distribution creates genuine strategic value for manufacturers and meaningful access improvements for specific patient populations, while simultaneously revealing limitations that should temper expectations for universal transformation. The analysis of five therapeutic areas with TrumpRx participation—obesity/GLP-1, type 2 diabetes, rheumatoid arthritis, migraine, and hepatitis C—reveals that access benefits distribute asymmetrically across four tiers: approximately 88 million Americans experience high benefit through savings exceeding $500 monthly (with hepatitis C patients saving up to $22,495 on curative therapy), 32 million Medicare Standard Part D beneficiaries experience moderate benefit through savings of $125 monthly, 210 million experience neutral impact because their existing coverage already provides optimal access, and 13 million Medicare Low Income Subsidy beneficiaries face negative impact from TrumpRx pricing that exceeds their subsidized copays. Therapeutic area viability is highly concentrated based on expert panel assessment: only two therapeutic areas (obesity and erectile dysfunction) achieve high viability status, five fall in the medium tier, and three demonstrate low viability for DTC distribution. DTC value derives from market expansion, capturing the 38.9 million Americans unable to access GLP-1 therapies through traditional channels, where even a conservative 2 percent capture rate

would generate billions in incremental annual revenue. Beyond market expansion, DTC platforms generate additional strategic value through adherence improvements and patient lifetime value extension, with potential revenue retention of $35.7 million annually from a 25 percent adherence improvement and $71.4 million in incremental lifetime value across 100,000 patients.

The identification of racial, ethnic, and geographic paradoxes in DTC accessibility represents perhaps the most important finding for health equity considerations. Populations with the greatest pharmaceutical access needs often face the greatest digital access barriers, creating structural limitations on DTC distribution's potential to increase patient access. This suggests that DTC should be understood as a segment-specific channel requiring complementary policy interventions rather than a universal solution to pharmaceutical access challenges.

As the pharmaceutical industry continues its evolution toward direct patient relationships, commercial leaders and policymakers alike must approach these emerging channels with clear-eyed assessment of their genuine benefits and inherent limitations. The findings presented here provide a framework for such assessment, grounded in quantitative analysis of market factors and patient segment characteristics rather than the promotional narratives that often accompany emerging distribution innovations.

REFERENCES

1. Eli Lilly and Company. LillyDirect launches to provide access to Lilly medications for certain chronic diseases. Press release. January 4, 2024.
2. Pfizer Inc. PfizerForAll launches as new direct-to-patient platform. Corporate announcement. March 2024.
3. The White House. Delivering most-favored-nation prescription drug pricing to American patients. Executive Order 14297. May 12, 2025.
4. Drug Channels Institute. The gross-to-net bubble reached a record $300+ billion in 2023. Drug Channels. February 2024.
5. IQVIA Institute for Human Data Science. The use of medicines in the U.S. 2024: Usage and spending trends and outlook to 2028. April 2024.
6. Eli Lilly and Company. 2024 Annual Report. Securities and Exchange Commission Form 10-K. February 2025.
7. U.S. Department of Health and Human Services. TrumpRx fact sheet: Participating manufacturers and pricing commitments. November 2025.
8. TrumpRx.gov. Medication pricing and availability. Accessed February 2026.
9. Managed Healthcare Executive. Here are the Pfizer drugs to be offered through TrumpRx. January 2026.
10. Kaiser Family Foundation. KFF health tracking poll: The public's views on prescription drug costs and drug pricing policy. December 2024.
11. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System prevalence data. 2024.
12. U.S. Census Bureau. American Community Survey 1-year estimates: Computer and internet use. 2023.
13. Federal Reserve Bank of Atlanta. The telehealth divide: Digital inequity in rural health care deserts. October 2024.
14. 46brooklyn Research. January 2026 brand drug price changes analysis. January 2026.

15. Cutler DM, Everett W. Thinking outside the pillbox: medication adherence as a priority for health care reform. N Engl J Med. 2010;362(17):1553-1555.

16. Iuga AO, McGuire MJ. Adherence and health care costs. Risk Manag Healthc Policy. 2014;7:35-44.

17. Avalere Health. Medication adherence in medically underserved areas. Avalere analysis. May 2024.

18. National Association of Insurance Commissioners. Trends in telehealth and its implications for health disparities. CIPR study. March 2022.

19. Health Resources and Services Administration. Designated health professional shortage areas statistics. 2024.

20. PMC. Disparities in digital access among American rural and urban households and implications for telemedicine-based services. J Rural Health. 2023;39(1):195-204.

21. Bessette LG, Anderson TS. Generalizability of clinical trials of novel weight loss medications to the US adult population. JAMA Intern Med. Published online November 25, 2024. doi:10.1001/jamainternmed.2024.6340

22. Hughes S, Rapfogel N. Following the money: untangling U.S. prescription drug financing. Center for American Progress. October 12, 2023.

Leveraging AI-Driven Multimodal Analysis to Accelerate Scientific Intelligence from a Major Cardiology Congress

Lori Klein, PharmD – Inizio Ignite, Putnam; Jo Ann Saitta, MS – Inizio Ignite, Putnam; Michael J. Harvey, PhD – Inizio Ignite, Putnam; Alexandra C. Dornbush, PhD – Inizio Ignite, Putnam

Abstract: The rapid expansion of scientific output at major medical congresses challenges life science teams seeking timely insights to inform strategy and medical engagement. At the 2025 American College of Cardiology (ACC) Annual Scientific Session, more than 11,000 abstracts and conversations from social media exceeded the limits of traditional manual review. This case study describes an AI-enabled, human-guided workflow integrating automated data extraction and large language model (LLM) analysis to identify emerging scientific and sentiment trends across priority cardiovascular topics. Abstracts and professional social media posts were analyzed using GPT-based models to characterize themes and synthesize cross-source insights. The approach identified actionable trends across lipoprotein(a), glucagon-like peptide-1 therapies, and cardiomyopathies, demonstrating accelerated insight generation and strategic value for Medical Affairs and commercial teams.

1. INTRODUCTION

Pharmaceutical companies make significant investments in major scientific congresses to generate and showcase data, making it essential to understand not only the evidence presented but also how it is received and interpreted across professional communities. A critical but often underemphasized strategic question is how to improve the quality and timing of upstream strategic decisions, particularly as new studies, insights, and stakeholder sentiment emerge that can alter a product's trajectory. The earlier an incorrect assumption, weak signal, or missed opportunity is identified, the greater the downstream impact. Conversely, late signal detection can lead to strategic misalignment with emerging science and professional sentiment.

Signals that inform pharmaceutical decision-making vary in both form and maturity. Peer-reviewed publications generally reflect validated findings shaped by extended publication timelines, capturing where the field currently stands. In contrast, scientific congresses and their associated abstracts often represent the earliest public articulation of emerging hypotheses, novel mechanisms, and preliminary clinical observations. They provide forward-looking visibility into areas of evolving scientific interest—capturing where the field could be heading *next* rather than where it stands *now*.

In recent years, scientific and clinical discourse has increasingly extended to professional social media platforms X and LinkedIn. These platforms function not only as channels for disseminating new publications and congress findings but also as forums for real-time interpretation and debate among clinicians and researchers. Social data serves as a complementary source of early perspective that, alongside scientific abstracts, strengthens

early signal detection without substituting for validated evidence.

While integrating scientific abstracts with real-time social conversations for early signal detection is strategically compelling, the volume and velocity of these data have historically posed practical challenges. Large scientific congresses generate thousands of abstracts alongside rapidly evolving social commentary, which is beyond the limits of manual review. Advances in natural language processing (NLP) have enabled systematic parsing of large text corpora to extract structured insights from scientific and social data.[1,2] More recently, LLMs have expanded these capabilities, improving speed, scale, and analytic capacity to better match the complexity of modern information environments.[3,4] Beyond analyzing large volumes of unstructured data, LLMs can integrate information across diverse sources and generate cross-source insights.[5] Their adaptability reduces the need for task-specific development,[6] enabling subject matter experts (SMEs) to explore and interpret unstructured data directly.

Building on these AI capabilities, this case study illustrates how AI can support SMEs in generating strategic insight. AI enhances speed and scale but does not operate autonomously; human expertise provides oversight, interpretation, and scientific grounding. AI enables earlier learning by identifying emerging signals that may inform portfolio strategy, evidence generation planning, and competitive positioning. The analysis applies an AI-enabled, human-guided workflow to the 2025 ACC Annual Scientific Session, demonstrating how LLMs can facilitate earlier insight from congress-scale scientific data and support more efficient evaluation of emerging trends relevant to strategic and Medical Affairs decision-making.

2. METHODS

2.1 Study Objectives and Analytical Scope

The objective of this study was to evaluate whether an LLM-enabled, human-guided workflow could accelerate structured synthesis of scientific abstracts and associated professional discourse from a major medical congress to inform early-stage pharmaceutical decision-making. The analysis focused on content generated in association with the 2025 ACC Annual Scientific Session, held on March 29 – April 1, 2025.

The analysis was organized around three predefined topic areas reflecting high strategic relevance in cardiovascular medicine. The associated research questions were developed by Therapeutic Area SMEs to address key pharmaceutical insight priorities commonly pursued by Medical Affairs and Commercial teams in congress analyses and scientific summaries.

Topic areas and associated questions

1. **Lipoprotein(a) (Lp(a))**
    a. How is Lp(a) currently perceived in terms of clinical relevance, and where is it positioned within the broader cardiovascular treatment paradigm?
    b. Do discussions around Lp(a) at ACC 2025 primarily focus on its role in primary prevention, secondary prevention, or both?

2. **Glucagon-like peptide-1 (GLP-1) therapies in cardiovascular disease**
    a. What themes and sentiment are emerging around GLP-1 therapies in relation to cardiovascular disease?

3. **Transthyretin amyloid cardiomyopathy (ATTR-CM)**

    a. What are the most frequently mentioned investigational therapies for ATTR-CM at ACC 2025? What is the tone of these mentions (e.g., positive, skeptical, neutral)?

    b. What themes and sentiment are emerging around clinical integration of new treatment options for ATTR-CM?

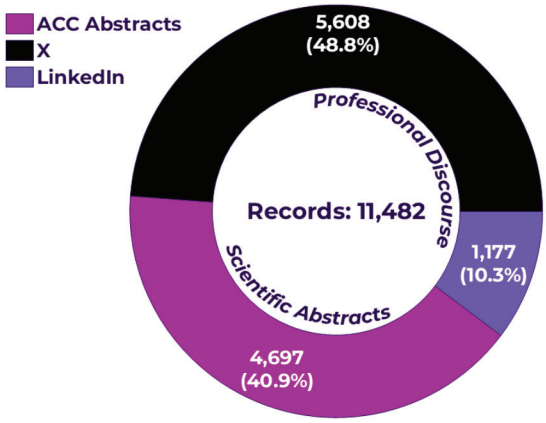


Figure 1. *Overview of scientific & digital data*

2.2 Data Sources and Collection

Two primary data sources were analyzed: scientific abstracts presented at ACC 2025 and publicly available professional discourse from LinkedIn and X (formerly Twitter).

Scientific Abstracts

The ACC 2025 congress program included more than 4,000 scientific abstracts (Figure 1), spanning clinical trials, observational studies, basic science, and health services research across the full spectrum of cardiovascular medicine. Topic-specific search terms corresponding to each of the three predefined topic areas (Lp(a), GLP-1 therapies, ATTR-CM) were used to identify relevant abstracts within the congress program.

Abstracts were extracted as images and converted into machine-readable text for analysis. Abstract retrieval included all categories in scope *except* Clinical Case Studies, given the forward-looking nature of the research questions.

Professional Discourse

Publicly available professional discourse referencing ACC 2025 was collected from LinkedIn and X using API-based retrieval. Predefined search terms (including Lp(a), Lipoprotein-A, ATTR, Transthyretin amyloid cardiomyopathy**,** GLP-1, GLP1, Glucagon-like peptide-1) and congress-related identifiers (including #ACC2025, #ACC25, #CardioTwitter, #CVPrev and therapeutic area keywords) were used to capture posts associated with the conference and relevant topic areas.

Social data were collected from March 28-April 1, 2025, spanning one day before and after ACC 2025. The dataset comprised approximately 7,000 social media posts, including over 5,000 posts from X and over 1,000 posts from LinkedIn (Figure 1). For each question, social content was processed using LLMs to remove duplicate entries, clearly non-informative text, and promotional material. Once the data was filtered, analyses focused on thematic content and qualitative tone. Engagement metrics (e.g., likes, reposts, amplification) were not included in our analysis.

2.3 AI-Enabled, Human-Guided Thematic Analysis

Analysis was conducted using ClarityNav, a proprietary AI-enabled insights platform developed by Inizio Ignite, Putnam as part of the InizioNavigator.ai suite. ClarityNav integrates a custom trained version of OpenAI's ChatGPT Enterprise (model 4o at the time of the analysis) via API calls with structured analytical workflows designed for large volumes of multimodal unstructured and structured data sources. The same workflow was applied independently to each of the three topic areas.

Step 1: Theme extraction from abstracts. Within each topic-specific abstract subset, ClarityNav was used to review text and identify recurring concepts and areas of scientific emphasis. Based on patterns observed across abstracts, the model clustered related content into preliminary themes (e.g., therapeutic interventions, diagnostic innovations, risk stratification, healthcare disparities). These initial clusters often reflected granular semantic distinctions within the text resulting in more narrowly defined categories.

Step 2: Expert refinement. Therapeutic Area SMEs reviewed the LLM-generated themes to ensure clinical coherence and relevance to the predefined research questions. During this stage, closely related or overly narrow themes were consolidated, uninformative groupings were removed, and theme definitions were refined by SMEs to connect clearly to the predefined research questions. This human-led refinement resulted in a finalized thematic framework for each topic area.

Step 3: Cross-source comparison. Once themes were established, the finalized thematic framework was applied to the corresponding social media dataset. Social posts were evaluated with ClarityNav against the locked scientific themes to assess areas of thematic alignment, divergence, and relative prominence across sources. This structured comparison enabled identification of topics emphasized consistently across both scientific and professional discourse as well as themes that were amplified, selectively highlighted, or received limited attention in social commentary.

Step 4: Tone characterization. Tone was characterized based on language patterns in both abstracts and professional discourse using ClarityNav. Data were described using terms such as *enthusiasm* (expressions of excitement or advocacy for clinical adoption), *optimism* (positive outlook on clinical translation), *caution* (measured commentary noting implementation barriers or evidence gaps), *skepticism* (expressions of doubt about feasibility or benefit), and *uncertainty* (question-oriented framing or calls for further evidence). Tone characterization was descriptive and interpretive; it did not involve automated numerical sentiment scoring.

Step 5: Structured signal evaluation. To ensure consistent interpretation across the three topic areas, findings were evaluated against three predefined criteria: (1) repeated thematic coverage across scientific abstracts and professional discourse rather than isolated mentions; (2) sustained volume of discussion, indicating concentrated attention rather than episodic reference; and (3) observable alignment or divergence between abstract-level findings and professional interpretation. These criteria were applied across all three topics to distinguish persistent signals from incidental or low-volume observations.

Through this workflow, LLMs were used to accelerate structured pattern recognition and clustering across large volumes of unstructured text, while final interpretation, thematic

consolidation, and strategic contextualization were performed by Therapeutic Area SMEs. Using this approach, thematic synthesis, cross-source comparison, and tone characterization were completed within 2 weeks of the conclusion of ACC 2025.

3. RESULTS

The combined dataset comprised approximately 11,000 records, including over 4,000 scientific abstracts and approximately 7,000 social media posts. Together, they provided a structured view of emerging scientific signals and their reception within professional discourse.

3.1 Lp(a)

Themes from scientific abstracts

As shown in Figure 2, more than 40 abstracts addressed Lp(a) at ACC 2025. One major cluster addressed Lp(a)'s mechanistic role as an independent, genetically determined risk factor for myocardial infarction, stroke, aortic stenosis, and major adverse cardiovascular events (MACE). Studies reinforced its relevance in both primary and secondary prevention, with particular attention to population-specific risk in South Asian, autoimmune, transplant, and vasospastic angina cohorts. A second cluster centered on risk prediction. Multiple abstracts explored Lp(a)'s integration into risk stratification models, including the AHA PREVENT equations. In one study, combined Lp(a) and CRP measurement improved 30-year Atherosclerotic Cardiovascular Disease (ASCVD) risk prediction in younger adults. A separate analysis reported that elevated Lp(a) ranked among the top predictors in a GPT-based MACE risk model, reflecting its emerging inclusion in AI-enabled clinical decision tools. Investigational therapeutics were also represented. The ALPACA trial reported that the siRNA therapy lepodisiran reduced Lp(a) levels by more than 90%, marking a significant development within the field.

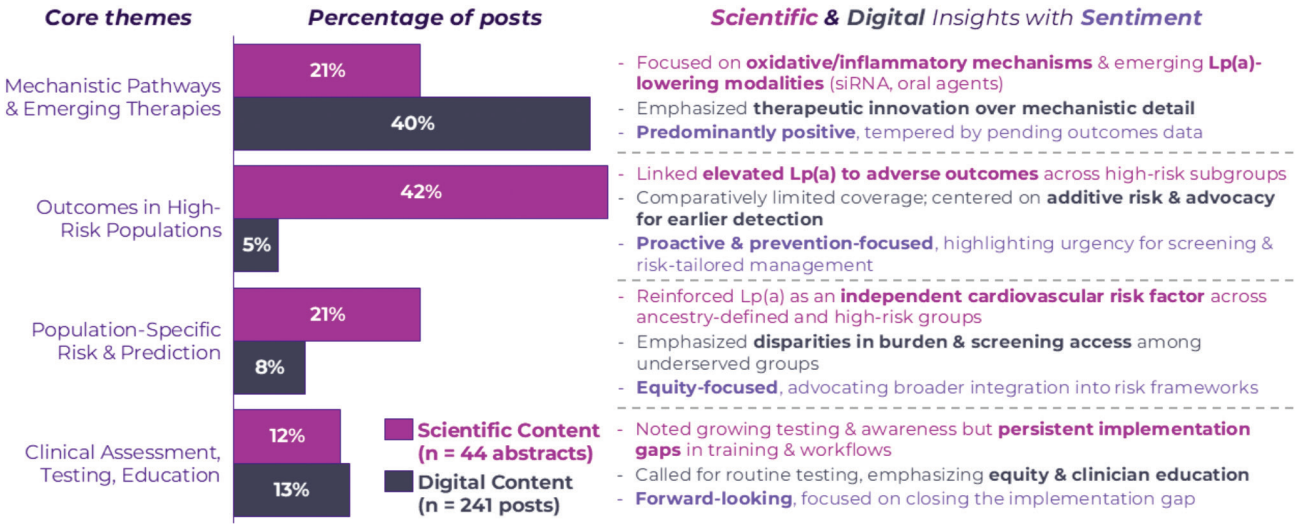


Figure 2. *Lp(a) cross-source thematic distribution & sentiment*

Cross-source comparison

Applying the finalized scientific themes to the professional discourse dataset revealed clear differences in thematic emphasis. The scientific program spanned mechanistic research, risk prediction, oxidation biology, and diverse population studies. Professional conversations, however, concentrated disproportionately on late-stage therapeutic innovations, particularly recently presented clinical trial data. Among these, the ALPACA trial and lepodisiran received notable attention, reflecting the prominence of the findings at the meeting.

At the same time, abstracts addressed both the therapeutic potential of Lp(a) and the absence of outcomes data demonstrating cardiovascular event reduction. Questions raised within the scientific program – how much Lp(a) lowering is needed, in which patients, and over what timeframe – were comparatively underrepresented in professional discourse.

Tone

Professional tone around Lp(a) was predominantly enthusiastic. Clinicians, researchers, and industry commentators expressed urgency around broader testing, therapeutic development, and patient education. However, a secondary thread of uncertainty was present, particularly among specialists who questioned the timeline to outcomes evidence and the readiness of clinical guidelines. This combination - strong enthusiasm alongside unresolved questions – is consistent for a field approaching an inflection point.

Taken together with the cross-source comparison, this pattern met the predefined signal evaluation criteria: Lp(a) appeared repeatedly across both data sources, discussion was sustained, and a measurable gap existed between abstract-level findings and professional interpretation.

3.2 GLP-1 receptor agonists

Themes from scientific abstracts

Abstracts addressing GLP-1 receptor agonists spanned a wide range of cardiovascular applications (Figure 3). The SOUL trial of oral semaglutide served as a focal point, demonstrating a 14% reduction in MACE among high-risk patients with type 2 diabetes (T2D), ASCVD, and/or chronic kidney disease (CKD). This was the first such outcome reported for an oral GLP-1 and was interpreted as validating the cardiovascular potential of the class while emphasizing the real-world advantages of oral delivery. Building on the SOUL trial, the STRIDE trial in peripheral artery disease (PAD) extended the evidence base, showing that semaglutide improved walking distance, symptoms, and ankle-brachial index in patients with limited pharmacological options and high functional burden.

Beyond major outcomes trials, several abstracts explored GLP-1 applications in rhythm management. Studies linked GLP-1 use to reduced risks of atrial fibrillation, cardioversion, and stroke in patients with T2D. A real-world analysis reported reduced incidence of ventricular arrhythmias, cardiac arrest, and cardiogenic shock, suggesting a role in broader electrophysiological risk reduction. Exploratory abstracts also presented early data on GLP-1 use in cardio-oncology, indicating potential to mitigate chemotherapy-induced cardiotoxicity, particularly from anthracyclines and 5-fluorouracil. Additional analyses showed GLP-1s outperforming bariatric surgery across multiple cardiovascular endpoints in obese patients, highlighting relevance in non-diabetic populations.

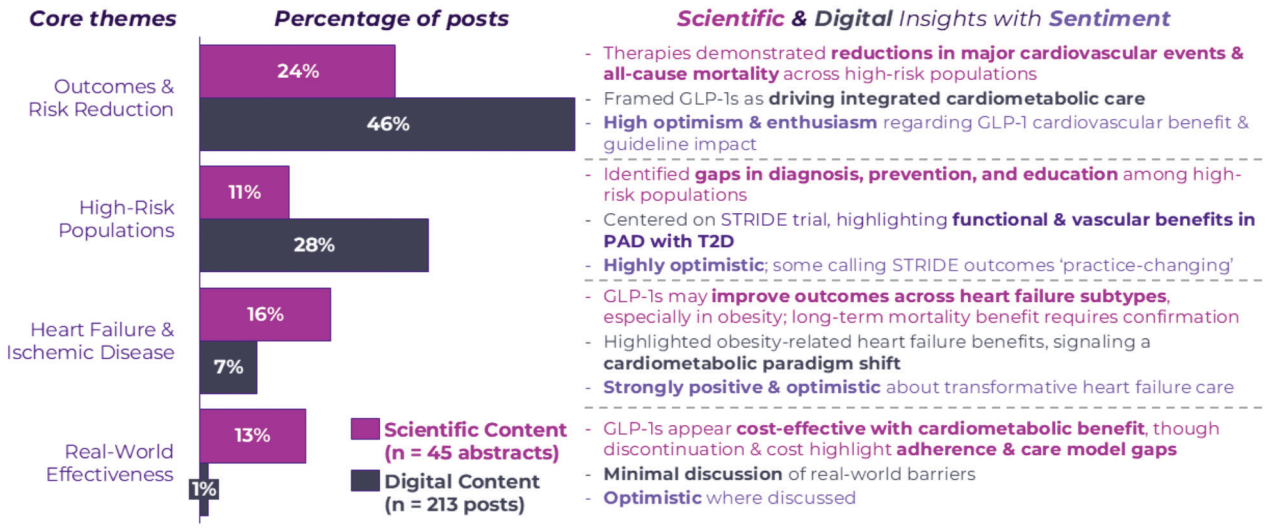


Figure 3. *GLP-1 cross-source thematic distribution & sentiment*

Cross-source comparison

Professional discourse showed strong alignment around major trial outcomes alongside selective emphasis. The SOUL and STRIDE trials, which had prominent findings in ACC 2025, were amplified in professional commentary. The oral formulation of semaglutide drew additional attention in discussions of adherence and access. These focal points were consistent across both scientific and social sources.

In contrast, several areas with substantial abstract coverage were comparatively underrepresented in professional discourse. Heart failure integration, real-world effectiveness, and health system implementation - each addressed by multiple abstracts - did not feature prominently in social commentary. Together, these findings indicate strong alignment around major trial outcomes, alongside comparatively limited attention to implementation-focused themes.

Tone

Professional tone around GLP-1 therapies was overwhelmingly positive. Commentary framed GLP-1 receptor agonists as foundational cardiometabolic therapies rather than glucose-lowering agents with incidental cardiovascular benefits. Skepticism was notably absent, and discussion frequently emphasized the transformative implications of recent trial data, with clinicians describing SOUL as "practice-changing" and positioning STRIDE as a breakthrough in PAD management. The near-uniform enthusiasm, combined with strong trial-level evidence, suggests that professional consensus around the cardiovascular role of GLP-1s is consolidating rather than emerging.

This pattern met the predefined signal evaluation criteria: repeated cross-source coverage, sustained discussion, and measurable divergence between abstract-level findings and professional interpretation.

3.3 ATTR-CM

Themes from scientific abstracts

ATTR-CM abstracts at ACC 2025 reflected a field in which investigational therapies are rapidly advancing (Figure 4), with two new agents serving as key focal points. In the ATTRibute-CM trial, acoramidis (Attruby/ Beyonttra) demonstrated greater than 90% transthyretin (TTR) stabilization along with reductions in cardiovascular-associated hospitalizations and mortality. Vutrisiran (Amvuttra), a new RNAi therapy, was highlighted for its novel mechanism of silencing TTR production at the source, favorable quarterly subcutaneous dosing, and a broad label covering cardiovascular event reduction.

Innovation in diagnostics emerged as a parallel theme. Abstracts described increased use of AI-based ECG tools, genetic testing, and point-of-care ultrasound to facilitate earlier diagnosis. Several studies emphasized the persistent challenge these technologies seek to address – diagnostic delays remain common and frequently result in patients presenting at later stages of disease when available treatments may be less effective.

Cross-source comparison

Commentary on ATTR-CM was broadly aligned with the presented clinical data but was tempered by concerns around implementation. Posts highlighted positive trial outcomes for acoramidis and the mechanistic approach of vutrisiran, while frequently raising concerns about pricing and access. The estimated annual cost of $470,000 for vutrisiran was commonly cited as a potential barrier for patient access, with discussion emphasizing the risk that promising therapies could face insurance barriers or restrictive reimbursement policies.

The problem of diagnostic delay was also amplified. Posts emphasized the importance of recognizing early warning signs such as bilateral carpal tunnel syndrome, lumbar spinal stenosis, and unexplained neuropathy. This emphasis was consistent with the scientific program's focus on improving screening but extended the discussion to multidisciplinary care models involving cardiologists, neurologists, and surgeons.

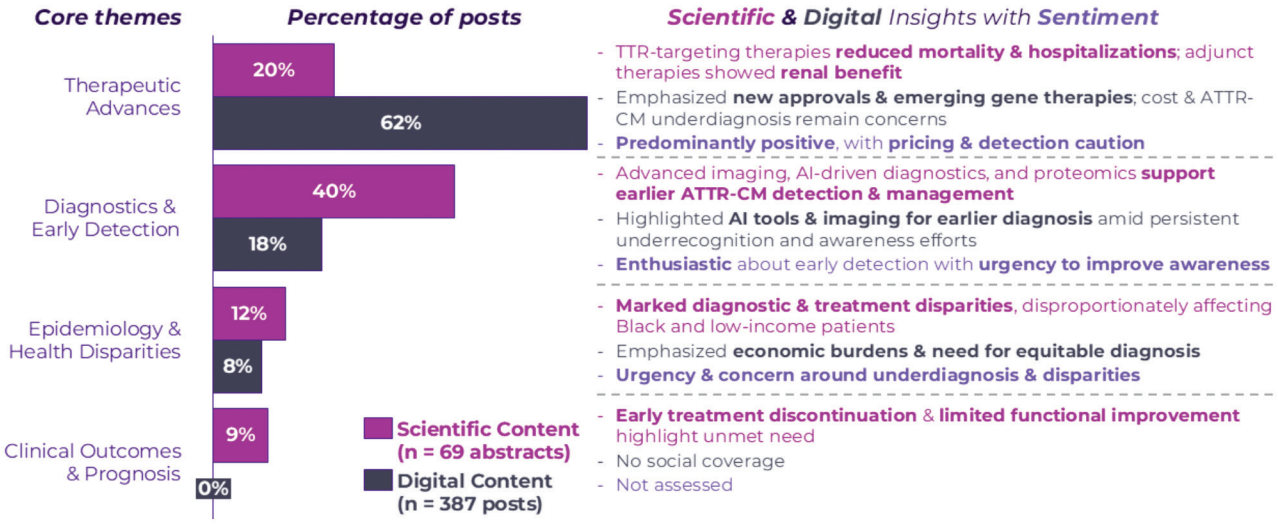


Figure 4. *ATTR-CM cross-source thematic distribution & sentiment*

Tone

Professional tone around ATTR-CM was optimistic regarding scientific progress but cautious regarding real-world implementation. Clinical trial data were viewed positively, although discussion frequently reflected concern about whether diagnostic pathways, specialist access, and reimbursement structures are adequately prepared to support widespread adoption of these new therapies.

In ATTR-CM, strong confidence in emerging therapies was coupled with sustained concern about real-world implementation, distinguishing it from the other two topics: in Lp(a), enthusiasm outpaced evidence, whereas in GLP-1s, enthusiasm and evidence were closely aligned.

3.4 Summary

Across topic areas, cross-source comparison showed that professional discourse was consistently narrower than the scientific program. In Lp(a), more than 40 abstracts addressed diverse aspects of the biomarker, yet social discussion concentrated on ALPACA and lepodisiran. For GLP-1s, abstracts spanned ASCVD, PAD, heart failure, arrhythmia management, and cardio-oncology, whereas professional commentary focused on SOUL and STRIDE with comparatively limited attention to heart failure integration and real-world effectiveness. In ATTR-CM, scientific content covering gene therapy, diagnostics, care disparities, and treatment sequencing was again broader than the professional conversation, which centered on specific approved or late-stage therapies. Across all three areas, professional discourse disproportionately amplified late-stage clinical data and named therapeutic assets, while earlier-stage science and broader disease considerations received comparatively less attention.

Beyond differences in thematic breadth, tonal analysis revealed distinct patterns for each therapeutic area. For GLP-1s, overwhelmingly enthusiastic framing and minimal skepticism suggest consolidating professional consensus. In Lp(a), enthusiasm was high but accompanied by persistent uncertainty, particularly regarding the timeline to outcomes evidence. In ATTR-CM, optimism about scientific progress coexisted with caution about real-world delivery.

4. DISCUSSION

This case study evaluated whether an AI-enabled, human-guided workflow could accelerate structured synthesis of congress-scale scientific abstracts and professional discourse to support earlier strategic insight in pharmaceutical decision-making. Across three predefined cardiovascular topics, consistent structural and tonal patterns emerged in how scientific findings were reflected in professional conversation. Prominent clinical developments drew concentrated attention, while broader mechanistic and implementation-oriented research received comparatively less visibility. At the same time, variation in tone reflected differing levels of confidence and perceived field maturity. Applied to ACC 2025, this approach demonstrated how structured synthesis of scientific and professional discourse can surface relevant signals for pharmaceutical strategy. These findings illustrate how congress-scale scientific and professional discourse can reveal emerging signals before narratives consolidate, enabling earlier strategic awareness of where a field may be heading.

Thematic Patterns and Cross-Source Comparisons

The concentration patterns observed across the topic areas have direct strategic implications for pharmaceutical teams. Professional discourse did not mirror the full breadth of the scientific

program; instead, attention clustered around a subset of prominent developments. While this emphasis is understandable given the relevance of pivotal data, it creates the possibility that other meaningful advances receive comparatively less visibility.

Cross-source comparison provides a structured way to evaluate this imbalance. When scientific depth and professional attention move in parallel, emerging trends may reflect durable momentum. However, divergence between the two warrants closer examination. Topics receiving concentrated visibility despite limited evidentiary maturity may require disciplined validation. Conversely, areas with substantial scientific activity but limited amplification may represent under-recognized inflection points. Identifying such divergence early supports a more balanced assessment of emerging trends.

System-Level Barriers Across Topics

Beyond selective amplification, the scientific program repeatedly highlighted practical challenges that extended across therapeutic areas. These included diagnostic access and timing (ATTR-CM diagnostic delays, the role of AI-enabled screening tools); treatment sequencing (how to position GLP-1s within broader cardiometabolic regimens, how to incorporate Lp(a) testing into risk assessment pathways); cost and reimbursement (ATTR-CM pricing, payer uncertainty around novel mechanisms); evidentiary maturity (unresolved questions regarding outcomes data and guideline readiness in Lp(a)); and equity (disparities in ATTR-CM diagnosis by race and socioeconomic status, geographic barriers to specialist care).

The recurrence of these issues across multiple areas of cardiology suggests that they reflect systemic challenges rather than isolated concerns. Importantly, these themes surfaced alongside therapeutic advances, indicating that scientific progress and implementation readiness do not always advance in parallel. Early identification of such friction points may help teams anticipate barriers before they materially affect adoption.

Tonal Patterns: Indicators of Maturity and Confidence

Across the three topic areas, distinct tones reflected differing levels of consensus, enthusiasm, uncertainty, and caution. In Lp(a), enthusiasm was prominent but accompanied by persistent uncertainty, particularly regarding outcomes data and guideline integration. This combination reflects an area gaining momentum while key pieces of evidence are still needed. In GLP-1 therapies, abstracts and discourse were largely aligned and characterized by overwhelming enthusiasm and minimal skepticism. ATTR-CM presented a different balance: confidence in therapeutic progress was evident, yet discussion frequently emphasized diagnostic limitations, reimbursement considerations, and implementation challenges. Scientific advancement was viewed positively, but real-world readiness tempered optimism. Taken together, these tonal differences suggest that sentiment may serve as a qualitative indicator of field maturity and perceived readiness for clinical integration.

Implications for Pharmaceutical Decision -Making

For pharmaceutical teams, earlier identification of emerging barriers and evolving narratives–before a therapy reaches market–enables more informed strategic dialogue and more timely evidence and portfolio decisions while development pathways remain flexible. As demonstrated in this study, structured cross-source analysis can reveal where uncertainty, debate, or unmet need is gaining

attention, distinguishing areas of concentrated momentum from those supported by meaningful scientific activity but receiving comparatively less attention. Even modest signals within scientific programs or discourse may carry disproportionate implications, underscoring the need for disciplined validation and expert interpretation before translating attention into strategic commitment. Incorporating these signals during development reduces the risk of misalignment between clinical evidence, stakeholder expectations, and real-world adoption at launch.

Structured synthesis of scientific and professional discourse can inform tailored strategic planning and field execution, as shown in the AI-enabled pharmaceutical strategy feedback loop in Figure 5. Those activities, in turn, can influence real-world outcomes, including changes in healthcare professional beliefs, engagement patterns, and clinical behavior. Evaluating these outcomes and reintegrating them into subsequent analysis enables refinement of future strategy. In this way, insight generation becomes part of a continuous feedback loop linking analysis, strategy development, execution, and measurable impact. Closing this loop supports a more iterative, evidence-informed process in which successive cycles of analysis enable increasingly precise engagement over time.

Importance of the AI-Enabled, Human-Guided Framework

The value of this approach lies not in automation alone, but in the combination of computational scale and expert oversight. LLM-supported analysis enabled rapid synthesis of thousands of abstracts and professional posts, surfacing patterns that would be difficult to assemble manually within comparable timeframes. However, interpretation and strategic judgment remained firmly human-directed.

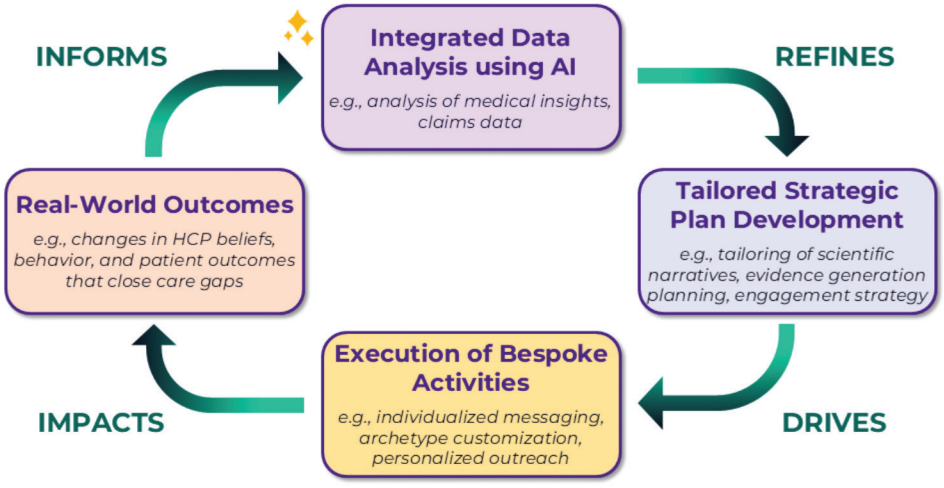


Figure 5. *AI-enabled pharmaceutical strategy feedback loop*

In pharmaceutical settings—where evidence standards, regulatory scrutiny, and clinical consequences are significant—acceleration must be paired with validation. Human guidance ensures emerging patterns are evaluated against clinical plausibility and development context. Rather than replacing expertise, AI serves as a tool to expand its reach. By reducing the time required for large-scale information synthesis, this approach creates space for more deliberate scientific discussion and cross-functional alignment. Applied in this case study, it translated congress-scale information into timely, decision-relevant insight while preserving rigor.

Limitations

This analysis relied on publicly available discourse, which may not capture the full range of clinical perspectives or offline discussions at the congress. Engagement metrics were not incorporated, limiting assessment of the relative reach or influence of individual posts. The evaluation window was confined to the congress period and therefore does not reflect longitudinal evolution of themes or sentiment. Tone characterization relied on LLMs and expert cross-validation rather than quantitative sentiment scoring, which may limit comparability with numerically scored studies.

Future Evolution

The present case study reflects an episodic application of AI-enabled synthesis to a single major congress. Future development could extend this approach toward more continuous monitoring of scientific publications, congress outputs, and professional discourse, enabling earlier detection of emerging themes as they evolve over time.

Integration with additional data sources, including real-world utilization and claims data, represents a potentially fruitful opportunity. Linking early scientific and sentiment signals to subsequent changes in treatment adoption or utilization would allow evaluation of how discourse translates into measurable clinical impact. Such integration would strengthen the feedback loop between analysis, strategic planning, and observed outcomes.

As these capabilities mature, semi-automated or agent-assisted workflows may support ongoing signal detection. However, governance, compliance, and expert oversight will remain essential to ensure interpretative rigor and responsible application within regulated pharmaceutical environments.

5. CONCLUSION

Major scientific congresses generate volumes of information that exceed the limits of traditional manual review, yet early interpretation of these signals can meaningfully shape pharmaceutical strategy. This case study demonstrates that an AI-enabled, human-guided workflow can synthesize congress-scale scientific abstracts and professional discourse in a structured and time-sensitive manner without compromising interpretive rigor.

Across Lp(a), GLP-1 therapies, and ATTR-CM, integrating thematic extraction, cross-source comparison, and tone characterization revealed patterns that would have been difficult to identify through sequential and manual review alone. Differences in scientific breadth, professional emphasis, evidentiary maturity, and implementation readiness became visible when abstracts and professional discourse were analyzed together. These cross-source dynamics clarified where enthusiasm aligned with evidence, where uncertainty persisted, and where implementation barriers may influence real-world adoption.

The value of this approach lies not in automation alone, but in the integration of LLM-enabled scale with expert oversight. LLMs enabled rapid synthesis across thousands of data points, while Therapeutic Area experts ensured that emerging patterns were evaluated within appropriate clinical and strategic context. When applied thoughtfully, this framework extends expert judgment, shortens the interval between signal emergence and strategic awareness, and supports more responsive, evidence-grounded decision-making.

REFERENCES

1. Savova GK, Masanz JJ, Ogren PV, Jiaping Z, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507-513. Available from: https://doi.org/10.1136/jamia.2009.001560

2. Paul M, Dredze M. You are what you tweet: Analyzing twitter for public health. Proc Int AAAI Conf Weblogs Soc Media. 2011;5(1):265-272. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14137/13986

3. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. Proc EMNLP-IJCNLP. 2019:3615-3620. Available from: https://aclanthology.org/D19-1371/

4. Nguyen DQ, Vu T, Nguyen AT. BERTweet: a pre-trained language model for English Tweets. Proc EMNLP Syst Demonstrations. 2020:9-14. Available from: https://aclanthology.org/2020.emnlp-demos.2/

5. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459-9474. Available from: https://dl.acm.org/doi/abs/10.5555/3495724.3496517

6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-1901. Available from: https://dl.acm.org/doi/abs/10.5555/3495724.3495883

ABOUT THE AUTHORS

Lori Klein, PharmD, is a Partner and the Medical and Scientific Affairs Practice Lead at Inizio Ignite, Putnam, a strategic consulting firm serving pharmaceutical and life science companies. She brings over 25 years of industry experience spanning medical affairs strategy, Lori is a member of Inizio Advisory Executive AI Working Group and Chairs Putnam's AI Working Group. She is an author of the MAPS GenAI white paper, speaks frequently on AI in Medical Affairs, and has experience partnering with clients to design bespoke AI solutions. Lori is active in MAPS as the Data Gap Identification Competency Lead, AI lead for the Medical Insights Competency, and a frequent presenter on data analytics and AI. Lori holds a PharmD from the University of Minnesota and completed a postdoctoral fellowship in critical care and infectious diseases.

Jo Ann Saitta, MS, is a Partner and Head of the Global Data Strategy, Analytics & AI Practice at Inizio Ignite, Putnam. She leads engagements with life science clients to apply data, analytics, and AI to solve complex problems and drive sustainable business growth. Jo Ann has over 20 years of C-level strategy experience in the healthcare and wellness technology field. Her work focuses on digital technology and data-driven business transformation initiatives, including new operational models, mergers and acquisitions, product development, and strategic partnerships. She is a thought leadership speaker, most recently at FiercePharma Week, Digital Pharma Advances, and Age of AI. She holds an MS in Computer Science from the New Jersey Institute of Technology, and bachelor's degrees in Computer Science and Political Science from Rutgers University.

Michael J. Harvey, PhD, is a Manager at Inizio Ignite, Putnam, where he focuses on pharmaceutical commercialization strategy, forecasting, and decision science. He holds a PhD in Health Services Organization and Policy from the University of Michigan, with a specialty in Decision Sciences and Operations Research, and a MRes in Medicine from the University of St Andrews.

Alexandra C. Dornbush, PhD, is a Life Sciences Consultant at Inizio Ignite, Putnam. She holds a PhD in Chemical and Physical Biology from Vanderbilt University. Alexandra focuses on supporting pharmaceutical organizations in achieving strategic objectives through actionable insight generation, with an emphasis on informing Medical Affairs strategy and cross-functional decision-making.

Enabling AI-Driven Medical Insight Through Modular Analytics Architecture: Lessons from Field-Based Medical Affairs

Jane Urban and Abhishek Trigunait

Article Type: Review / Expert Opinion

1. INTRODUCTION

Medical affairs organizations operate in an environment defined by accelerating scientific discovery, expanding data availability, and increasing expectations for timely, evidence-based engagement with healthcare professionals. Advances in genomics, real-world evidence, digital health, and clinical research have produced volumes of information that exceed the capacity of traditional analytics and manual review processes.[1,2] At the same time, pharmaceutical organizations are managing more specialized therapies, smaller patient populations, and increasingly complex standards of care.[3]

Within this landscape, Medical Science Liaisons (MSLs) serve as field-based scientific experts who engage clinicians in peer-to-peer dialogue, synthesize emerging evidence, and surface real-world insights that inform medical strategy. However, the tools supporting these activities have not evolved at the same pace as the scientific ecosystem, creating a gap between information availability and its practical use in scientific engagement.[1]

A foundational principle of Medical Affairs is the preservation of scientific exchange, which is distinct from promotional activity in both regulatory framework and purpose. Engagement must remain balanced, evidence-based, and grounded in scientific rigor.

Artificial intelligence (AI) can enhance how medical affairs teams access, synthesize, and apply scientific information by accelerating insight generation and reducing the time between evidence emergence and informed action, while maintaining human oversight and judgment.[4,5] As noted in MAPS 2025 Executive Insights, many organizations are shifting from activity metrics toward outcomes-focused KPIs, and modular AI architectures can support this transition by linking qualitative field insights to strategic priorities.[15]

This paper examines how modular, vendor-agnostic analytics architectures can enable responsible and scalable application of AI within medical affairs, particularly in supporting MSL workflows while preserving scientific rigor, regulatory compliance, and human oversight.[6,7]

2. THE MEDICAL AFFAIRS ANALYTICS PROBLEM:

Why Traditional Approaches Fall Short for MSLs

Medical affairs organizations face a distinct analytics challenge compared with other pharmaceutical functions. While the core metric of success in Commercial analytics is frequently tied to tangible outputs like prescription volume or market share, the 'value' in Medical Affairs is often found in the nuance of a scientific objection or a clinician's hypothesis regarding a sub-population. Consequently, medical analytics must prioritize contextual depth rather than transactional breadth, treating detailed field insights as high-value signals rather than simple activity metrics.[1,2]

This distinction is most evident in the Medical Science Liaison (MSL) role. MSLs are field-based scientific experts, often with clinical or academic backgrounds, who engage healthcare professionals in peer-to-peer scientific dialogue and document detailed narratives of these interactions. These notes may capture perspectives on emerging evidence, treatment patterns, safety considerations, dosing practices, comorbidities, and evolving standards of care.[1]

As a result, medical affairs generates large volumes of unstructured data. CRM systems supporting MSL teams often contain extensive free-text notes, literature references, follow-up questions, and longitudinal conversation histories. Although this information represents a valuable source of real-world scientific intelligence, it remains difficult to analyze systematically using traditional analytics approaches, which are optimized for structured and repeatable measures rather than narrative and context-dependent inputs.[2,3]

Natural language processing techniques have been applied to medical affairs data, but historically these approaches required extensive manual feature engineering, predefined taxonomies, and rigid classification schemes. Such methods are slow to adapt as scientific language evolves and are often poorly suited to capturing nuance in clinician sentiment or emerging hypotheses.[3,4] These limitations are compounded by the time-sensitive nature of medical affairs work, where MSLs may respond to unsolicited medical inquiries involving safety concerns or complex clinical scenarios that require accurate and timely interpretation.[1]

These analytical limitations are reinforced by stringent compliance requirements. Scientific exchange must remain balanced, evidence-based, and appropriately documented, limiting the applicability of point solutions that prioritize optimization over scientific rigor.[8,9] Additionally, the heterogeneity of medical affairs stakeholders—ranging from academic thought leaders to community clinicians—makes standardized analytical outputs insufficient for field-based scientific engagement.[1]

Together, these factors create a persistent gap between data availability and actionable insight within medical affairs. Addressing this gap requires analytical approaches that treat qualitative data as a first-class input, support real-time synthesis and interpretation, and embed intelligence directly into the workflows of field-based scientific experts.[4,6]

3. A MODULAR ARCHITECTURE FOR AI-ENABLED MEDICAL AFFAIRS

Addressing the analytical challenges facing medical affairs requires an architectural approach that supports flexibility, governance, and scalability while respecting the scientific, regulatory, and operational realities of field-based medical teams. A modular, AI-enabled analytics architecture provides a practical foundation for achieving these goals without disrupting established workflows or compromising compliance.[5]

Because MSL interactions are peer-to-peer and exploratory, the data generated is inherently more complex than the structured logs typical of commercial sales environments. Standard CRM analytics are often unable to capture the subtle shifts in scientific sentiment contained within long-form narratives. A modular architecture addresses this challenge by enabling large language models to act as interpreters of unstructured data, identifying emerging scientific themes across field interactions without relying on rigid, predefined taxonomies.

3.1 Architectural Principles

At the core of this approach is a modular, vendor-agnostic design that decouples data ingestion, transformation, analytics, and user interfaces. Rather than embedding analytical logic within individual applications, each layer operates as an independent component with well-defined interfaces. This separation allows organizations to evolve capabilities—such as AI models or visualization tools—without requiring wholesale system redesign.[5]

This flexibility is particularly important given the rapid pace of innovation in artificial intelligence. New large language models, orchestration frameworks, and analytical techniques continue to emerge, often with substantial improvements in capability. Architectures that tightly couple workflows to a single model or vendor risk rapid obsolescence or costly reengineering.[5]

Field-native intelligence is a second foundational principle. Medical affairs insights are generated and consumed in distributed, time-constrained environments. MSLs frequently prepare for scientific discussions on short notice, respond to unsolicited medical inquiries, and synthesize evidence between engagements. Analytical systems that require users to leave core workflows or rely on centralized interpretation introduce friction and limit adoption. Intelligence must therefore surface contextually within the tools MSLs already use.[1]

A critical requirement in Medical Affairs is preservation of the scientific firewall. While commercial analytics often seeks to optimize engagement frequency, medical affairs analytics must protect the integrity of scientific exchange. A modular architecture supports this distinction by ensuring that data sources used for insight generation remain separated from promotional datasets, reinforcing compliance guardrails that might otherwise rely on manual oversight.

Finally, the architecture must be explicitly human-centered. AI functions as an augmentative layer that enhances scientific judgment rather than replacing it. Transparency, explainability, and auditability are therefore treated as design requirements, ensuring that analytical outputs can be trusted, interrogated, and governed appropriately within a regulated scientific environment.[9]

3.2 Data and Transformation Layers

The foundation of the architecture is a comprehensive data layer that includes structured data—such as CRM metadata, interaction classifications, and operational attributes—as well as unstructured and semi-structured sources, including free-text MSL notes, scientific publications, regulatory guidance, clinical trial summaries, internal medical content, and real-world evidence narratives.[2]

Preserving data fidelity at this stage is essential. Raw data should be retained alongside transformed representations to support auditability, traceability, and re-analysis as scientific questions evolve. Role-based access controls ensure that sensitive information is available only to appropriate users, reflecting the permissions and responsibilities of medical affairs personnel.[10]

Above the raw data layer sits a transformation and enrichment layer that prepares information for analytical use. Traditional data engineering processes—including normalization, entity resolution, versioning, and metadata tagging—provide the structural foundation for reliable analytics, while AI-enabled enrichment expands the scope and efficiency of these processes.[2]

Natural language models can extract entities, concepts, relationships, and themes from unstructured text at scale, reducing reliance on manually curated taxonomies. Because these models adapt to evolving scientific language and emerging mechanisms of action, they are better suited to capturing nuance than static rule-based systems. The result is a data foundation that remains structured enough for analytical reuse while retaining the flexibility needed to accommodate scientific complexity.[3]

3.3 Analytics and AI Layer

The analytics layer builds on the enriched data foundation to support a range of analytical approaches, from traditional statistical methods to advanced AI-driven techniques. Classical analytics—including descriptive statistics, trend analysis, rule-based logic, and validated reporting—remain essential, particularly for compliance-sensitive use cases and formal medical reporting.[2]

AI-enabled techniques complement these methods by enabling synthesis, pattern detection, and probabilistic reasoning across high-dimensional data. Large language models, for example, can synthesize unstructured medical text, generate balanced summaries of scientific literature, and identify relationships across disparate sources.[3,17]

A key design consideration is implementing AI capabilities through task-specific agents rather than monolithic models. Agents can be optimized for discrete functions—such as literature summarization, entity extraction, regulatory monitoring, or field insight aggregation—and orchestrated within defined workflows. This approach improves explainability, facilitates validation, and allows individual components to be updated or replaced as capabilities evolve.[5]

Industry trends reinforce this direction. Gartner predicts that by 2028, over 40% of enterprises will adopt hybrid multi-agent systems, supporting architectures in which specialized agents handle discrete functions such as literature synthesis and regulatory monitoring.[16]

AI within this layer is designed to support decision-making rather than automate it. Outputs should include source references, reasoning context, and confidence indicators, enabling users to understand how conclusions were reached and reinforcing human oversight in regulated environments.[9]

3.4 Interface and Workflow Integration

Analytical value is ultimately realized at the point of use. For medical affairs, this requires interfaces that integrate seamlessly into existing workflows rather than introducing new destinations for insight consumption. Workflow-embedded intelligence reduces cognitive load, accelerates adoption, and ensures that analytical outputs are available when and where decisions are made.[4]

Interfaces may take multiple forms depending on user role. For MSLs, AI-enabled briefings embedded within CRM systems can synthesize recent publications, prior interactions with a clinician, regional scientific trends, and relevant regulatory updates ahead of a scientific discussion. For medical managers and leadership, dashboards and conversational interfaces can surface aggregated field insights, highlight emerging knowledge gaps, and support strategic exploration.[4]

Personalization is a critical capability at this layer. Users require different levels of detail, scientific depth, and modes of interaction. AI systems can tailor outputs based on role, therapeutic area, and usage patterns, enabling scalable distribution of insight while maintaining relevance.[4]

As workflows evolve, the emphasis is also shifting from single prompt-response interactions to multi-agent orchestration. In this approach, specialized agents—focused on domains such as clinical trial data, safety monitoring, or literature analysis—collaborate to provide a more comprehensive view within existing workflows. This mixture-of-experts model mirrors the interdisciplinary nature of Medical Affairs, supporting cross-validation of insights across diverse datasets.

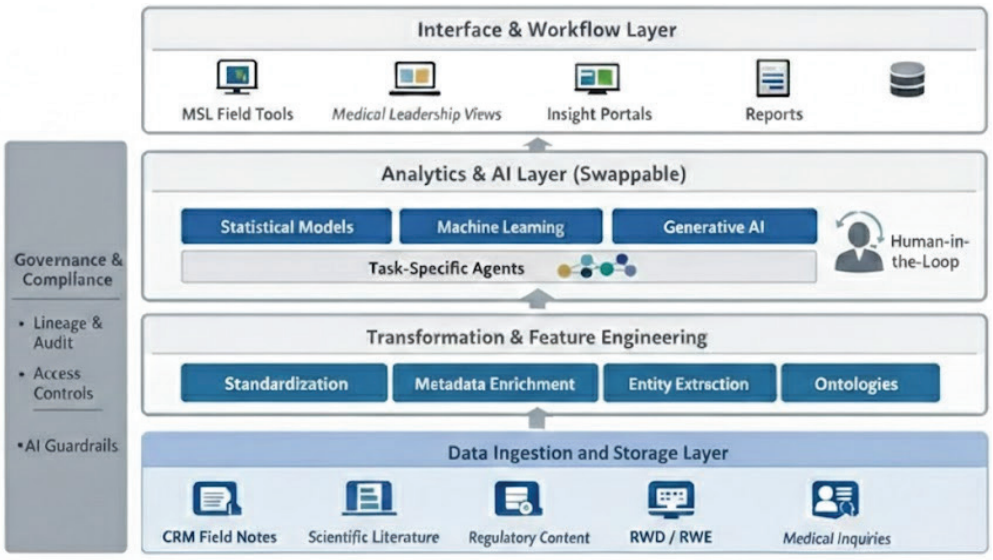
3.5 Governance Overlay

Governance spans all layers of the architecture, providing the controls necessary to ensure trust, compliance, and sustainability. In medical affairs, governance extends beyond data security to include scientific integrity, balanced evidence presentation, and regulatory alignment.[10]

Audit trails capture data lineage, transformation logic, model usage, and user interactions, enabling retrospective review and regulatory inspection when required. Access controls enforce role-based permissions, ensuring that sensitive information, such as internal strategy or early safety signals, is handled appropriately.[10]

Governance mechanisms also address AI-specific risks, including hallucination, bias, and overreliance on automated outputs. Systems should surface uncertainty, highlight conflicting evidence, and require human validation for high-impact decisions.[6]

By embedding governance into the architecture itself, organizations can support innovation while maintaining scientific rigor, regulatory compliance, and human judgment, an essential balance in the evolving landscape of medical affairs.[8]



Data ingestion, transformation, analytics, and interface layers are decoupled fo enable flexibility, governance, and scalability. Analytical logic and AI models are swappable without reengineering downstream workflows, supporting responsible experimentation and field-based insight delivery.

Figure 1. *Modular, vendor-agnostic analytics architecture for AI-enabled medical affairs.*

A layered architecture separating data ingestion, transformation, analytics, and workflow interfaces, with governance and human oversight embedded throughout.

4. APPLICATION OF A MODULAR AI ARCHITECTURE TO MEDICAL AFFAIRS USE CASES

Medical affairs functions present a compelling environment for AI-enabled analytics and decision support. Unlike many commercial roles, medical affairs teams—particularly Medical Science Liaisons (MSLs)—operate in highly variable, unstructured, and scientifically complex contexts. Their work requires deep domain expertise, rapid synthesis of evolving evidence, and careful adherence to regulatory and ethical standards. These characteristics highlight both the limitations of traditional analytics approaches and the potential value of modular, AI-enhanced architectures when thoughtfully applied.[1,10]

Consider a scenario in which an MSL is preparing for a scientific discussion with a leading oncologist on short notice. In a traditional workflow, the MSL might spend significant time manually searching literature and internal repositories for the latest evidence. In an AI-enabled modular environment, a briefing agent can synthesize recent publications, congress abstracts, and prior interaction history into a concise summary, allowing the MSL to focus more quickly on interpretation and engagement.

This section examines four medical affairs use cases where workflow-embedded analytics can improve effectiveness while preserving the central role of human judgment: literature synthesis, field insight aggregation, regulatory intelligence, and time-sensitive medical inquiry support.[3,4]

4.1 Evidence-Based Literature Synthesis for Scientific Exchange

One of the most consistent applications of AI within medical affairs is the synthesis of scientific literature. The volume of peer-reviewed publications, congress abstracts, preprints, and real-world evidence studies has grown beyond the capacity of any individual to track comprehensively. MSLs are expected to remain current across therapeutic areas, mechanisms of action, safety updates, and evolving standards of care, often while supporting multiple products or disease states.[3,17]

AI-enabled literature synthesis addresses this challenge by augmenting, rather than replacing, expert review. At a basic level, AI systems can generate concise summaries of individual publications, enabling rapid assessment of relevance and scientific contribution. More advanced implementations allow modular agents to search across large bodies of literature using defined criteria—such as population characteristics, endpoints, comparators, or competing therapies—and synthesize findings into structured thematic summaries.[3]

Evidence presentation in the form of literature synthesis must account for contradictory findings, limitations, and areas of uncertainty. This necessitates architectural design choices that prioritize transparency, traceability, and citation fidelity. AI-generated summaries should link directly to source material, surface divergent interpretations, and distinguish between established consensus and emerging hypotheses.[10]

When implemented within a modular architecture, literature synthesis agents can be reused across therapeutic areas while allowing configuration for disease-specific nuance. This approach supports scalability without sacrificing scientific rigor, enabling MSLs to focus on interpretation, dialogue, and clinical relevance rather than manual evidence aggregation.[5]

4.2 Aggregation and Interpretation of Field-Generated Scientific Insights

A second core application involves the aggregation and synthesis of qualitative insights generated through field medical interactions. MSLs routinely document scientific discussions with clinicians, capturing perspectives on treatment paradigms, real-world practice patterns, and emerging research. These notes represent a valuable source of real-world scientific intelligence within pharmaceutical organizations.[1]

Historically, however, the value of this information has been constrained by its unstructured nature. Field notes are difficult to aggregate, inconsistently coded, and challenging to analyze at scale, often leaving insights anecdotal, localized, or underutilized beyond individual territories.[2]

AI-enabled architectures can change this dynamic. Using natural language processing and large language models, qualitative field notes can be transformed into structured insight while preserving contextual nuance. At the individual level, MSLs may receive synthesized summaries of their engagements over time, highlighting recurring themes, frequently asked questions, or shifts in clinician sentiment. At the organizational level, aggregated insights can reveal regional or national trends, identify emerging scientific narratives, and surface opportunities for targeted education or evidence generation.[3]

Modular architectures also allow these capabilities to be selectively exposed. Medical leadership may view aggregated trends, while individual MSLs retain access to territory-

specific insights. With appropriate governance and filtering, select outputs can inform cross-functional alignment with clinical development or commercial teams without compromising scientific independence.[10]

4.3 Continuous Regulatory and Policy Intelligence

Regulatory intelligence is another domain where AI-enabled synthesis provides meaningful value to medical affairs teams. Regulatory guidance, label updates, safety communications, and policy changes evolve continuously across jurisdictions, making manual monitoring time-consuming and increasing the risk of delayed interpretation.[8]

AI agents embedded within a modular architecture can be configured to monitor regulatory sources, flag relevant updates, and generate tailored summaries aligned to specific therapeutic areas or products. These agents can incorporate complementary context—such as related therapies, class effects, or evolving standards of care—supporting more informed interpretation.[8]

For medical affairs teams, timely regulatory intelligence directly informs scientific exchange, inquiry response, and risk management. By embedding regulatory monitoring into existing workflows, organizations can reduce latency between regulatory change and field awareness while maintaining appropriate oversight and review processes.[8]

4.4 Support for Time-Sensitive Medical Inquiries

Finally, medical affairs teams routinely manage time-sensitive inquiries involving product safety, off-label questions, adverse events, or complex clinical scenarios. In these contexts, speed and accuracy are paramount, and delays can have clinical, regulatory, or reputational consequences.[4]

AI-enabled decision support can assist by rapidly synthesizing relevant evidence, prior inquiries, internal guidance, and regulatory constraints to support initial triage and response preparation. Importantly, such systems must be designed explicitly as support tools, not autonomous responders. Human review, medical judgment, and compliance oversight remain non-negotiable.[4,10]

Within a modular architecture, inquiry support agents can draw from multiple layers—literature repositories, field insight databases, regulatory content, and internal medical guidance—while maintaining strict access controls and auditability. This enables faster response preparation without compromising scientific integrity or compliance obligations.[10]

4.5 Implications for Medical Affairs Operating Models

Across these use cases, AI creates value by reducing cognitive and operational friction rather than replacing medical expertise. By embedding synthesis and contextual intelligence into workflows, medical affairs teams can shift time from manual information gathering toward higher-value scientific engagement.[12]

Modularity is essential because medical affairs use cases vary across disease biology, clinical practice, regulatory environments, and clinician needs. Point solutions that standardize workflows too rigidly often fail to gain adoption. In contrast, modular, vendor-agnostic architectures allow capabilities to be assembled, adapted, and governed in alignment with real-world medical practice.[5]

When implemented with appropriate controls, AI-enhanced architectures can improve responsiveness and scientific agility while preserving the core principles of balance, credibility, and patient-centered decision-making.[10,12]

5. GOVERNANCE, COMPLIANCE, AND TRUST IN AI-ENABLED MEDICAL AFFAIRS

As artificial intelligence becomes embedded in medical affairs workflows, governance and trust become foundational requirements. Medical affairs operates in a highly regulated environment where scientific credibility, patient safety, and compliance obligations are paramount. Unlike many commercial analytics applications, medical affairs use cases frequently involve unstructured scientific evidence, sensitive field interactions, and time-sensitive inquiries, requiring safeguards that reflect these risks.[10]

Our modular approach aligns with emerging regulatory guidance emphasizing risk-based frameworks and human oversight to protect patient safety in AI-assisted decision-making.[13,14]

5.1 Auditability, Content Integrity, and Scientific Traceability

A central governance requirement is maintaining a clear and auditable trail from source data to analytical output. When AI systems synthesize literature, summarize field interactions, or support inquiry responses, insights must be traceable to their originating sources, supporting internal review, external audit, and scientific defensibility.[10]

This requires controls that preserve underlying data and analytical logic. Field-generated content must be protected against unauthorized edits, with versioning and access controls that maintain data integrity. AI-generated summaries should reference source documents directly, enabling reviewers to verify accuracy and context rather than relying on opaque outputs.[10]

For literature synthesis, linking outputs to original publications—including citation metadata and publication dates—allows medical affairs professionals to evaluate the strength, recency, and relevance of the evidence.[3]

5.2 Role-Based Access and Permissions in Medical Affairs Workflows

Medical affairs organizations often operate with complex access and permission structures. MSLs may require broad access to scientific content and medical inquiries, while strict separation must be maintained between scientific exchange and promotional activity.[10]

AI-enabled systems must reflect these distinctions through granular role-based access controls. Some users require detailed scientific data, while others need only aggregated insights. Consistent enforcement of permissions across data, analytics, and interface layers is essential to prevent both over-restriction and compliance risk. Modular architectures that centralize access control while allowing flexible consumption help mitigate these challenges.[5]

5.3 Managing Hallucination and Bias in Literature Synthesis

AI-assisted literature review introduces risks related to hallucination, bias, and incomplete representation of evidence. In medical affairs contexts, these risks are consequential because synthesized content may inform scientific dialogue or inquiry responses.

Mitigation requires constraining outputs to verifiable source material, embedding citation requirements, and surfacing conflicting

findings and uncertainty rather than converging prematurely on a single narrative.[6] Retrieval-augmented generation (RAG) techniques support this approach by anchoring outputs to validated corpora, treating language models as reasoning engines rather than knowledge repositories.[16]

Access to subscription-based literature presents an additional consideration. AI-enabled systems must respect licensing constraints while enabling synthesis of authorized content within secure environments.

5.4 Continuous Regulatory Monitoring and Alignment

Regulatory requirements affecting medical affairs continue to evolve across jurisdictions. AI-enabled systems must therefore accommodate ongoing updates to source repositories, synthesis logic, and workflow constraints to remain aligned with current guidance.[8]

Embedding regulatory intelligence within the broader analytics architecture helps ensure that AI-supported insights reflect current standards. Guardrails should also prevent generation of content beyond approved scientific boundaries while still enabling rapid access to relevant evidence.[8,10]

5.5 Transparency, Explainability, and Human Oversight

Trust in AI-enabled systems depends on transparency and explainability. Users must understand how insights are generated, what data sources are used, and where AI support ends and human judgment begins.[9]

Human-in-the-loop design is therefore essential. AI systems should assist in synthesis and interpretation rather than act as autonomous decision-makers. Review mechanisms, escalation paths, and override capabilities must be defined, particularly for high-impact or time-sensitive scenarios.[4,10]

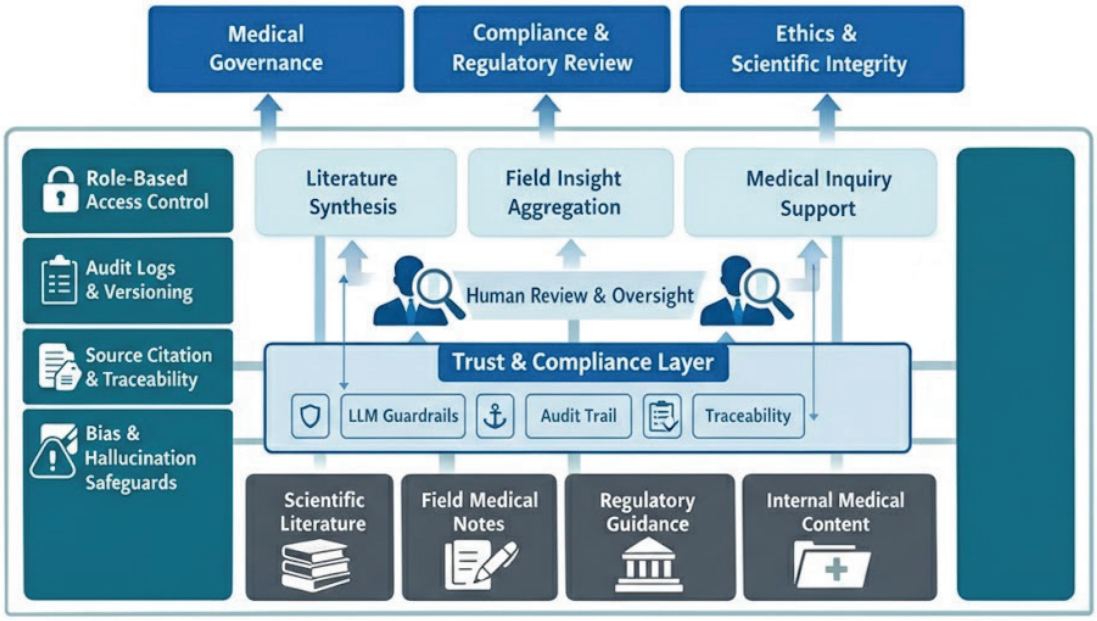


Figure 2. *Governance and human-in-the-loop safeguards for AI-enabled medical affairs. Illustration of audit trails, access controls, explainability, and regulatory oversight across AI-supported workflows.*

By combining architectural safeguards with organizational oversight, medical affairs teams can leverage AI responsibly while preserving scientific integrity. Governance, in this context, enables sustainable adoption rather than limiting innovation.[10]

6. ORGANIZATIONAL AND TALENT IMPLICATIONS FOR AI-ENABLED MEDICAL AFFAIRS

The adoption of AI-enabled analytics in medical affairs is not solely a technical challenge; it also requires organizational and cultural change. Medical affairs teams—and Medical Science Liaisons (MSLs) in particular—differ meaningfully from commercial or centralized analytics teams in professional background, working style, and expectations. These differences influence how AI tools must be designed, introduced, and governed.[1,12]

6.1 The Unique Profile of Medical Science Liaisons

MSLs are often recruited from clinical, academic, or research backgrounds and may have limited prior exposure to enterprise analytics platforms. Their training emphasizes scientific rigor, hypothesis-driven inquiry, and critical evaluation of evidence. As a result, comfort with technology varies, while expectations for accuracy, balance, and methodological transparency remain consistently high.[12]

This creates both opportunity and constraint. MSLs are well positioned to benefit from AI-supported synthesis and pattern recognition, but tools that require steep learning curves or provide opaque outputs are unlikely to gain trust or sustained adoption. AI systems must therefore support scientific workflows without introducing unnecessary technical complexity.[9,12]

6.2 AI as a Scientific Co-Pilot Rather Than a Replacement

Within medical affairs, AI is most effective when positioned as a scientific co-pilot rather than a replacement for expert judgment. Evidence discovery, literature review, and contextual understanding are core components of the MSL role. AI can accelerate access to relevant information and broaden situational awareness while leaving interpretation and decision-making to medical professionals.[3,12]

For example, AI-enabled systems can support exploratory evidence searches by identifying related mechanisms of action, adjacent therapeutic areas, or emerging hypotheses that may not be immediately apparent through traditional keyword-based searches. By providing broader ecosystem context, AI helps MSLs situate individual studies within a larger scientific narrative.[3]

6.3 Navigating Skepticism and Building Trust

Medical affairs professionals often approach AI with healthy skepticism rooted in scientific training and professional responsibility. Claims must be supported by evidence, methodologies must be defensible, and uncertainty must be acknowledged rather than obscured.[9,10]

Adoption strategies that recognize this context are more likely to succeed. Transparency is essential: users must understand how insights are generated, what data sources are included, and where limitations exist. Systems that provide citations, expose assumptions, and surface reasoning pathways are more likely to earn trust than those that present conclusions without explanation.[9]

Education also plays an important role. Training should address not only how to use AI tools, but how to evaluate them—recognizing appropriate use cases, identifying bias, and integrating AI-supported insights responsibly into scientific dialogue.[10,12]

6.4 Leadership Readiness and Cultural Enablement

Medical affairs leadership plays a critical role in shaping AI adoption. Leaders must balance openness to innovation with responsibility for scientific integrity and compliance. Many are both intellectually curious and methodologically conservative, seeking to explore new capabilities while demanding rigorous validation.[11]

Leadership-driven experimentation within defined guardrails can accelerate adoption. When leaders model thoughtful engagement with AI—using it to explore evidence landscapes or prepare for scientific discussions—they reinforce the role of AI as a support for scientific work rather than a substitute for expertise.[12]

At the organizational level, successful adoption requires alignment across medical, analytics, and technology teams. Involving MSLs and medical leaders in design, evaluation, and iteration helps ensure that tools reflect real-world scientific practice and are more likely to gain sustained adoption.[7]

6.5 Implications for the Future Medical Affairs Workforce

Over time, AI-enabled architectures are likely to influence how medical affairs roles evolve. As routine evidence aggregation and summarization become more automated, the relative value of synthesis, interpretation, and scientific dialogue will continue to grow. MSLs may spend less time locating information and more time contextualizing evidence, engaging in deeper discussions with clinicians, and identifying unmet scientific needs.

This evolution reinforces—not diminishes—the importance of the MSL role. By reducing cognitive and operational friction, AI enables medical affairs professionals to operate closer to the top of their expertise. Organizations that support this shift through thoughtful tool design, targeted training, and cultural alignment will be better positioned to realize the full potential of AI-enhanced medical affairs.

7. DISCUSSION AND FUTURE OUTLOOK

The accelerating pace of scientific discovery and the increasing complexity of therapeutic development and clinical practice have placed growing demands on medical affairs organizations. Traditional analytics and information management approaches—while foundational—are no longer sufficient on their own to support the depth, speed, and contextual nuance required for modern scientific exchange.[1,2] In this environment, artificial intelligence represents an evolution in how medical affairs teams generate, synthesize, and apply evidence.[3]

AI provides particular value in synthesis, second-order inquiry, and balanced evidence dissemination. These capabilities help teams navigate large and heterogeneous bodies of literature, surface emerging or conflicting perspectives, and reduce the operational burden associated with manual evidence aggregation.[4,5] When designed with appropriate safeguards, AI-enabled systems can strengthen scientific rigor by making uncertainty and limitations more visible rather than obscured.[6]

Architectural choices remain critical. Given the rapid evolution of AI technologies, tightly coupled systems introduce long-term risk.[8] Modular, swappable-layer architectures provide a practical alternative, enabling organizations to adapt to technological change while preserving governance, traceability, and operational continuity.[9] When aligned with medical affairs workflows and values, these architectures support greater scientific agility while remaining grounded in the mission of advancing patient care and evidence-based decision-making.[10]

The question is no longer whether AI will influence Medical Affairs, but how organizations will build the architectural and governance foundations necessary to guide that transformation responsibly.

REFERENCES

1. MassBio. The evolving role of medical affairs in biopharma [Internet]. Cambridge (MA): Massachusetts Biotechnology Council; 2023 [cited 2026 Feb 13]. Available from: https://massbioportalimages.blob.core.windows.net/massbioportalimages/38fdabc45178ea11a811000d3a378f47/xspfpaX.y6gZEMhFYTc.2.pdf

2. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst [Internet]. 2014;2:3 [cited 2026 Feb 13]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4341817/

3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med [Internet]. 2023;29:1930–1940 [cited 2026 Feb 13]. Available from: https://www.nature.com/articles/s41591-023-02448-8

4. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med [Internet]. 2020;3:17 [cited 2026 Feb 13]. Available from: https://www.nature.com/articles/s41746-020-0221-y

5. Bass L, Clements P, Kazman R. Software architecture in practice. 4th ed. Boston (MA): Addison-Wesley; 2021.

6. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv [Internet]. 2023;55(12):1–38 [cited 2026 Feb 13]. Available from: https://dl.acm.org/doi/10.1145/3571730

7. Kitsios F, Stefanakakis S, Kamariotou M, Dermentzoglou L. Artificial intelligence and business strategy towards digital transformation: a research agenda. Sustainability [Internet]. 2021;13(4):2025 [cited 2026 Feb 13]. Available from: https://www.mdpi.com/2071-1050/13/4/2025

8. European Medicines Agency. EMA regulatory science to 2025: strategic reflection [Internet]. Amsterdam: European Medicines Agency; 2020 Jan 20 [cited 2026 Feb 13]. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf

9. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning [Internet]. arXiv; 2017 [cited 2026 Feb 13]. Available from: https://arxiv.org/abs/1702.08608

10. European Commission. Ethics guidelines for trustworthy AI [Internet]. Brussels: European Commission; 2019 [cited 2026 Feb 13]. Available from: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

11. Davenport TH, Bean R. Data and analytics leadership annual executive survey 2023 [Internet]. NewVantage Partners; 2023 [cited 2026 Feb 13]. Available from: https://static1.squarespace.com/static/62adf3ca029a6808a6c5be30/t/63a477e23d7f9162b7771d6b/1671722982363/WAVESTONE+NVP--+Data+%26+Analytics+Executive+Leadership+2023--Survey+Findings.pdf

12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med [Internet]. 2019;25:44–56 [cited 2026 Feb 13]. Available from: https://www.nature.com/articles/s41591-018-0300-7

13. U.S. Food and Drug Administration. Guiding principles of good AI practice in drug development [Internet]. Silver Spring (MD): FDA; 2026 Jan 14 [cited 2026 Feb 13]. Available from: https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development

14. U.S. Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: guidance for industry and other interested parties [Internet]. Silver Spring (MD): FDA; 2025 Jan 6 [cited 2026 Feb 13]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

15. Medical Affairs Professional Society (MAPS). MAPS 2025 Americas Meeting: executive summary on AI and strategic impact [Internet]. 2025 Apr [cited 2026 Feb 13]. Available from: UPDATE_1.21.25-STRATEGY_Value-and-Impact-in-Medical-Affairs_SG-pwp.pdf

16. Gartner. Gartner identifies the top strategic technology trends for 2026 [Internet]. Stamford (CT): Gartner; 2025 Oct 20 [cited 2026 Feb 13]. Available from: https://www.gartner.com/en/articles/top-technology-trends-2026

17. Wang Y, Zhao Y, Zhang H, et al. Large language models in medicine: applications, challenges, and future directions. Int J Med Sci [Internet]. 2024;22:2792–2806 [cited 2026 Feb 13]. Available from: https://www.medsci.org/v22p2792.htm

Decoding Healthcare Organization Influence: A Metric for Institutional Control in Treatment Decisions

Eunseop Kim, Sr. Advisor, Consumer Analytics, Eli Lilly and Company

Abstract

Measuring institutional control—the extent to which healthcare organizations (HCOs) influence prescribing decisions—is critical for pharmaceutical strategy but remains difficult to quantify. This study proposes a model-based framework to quantify this influence as a standardized metric of organizational predictability. By employing a Bayesian hierarchical approach, the methodology captures variability in treatment choices while accounting for both patient-level characteristics and organizational heterogeneity. To ensure technical reliability, a simulation experiment was conducted to validate the framework's performance across varying degrees of institutional influence, followed by an application to a large-scale dataset of therapy escalation. Results demonstrate that this approach effectively differentiates institutional control levels across thousands of organizations, providing a scalable, probabilistic measure of influence. By identifying which organizations exert the most control over treatment pathways, this framework enables management to optimize commercial engagement, refine targeting strategies, and benchmark performance across diverse healthcare systems, offering a principled tool for enhancing strategic resource allocation.

1 Introduction

In the current pharmaceutical landscape, the consolidation of providers into large, integrated HCOs has introduced a significant layer of complexity for decision-making. The organizational landscape of U.S. healthcare has undergone a significant structural transition; as of 2024, private practices account for less than half of the physician workforce. Specifically, only 42.2% of physicians now remain in private practice, representing an 18-percentage point drop since 2012. This shift is largely driven by the expansion of hospital systems; the share of physicians in hospital-owned practices has climbed to 34.5%, a rise complemented by a doubling of those directly employed by or contracting with hospitals to 12.2% [1, 2].

As these systems move toward centralized governance, decision-making processes and supporting analytics must navigate an "institutional black box" where traditional provider-level models often fail to account for organizational mandates. Understanding the degree of control these organizations exert over their affiliated healthcare providers (HCPs) is a critical analytical requirement for effective resource allocation and strategic planning. This research addresses this challenge by introducing a principled, model-based framework to quantify institutional control—defined as the extent to which an organization influences or dictates the prescribing behavior of its providers across similar clinical scenarios.

Central to this framework is a patient-centric view of treatment variability. Rather than relying solely on aggregate volume metrics, the methodology focuses on the patient journey and how individuals with similar clinical profiles are managed across different institutional settings. By centering the analysis on the patient, the framework aims to distinguish between organizational influence and individual clinical autonomy. The operating assumption is as follows: an HCO with high control will exhibit consistent, predictable treatment patterns for patients with

comparable clinical needs, suggesting that organizational protocols, such as system-wide treatment guidelines or formulary restrictions, are the primary drivers of care. Conversely, a low-control environment exhibits greater variability across providers for the same patient types, reflecting a decentralized structure where individual HCPs maintain significant decision-making independence. From an analytical perspective, this distinction is vital; it ensures that the patient journey is not merely modeled as a series of isolated provider preferences but is understood within the broader context of the institutional forces shaping care delivery.

Quantifying institutional control offers significant utility for commercial analytics and organizational behavior modeling. A data-driven metric allows for a more sophisticated, behavior-based segmentation of the market, moving beyond simple volume or geographical metrics. This enables analytics functions to provide more tailored recommendations for engagement: in high-control environments, data might suggest a "top-down" strategy focusing on clinical leadership to ensure clinical alignment between system-wide protocols and current therapeutic evidence. In contrast, in organizations where provider autonomy remains the norm, engagement strategies would prioritize traditional "bottom-up" outreach to individual HCPs. By integrating this level of predictability into existing analytic frameworks, organizations can optimize field force deployment and refine omnichannel strategies to better align with the actual decision-making structure of an account.

To demonstrate the practical application of this framework, this investigation utilizes a pilot study within a major chronic disease therapeutic area, specifically examining the critical clinical event of therapy escalation. This event is defined as the progression of a patient from first-line treatments to advanced therapies By analyzing these escalation events across a broad national landscape of HCOs, the methodology identifies the proportion of prescribing heterogeneity attributable to organizational influence. While the pilot centers on a specific therapeutic transition, the underlying framework is built to be scalable and extensible, providing a consistent, objective lens through which analytics teams can view institutional influence across diverse therapeutic areas and healthcare settings to improve the precision of resource allocation and strategic interventions.

2 Methodology

2.1 Conceptual Framework

Institutional control is conceptualized as the degree of structured heterogeneity in prescribing behavior. In this framework, control is viewed not as a directly observable volume metric, but as a latent construct inferred from the consistency of clinical decisions across an organization's affiliated providers. In health services research, the organizational environment introduces systemic constraints, administrative incentives, and cultural norms that can significantly outweigh individual physician attributes in explaining practice variation [3].

Specifically, institutional control is assessed by analyzing the distribution of treatment choices across comparable clinical events. While high raw variability in an aggregate sense typically suggests low control, indicating that individual providers are making independent decisions based on personal judgment or idiosyncratic preference, this interpretation lacks the necessary nuance for institutional analytics. Identical levels of observed variability can emerge from fundamentally different underlying dynamics. One HCO may apply consistent, protocol-driven decisions across clearly defined patient subgroups, resulting in structured variability that reflects intentional clinical differentiation aimed at reducing unwarranted variation [4]. In such environments, the organization effectively minimizes idiosyncratic provider-level variation to achieve predictable, system-wide outcomes. In contrast, another HCO may exhibit the same level of variability due to uncoordinated, autonomous decision-making by individual providers, where the lack of centralized guidance leads to stochastic fluctuations in care delivery.

Without a rigorous model-based approach to

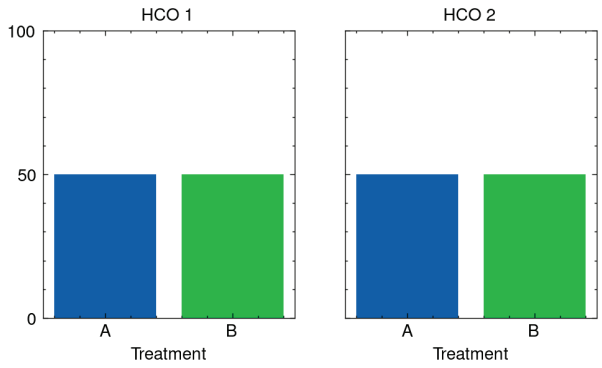


Figure 1: Prescribing distributions for two hypothetical HCOs ($N = 100$). In both instances, the marginal distribution of treatment choice is identical, with an even 50/50 split between Treatment A and Treatment B.

partition these sources of variation, there is a substantial risk of misclassifying an organization's actual degree of control. To illustrate this distinction, consider two hypothetical HCOs, each offering two treatments—A and B—and treating 100 patients. Suppose both exhibit the same aggregate distribution: 50 prescriptions of Treatment A and 50 of Treatment B (Figure 1). Now, assume that HCO 1 treats 50 male and 50 female patients and follows a strict protocol where all male patients receive Treatment A and all female patients receive Treatment B. In this case, the organization demonstrates absolute control through a deterministic rule; the observed heterogeneity is entirely explained by the interaction between the institutional mandate and the patient profile. Conversely, if HCO 2 treats the same patient mix but prescribes the treatments at random, the 50/50 split arises by chance, reflecting minimal control and high individual autonomy.

This example demonstrates that raw variability is a misleading metric for strategic analysis. Effective institutional analytics must move beyond a marginal view of aggregate prescribing share and toward a conditional view of organizational predictability. Decoupling observed heterogeneity from institutional influence enables identifying where treatment patterns are driven by individual provider preference and where they are dictated by the underlying organizational structure. This distinction is critical for commercial strategy, as it identifies which organizations possess the internal governance necessary to move from individual adoption to system-wide therapeutic integration.

2.2 Modeling Approach

The primary analytical unit of this framework is the therapy escalation event. A patient-centric approach is adopted, defining the event as the clinical transition from a first-line therapy to an advanced treatment. A critical temporal dimension of this transition is the time-to-event (TTE), which represents the duration between the initiation of first-line therapy and the date of the escalation event. This transition is assumed to represent a critical decision point where organizational influence is most likely to manifest through standardized treatment pathways or formulary mandates.

To formalize this, a hierarchical framework is considered where patient records $j \in \{1, \ldots, n_i\}$ are nested within HCOs $i \in \{1, \ldots, I\}$. Let $Y_{ij} \in \{1, \ldots, K\}$ denote the outcome variable indicating which of the K available treatments was prescribed to patient j in HCO i at the time of the escalation event. A Bayesian hierarchical multinomial logit model is employed to explicitly decompose the observed prescribing variance into its constituent parts while providing a principled mechanism for uncertainty quantification. This framework serves as an operating model for the derivation of an uncertainty-aware metric of institutional control for each organization, which is detailed in Section 2.3. This approach allows for the estimation of an interpretable control metric that accounts for the inherent uncertainty in organizational behavior, particularly across institutions with varying patient volumes. For HCO i and patient j, the treatment choice outcome is modeled as follows:

$$Y_{ij} \mid \boldsymbol{X}_{ij} \sim \text{Categorical}\left(p_{ij}^{(1)}, \ldots, p_{ij}^{(K)}\right),$$

$$\log\left(\frac{p_{ij}^{(k)}}{p_{ij}^{(K)}}\right) = \alpha^{(k)} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_i^{(k)} + \delta_i^{(k)},$$

$$
\begin{aligned}
\alpha^{(k)} &\sim \text{Normal}(0, \sigma_\alpha^2), \\
\boldsymbol{\beta}_i^{(k)} &\sim \text{Normal}(0, \boldsymbol{\Sigma}_\beta), \\
\delta_i^{(k)} &\sim \text{Normal}(0, \sigma_\delta^2), \\
\sigma_\alpha, \sigma_\delta &\sim \text{Half-Normal}(\sigma_0), \\
\boldsymbol{\Sigma}_\beta &\sim \text{Inverse-Wishart}(\boldsymbol{\Psi}, \nu), \\
i &= 1, \ldots, I, \quad k = 1, \ldots, K-1.
\end{aligned}
$$

In this specification, $p_{ij}^{(k)} = P(Y_{ij} = k \mid \boldsymbol{X}_{ij})$ denotes the probability that patient j at HCO i is prescribed treatment k. The feature vector $\boldsymbol{X}_{ij}$ comprises patient-level clinical and demographic attributes (e.g., age and gender), the TTE associated with the event, and various organization-specific features, including measures of institutional reach and payer-level dynamics. For notational simplicity, both patient- and organization-level features are consolidated into $\boldsymbol{X}_{ij}$; similarly, the coefficient vector $\boldsymbol{\beta}_i^{(k)}$ denotes both fixed and varying effects. The term $\alpha^{(k)}$ is a treatment-specific intercept, while $\delta_i^{(k)}$ captures the HCO-level random effect for treatment k, forming an $I \times (K-1)$ matrix that accounts for unobserved heterogeneity across HCO–treatment combinations. Treatment K is treated as the reference category for identifiability, and individual treatment probabilities are obtained by normalizing the logits:

$$
p_{ij}^{(k)} = \frac{\exp(\alpha^{(k)} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_i^{(k)} + \delta_i^{(k)})}{1 + \sum_{m=1}^{K-1} \exp(\alpha^{(m)} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_i^{(m)} + \delta_i^{(m)})},
$$

$$
p_{ij}^{(K)} = \frac{1}{1 + \sum_{m=1}^{K-1} \exp(\alpha^{(m)} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_i^{(m)} + \delta_i^{(m)})}.
$$

To perform Bayesian inference across the high-dimensional parameter space and ensure scalability across therapeutic areas with large-scale patient data, an optimization-based approximation via stochastic variational inference [5] with subsampling is adopted. Bayesian inference over the high-dimensional parameter space is conducted using an optimization-based approximation via stochastic variational inference [5] with subsampling to ensure scalability across therapeutic areas with large-scale patient data. This strategy is essential to accommodate the massive volume of escalation events while simultaneously estimating the thousands of organization-specific effects required to characterize institutional heterogeneity.

Specifically, the automatic differentiation variational inference [6] with a mean-field variational family is employed to approximate the posterior distribution. The posterior samples for all model parameters are obtained by maximizing the evidence lower bound, forming the basis for probabilistic uncertainty quantification of the proposed control metrics.

All experimentation and implementation are conducted using **aimz** [7], an internally developed Python package and probabilistic impact modeling framework, which provides scalable Bayesian inference and distributed predictive sampling capabilities designed for large-scale data.

2.3 A Metric for Institutional Control

The posterior samples obtained from the inference process can provide a high-resolution view of prescribing variability, but they require a structured summary to be actionable for institutional analysis. Motivated by the conceptualization of control as the degree of structured variability in prescribing behavior, a model-based metric is introduced, defined as the proportion of explained variance to total variance within each HCO. This approach allows for the estimation of an interpretable metric by adapting the R^2 framework proposed by [8] to a categorical response setting, where the model yields posterior distributions of predicted category probabilities.

For each patient j in HCO i and each posterior draw $s \in \{1, \ldots, S\}$, the predicted probability vector over the K treatment categories is denoted as:

$$
\boldsymbol{p}_{ij}^{(s)} = \left(p_{ij}^{(1)}, \ldots, p_{ij}^{(K)}\right)^{(s)}
$$

These vectors represent the model's estimate of treatment choice uncertainty and serve as the foundation for partitioning within-HCO variation. Specifically, the control metric for HCO i is computed for each posterior sample s, capturing

the proportion of variation in the model's predictions explained by the fitted structure relative to the total variance:

$$R_i^{2(s)} = \frac{\mathrm{Var}_{i,\mathrm{fit}}^{(s)}}{\mathrm{Var}_{i,\mathrm{fit}}^{(s)} + \mathrm{Var}_{i,\mathrm{res}}^{(s)}}$$

The components of this variance decomposition are defined as follows:

$$\begin{aligned}
&\mathrm{Var}_{i,\mathrm{fit}}^{(s)} = \mathrm{tr}\left(\widehat{\mathrm{Var}}\left(\boldsymbol{p}_{ij}^{(s)}\right)\right) \\
&= \frac{1}{n_i - 1}\sum_{j=1}^{n_i} \mathrm{tr}\left(\left(\boldsymbol{p}_{ij}^{(s)} - \bar{\boldsymbol{p}}_{i\cdot}^{(s)}\right)\left(\boldsymbol{p}_{ij}^{(s)} - \bar{\boldsymbol{p}}_{i\cdot}^{(s)}\right)^T\right) \\
&= \frac{1}{n_i - 1}\sum_{j=1}^{n_i} \left(\boldsymbol{p}_{ij}^{(s)} - \bar{\boldsymbol{p}}_{i\cdot}^{(s)}\right)^T\left(\boldsymbol{p}_{ij}^{(s)} - \bar{\boldsymbol{p}}_{i\cdot}^{(s)}\right),
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}_{i,\mathrm{res}}^{(s)} &= \mathrm{tr}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\widehat{\mathrm{Var}}\left(Y_{ij} \mid \boldsymbol{p}_{ij}^{(s)}\right)\right) \\
&= \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{p}_{ij}^{(s)^T}\left(1 - \boldsymbol{p}_{ij}^{(s)}\right).
\end{aligned}$$

The explained variance, $\mathrm{Var}_{i,\mathrm{fit}}^{(s)}$, measures how much the predicted probabilities vary across patients due to the learned model parameters, while the residual variance, $\mathrm{Var}_{i,\mathrm{res}}^{(s)}$, captures the remaining uncertainty in the categorical outcome. The trace operator is applied to summarize the covariance of the categorical distribution into a scalar value. While other generalized variance measures could be considered, the determinant is not applicable in this context due to rank deficiency in the residual variance. By construction, the R^2 value ranges from 0 to 1, where a value close to 1 indicates high control (i.e., most variation is captured by the model), and a value near 0 indicates low control where most variation remains unexplained.

To ground the interpretation of this metric, consider the example in Figure 1. Using one-hot encoding for treatment choice, each outcome in the two HCOs can be represented as a set of 50 vectors, denoted by $\boldsymbol{Z}_{ij}$, taking the form (1, 0) or (0, 1), corresponding to the prescription of treatments A and B, respectively:

$$\begin{aligned}
Y_{ij} = A \Rightarrow \boldsymbol{Z}_{ij} = (1,0) \quad &\text{if A is prescribed,} \\
Y_{ij} = B \Rightarrow \boldsymbol{Z}_{ij} = (0,1) \quad &\text{if B is prescribed.}
\end{aligned}$$

Now assume a perfect model that predicts the outcome probabilities $\boldsymbol{p}_{ij}^{(s)}$ for each patient j in HCO i without any error. In HCO 1, where prescribing is deterministic, the predicted probabilities are $\boldsymbol{p}_{ij}^{(s)} = \boldsymbol{Z}_{ij} = (1,0)$ for all males and $\boldsymbol{p}_{ij}^{(s)} = \boldsymbol{Z}_{ij} = (0,1)$ for all females. Since there is no variability in the predicted probabilities for each event, $\mathrm{Var}_{i,\mathrm{res}}^{(s)} = 0$, which leads to $R_i^{2(s)} = 1$, indicating complete control. In contrast, in HCO 2, where prescribing is essentially random, the predicted probabilities from this perfect model are $\boldsymbol{p}_{ij}^{(s)} = (0.5, 0.5)$ for all patients, reflecting an equal likelihood of either treatment being chosen. Since these predictions do not vary across patients, $\mathrm{Var}_{i,\mathrm{fit}}^{(s)} = 0$, and hence $R_i^{2(s)} = 0$, indicating no model-explained variability in prescribing behavior and a complete lack of control.

Note that this is an illustrative example; in practice, no HCO is expected to exhibit such an extreme level of control1. Rather, control levels will vary across a spectrum among organizations. Moreover, as real-world models do not achieve perfect predictive accuracy, it is essential to include relevant features and interpret the control metric as a relative measure when evaluating different institutions. The resulting distribution of R^2 values for each HCO captures the uncertainty associated with the control estimate, allowing for robust comparisons even when patient volumes n_i vary. It should be emphasized that this metric is not a traditional measure of overall model fit; rather, it serves as a specialized diagnostic of institutional governance, identifying where organizational mandates effectively supersede individual provider autonomy to drive systematic treatment patterns.

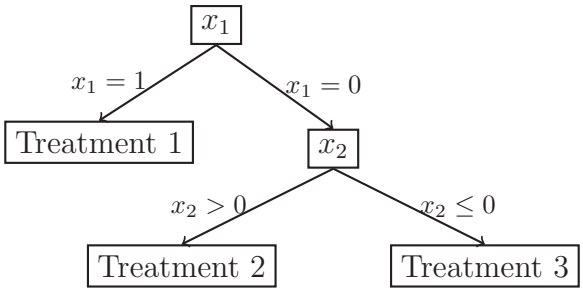


Figure 2: Deterministic logic tree for the clinical protocol used in the simulation study.

3 Results

3.1 Simulation Study

To validate the framework's ability to recover latent institutional governance, a simulation study is conducted involving 100 synthetic HCOs and a total of 100,000 patient records, with each HCO comprising approximately 1,000 patients. Each organization is assigned to one of three behavioral regimes (high, moderate, and low) representing the degree to which a centralized clinical protocol governs prescribing behavior. The protocol itself is designed as a decision tree, as illustrated in Figure 2, where the choice between three treatments is governed by two features: a binary indicator x_1 and a continuous variable x_2. A third feature, x_3, is generated as a distractor and is intentionally excluded from the protocol logic to test model discrimination.

Under the high control regime, representing 10% of the organizations, prescribing follows the deterministic logic with 90% adherence, leaving only a 10% margin for stochastic variation. The moderate regime, assigned to 70% of the HCOs, represents a more decentralized setting where the protocol is followed in only 50% of cases. Finally, the low control regime, comprising 20% of the sample, represents a setting where the structural protocol accounts for only 10% of decisions.

The simulation results confirm that the R^2 metric effectively recovers these latent structural signatures. As shown in Figure 3, the distributions of the R^2 values form three clearly separated peaks that reflect the varying simulated adherence levels. The high control group is

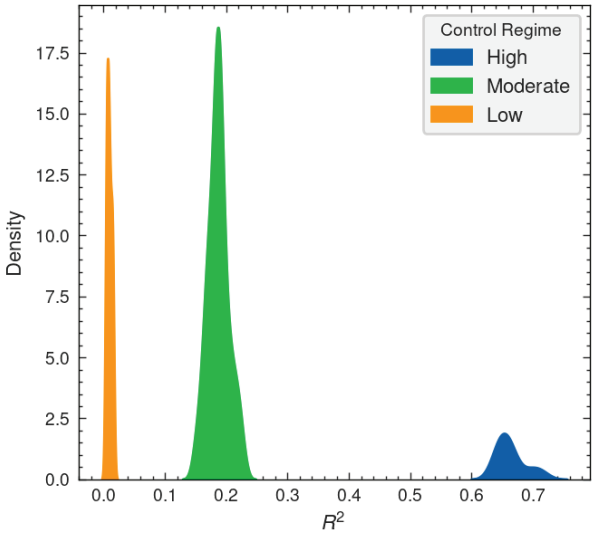


Figure 3: Density of posterior mean R^2 across simulated control regimes.

characterized by significantly elevated R^2 values, reflecting the model's success in capturing the systematic influence of x_1 and x_2. In contrast, the low control group exhibits R^2 values concentrated near zero, accurately identifying environments where prescribing decisions are decoupled from the observed features. By demonstrating a clear separation between these regimes, the simulation validates R^2 as a principled diagnostic tool for differentiating between systematic institutional mandates and idiosyncratic prescribing patterns.

3.2 Empirical Case Study

To evaluate the practical utility of the proposed framework, the methodology was applied to a large-scale dataset on a specific disease area. While the underlying model is therapeutic-agnostic, this application centers on a category characterized by significant shifts in clinical practice.

The construction of the analytical cohort followed a filtering process designed to isolate systematic institutional behaviors and ensure statistical robustness. First, the temporal scope was aligned with the current market landscape by including therapy progression events events over

a period chosen to reflect market dynamics. Second, the cohort was restricted to patients aged 18 years or older, focusing on the initial progression event to ensure the analysis reflects primary treatment decisions rather than subsequent switches. Finally, to ensure stable posterior estimates for the random effects, the analysis was restricted to large-scale, integrated healthcare systems with a high volume of escalation events during the study period. These procedures ensured comprehensive representation across the therapeutic landscape: every HCO in the final cohort managed events for all primary treatment categories as well as the reference group.

To maintain computational efficiency while preserving the complexity of the market, the analysis focused on the top 10 treatments by escalation volume. Any remaining low-volume or long-tail treatments were aggregated into an "Others" category, which serves as the reference group for the model.

The modeling process incorporated a multidimensional feature set designed to account for various clinical, demographic, and economic factors that might otherwise obscure the measured degree of institutional influence. Furthermore, a range of organizational features were incorporated to capture the impact of formulary constraints and existing institutional engagement.

The posterior distributions of the coefficients provide a high-resolution map of the decision-making drivers captured by the model. The SNR for a specific feature f and treatment k is defined as:

$$\text{SNR}_{f,k}^{(s)} = \frac{|\mu_{f,k}^{(s)}|}{\sigma_{f,k}^{(s)}},$$

where $\mu_{f,k}^{(s)}$ is the mean of the coefficients across the HCOs and $\sigma_{f,k}^{(s)}$ is the inter-institutional standard deviation for a given posterior sample s. As illustrated in Figure 4, some of the key features maintain SNRs above the 1.0 threshold. This indicates that for these treatments, the global clinical "Signal" outweighs the "Noise" of institutional variation. In contrast, TTE displays significantly lower SNR values across the majority of treatments, with almost all credible

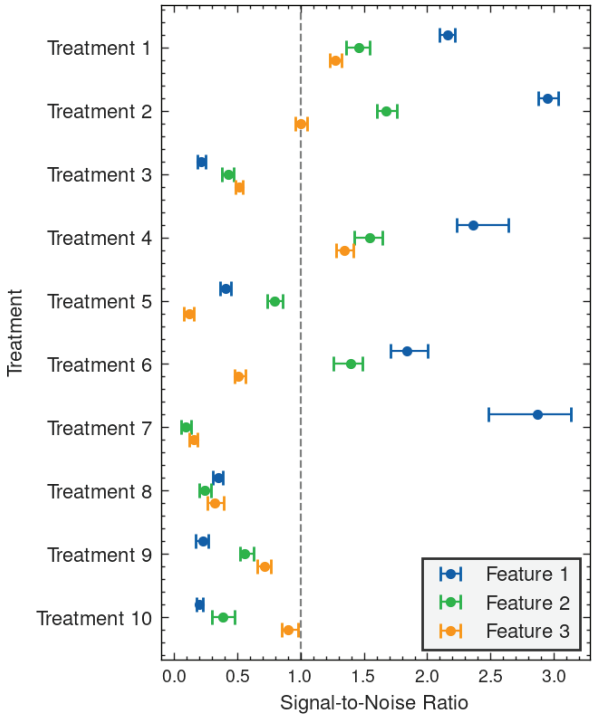


Figure 4: Signal-to-Noise Ratio of key features across the top 10 treatments. Points denote posterior mean SNR, and error bars indicate the 99% highest density interval of the SNR across the organizations.

intervals remaining below the 1.0 mark.

Figure 5 presents the results from the posterior predictive check for the model, comparing the observed distribution of treatment choices with the distribution predicted from posterior samples across the entire dataset. This indicates that the model closely reproduces the observed treatment distribution and validates the model's calibration, with only minor discrepancies for specific treatments. Although many of the estimated coefficients exhibit only modest individual effects, the model's overall fit appears to be largely driven by the random effects $\delta_i^{(k)}$, which account for unobserved heterogeneity across HCOs.

Following the validation of the model's calibration, the institutional control metric R^2 was derived for each organization. However, interpreting R^2 in isolation is challenging without a baseline measure of the raw prescribing diversity within each institution. To provide this context,

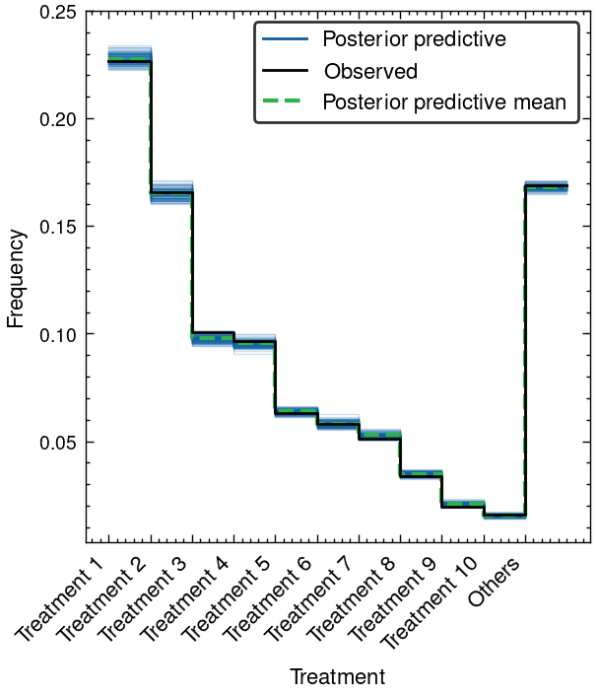


Figure 5: Posterior predictive check of treatment frequencies across the top 10 treatments and the reference category. The plot displays the observed counts for each treatment against the posterior predictive distribution of counts.

entropy is calculated for each HCO i as:

$$\text{entropy}_i = -\sum_{k=1}^{K} \hat{p}_i^{(k)} \log_2 \left(\hat{p}_i^{(k)} \right),$$

where $\hat{p}_i^{(k)}$ represents the observed proportion of patients prescribed treatment k within the organization. In this context, entropy serves as a model-independent measure of outcome variability. A value of zero denotes absolute uniformity where all patients receive the same treatment, while higher values reflect a more even distribution across available treatments, indicating greater heterogeneity in observed treatment patterns.

The resulting R^2 values represent the proportion of within-HCO variance explained by the model features relative to the total variance, providing the primary relative measure of institutional control. As summarized in Table 1, R^2 exhibits a weak negative correlation with both

Table 1: Correlation between institutional summary statistics.

Metric	HCO Size (n_i)	Entropy
R^2	-0.057	-0.163
Entropy	0.139	—

institutional size (-0.057) and entropy (-0.163). In contrast, entropy shows a positive correlation with HCO size (0.139). These relationships highlight the unique value of the R^2 metric: while entropy reflects the raw, unconditioned variability of observed treatment choices, R^2 isolates the structural influence that remains after accounting for patient profiles and unobserved heterogeneity. The low correlation between R^2 and HCO size—as well as its divergence from raw entropy—indicates that institutional control is not merely a byproduct of patient volume or treatment heterogeneity. Instead, it reflects distinct organizational behaviors and internal governance structures that drive systematic treatment patterns independently of scale.

The distribution of these R^2 values across the HCOs is illustrated in Figure 6. As shown, the metric exhibits a right-skewed distribution, with the majority of organizations concentrated in the lower range of the scale. Although the absolute magnitudes of R^2 are modest, their interpretation is most meaningful in a relative sense, serving as a comparative measure of institutional control across organizations. Importantly, each R^2 for an HCO is represented as a posterior distribution rather than a single point estimate. This distributional representation provides direct insight into the uncertainty surrounding each organization's estimated level of control, which is particularly informative for HCOs with lower patient volumes. This allows for a clear distinction between organizations that are confidently identified as high- or low-control environments and those whose position in the distribution may be driven by estimation uncertainty rather than systematic structural behavior.

By identifying organizations in the right tail of this distribution, the methodology isolates specific institutional environments where the

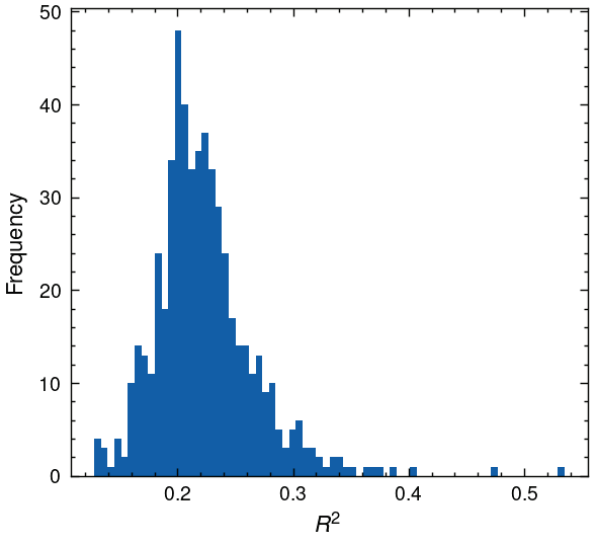


Figure 6: Distribution of the posterior mean R^2 across the HCOs.

model's structural features—and by extension, the organization's implicit or explicit mandates—most effectively override individual provider autonomy. This relative scaling provides a principled foundation for behavioral segmentation, allowing analytics teams to differentiate between HCOs where treatment pathways are systematic and those where decisions remain largely decentralized.

4 Conclusion

This research introduces a model-based framework for quantifying institutional control, moving beyond aggregate volume metrics to provide a principled measure of organizational predictability. By focusing on therapy escalation events and employing a Bayesian hierarchical approach, prescribing variance is partitioned to reveal the degree of influence that HCOs exert over treatment pathways. This methodology transforms an institutional black box into a transparent, uncertainty-aware metric that characterizes the internal governance of diverse healthcare systems.

An important modeling consideration is the assumption of independence of irrelevant alternatives. While this assumption facilitates tractable estimation, it may be violated in settings where treatment choices are highly substitutable. Such violations can bias model estimates and potentially attenuate R^2 values, thereby limiting the metric's ability to fully capture the underlying decision-making process. Future extensions that prioritize metric precision over interpretability could incorporate more flexible non-linear or black-box modeling architectures to relax these structural constraints.

A practical limitation concerns the sensitivity of the control metric to data richness. If critical drivers remain unobserved or poorly accounted for, the model may misattribute systematic behavior to noise, leading to an underestimation of control. This issue can be amplified in settings with low patient volumes or with a large number of available treatment choices. In such cases, the modeling framework may fail to recover the signal, rendering underlying institutional control statistically indistinguishable from randomness.

In addition, although this study focuses on the institutional layer, a simplifying assumption is the exclusion of explicit interactions between individual HCPs and their organizations. Investigating these multi-level dynamics represents a promising direction for future research, with the potential to further disentangle individual clinical autonomy from institutional mandate.

Furthermore, while this framework is currently grounded in discrete escalation events, future work could expand this to a more holistic patient journey perspective. Incorporating longitudinal treatment sequences and the transition between different therapy lines, rather than focusing on isolated decision points, would offer a more comprehensive view of how institutional control evolves and persists over the duration of a patient's care.

Despite these technical considerations, the strategic implications for pharmaceutical analytics and commercial resource allocation are significant:

- Optimized Commercial Strategy: By identifying high-control environments, organizations can shift from traditional "bottom-up" outreach to "top-down" strategies that align

with centralized decision-making protocols.

- Behavioral Segmentation: This framework enables a shift toward behavior-based segmentation, allowing for a more sophisticated understanding of market dynamics than simple geographic or volume-based metrics.

- Scalable Strategic Benchmarking: The extensibility of the proposed methodology ensures it can be applied across various therapeutic areas to benchmark institutional influence objectively.

Ultimately, the ability to quantify organizational influence over clinical decision-making provides a robust tool for enhancing strategic resource allocation, ensuring that commercial engagement is tailored to the actual structure of decision-making in an increasingly consolidated healthcare landscape.

References

[1] Carol K. Kane. Updated data on physician practice arrangements: Inching toward hospital ownership. *Chicago, IL: American Medical Association*, 2015.

[2] Carol K. Kane. Physician practice characteristics in 2024: Private practices account for less than half of physicians in most specialties. *Chicago, IL: American Medical Association*, 2025.

[3] Karen E. Lutfey, Stephen M. Campbell, Lisa D. Marceau, Martin O. Roland, and John B. McKinlay. Influences of organizational features of healthcare settings on clinical decision making: Qualitative results from a cross-national factorial experiment. *Health*, 16(1):40–56, 2012.

[4] Jon M. Richards, Trever B. Burgon, Diana Tamondong-Lachica, Jacob D. Bitran, Wilfredo L. Liangco, David R. Paculdo, and John W. Peabody. Reducing unwarranted oncology care variation across a clinically integrated network: A collaborative physician engagement strategy. *Journal of Oncology Practice*, 15(12):e1076–e1084, 2019.

[5] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[6] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

[7] Eunseop Kim. aimz: Scalable probabilistic impact modeling. `https://github.com/markean/aimz`, 2025.

[8] Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari and. R-squared for Bayesian regression models. *The American Statistician*, 73(3):307–309, 2019.

Large Language Models in Pharmaceutical Quality Management Systems: A Quality-First Implementation Framework

Chris Dayton, Chief Executive Officer, Quality Assured AI; Adam Pinkert, Sr. Advisor, Quality and Regulatory, Quality Assured AI, Principal (Trayenco Consulting); Nick Moench, Principal Machine Learning Architect, Quality Assured AI

Abstract: The integration of Large Language Models (LLMs) into pharmaceutical Quality Management Systems presents significant opportunity alongside substantial regulatory complexity. Traditional AI deployment approaches — optimizing general-purpose models for broad performance, then retrofitting regulatory controls — frequently fail to satisfy GxP requirements for deterministic audit trails, validated change control, and ALCOA+ data integrity principles.

This paper presents a quality-first implementation framework establishing GxP requirements as foundational design criteria from project inception. Five principles underpin credible LLM deployment: controlled system boundaries favoring closed architectures; retrieval-augmented generation constraining outputs to approved sources; robust human-in-the-loop oversight by qualified personnel; lifecycle control through locked-state model management; and comprehensive data governance addressing traceability and ALCOA+ compliance.

Case study evidence from a GMP facility remediation program demonstrates measurable outcomes including 30% reduction in deviation cycle time and elimination of a critical deviation backlog. A significant finding emerged from comparative performance analysis: AI-assisted novice investigators produced deviation narratives requiring less substantive revision than experienced writers working without AI assistance — suggesting well-designed AI systems can encode and transfer expert knowledge, addressing longstanding workforce development challenges.

When systems are properly architected, the majority of outputs require only minor refinement, with human reviewers providing genuine correction rather than rubber-stamp approval, affirming that optimal pharmaceutical AI strategy emphasizes human-AI collaboration over autonomous operation.

I. INTRODUCTION

The pharmaceutical industry faces a fundamental challenge when implementing artificial intelligence within GxP-regulated environments. Traditional approaches to AI deployment typically begin with general-purpose models optimized for broad performance, then attempt to retrofit regulatory controls after system design is complete. This adaptation strategy frequently encounters obstacles when confronting pharmaceutical quality requirements: deterministic audit trails, validated change control, explainable decisions, and ALCOA+ data integrity principles.

This paper examines an alternative approach that inverts this sequence. By establishing GxP requirements and regulatory constraints as foundational design criteria from project inception, system architecture and model selection decisions inherently support rather

than conflict with regulatory expectations. This quality-first methodology produces fundamentally different technical architectures than post-hoc adaptation approaches.

Drawing on practical implementation experience from quality-led AI development programs, this paper presents system-level considerations, architectural patterns, and risk-mitigation strategies for LLM deployment in pharmaceutical QMS applications, including deviation investigations, knowledge management, and quality decision-support. Real-world case examples demonstrate how architectural choices influence validation strategy, audit readiness, and quality-risk mitigation, illustrating how LLM systems can enhance efficiency while preserving GxP compliance.

II. REGULATORY FRAMEWORK AND COMPLIANCE REQUIREMENTS

Current Regulatory Landscape

In January 2026, the FDA and EMA issued joint guidance establishing ten core principles for AI in pharmaceutical applications.[1] These principles emphasize human-centric design, risk-based approaches, adherence to GxP standards, clear context of use, explainability, robust data governance, sound model design, performance assessment, lifecycle management, and user AI literacy. This guidance synthesizes existing pharmaceutical quality principles rather than creating entirely new requirements.

Key applicable regulations include FDA 21 CFR Parts 210/211 (cGMP)[2], 21 CFR Part 11 (electronic records),[3] EU EudraLex Volume 4 with Annex 11 (computerized systems)[4] and draft Annex 22 (AI use),[5] ICH Q9 (quality risk management),[6] and the EU AI Act[7] establishing risk-tiered frameworks.

Data Integrity and Validation Requirements

ALCOA+ principles (Attributable, Legible, Contemporaneous, Original, Accurate, Complete, Consistent, Enduring, Available) form critical constraints for LLM implementation.[8,9] For LLM systems, these translate to specific controls:

Attributability requires user authentication, session management, and audit trails logging, not merely outputs but complete interaction chains including prompts, model responses, RAG sources, and human reviews.

Contemporaneous recording demands real-time log generation with timestamp precision sufficient for regulatory audit.

Accuracy for probabilistic outputs focuses on constraining outputs to approved sources through RAG, communicating uncertainty appropriately, requiring SME validation, and monitoring for systematic inaccuracy patterns.

Completeness requires preserving model version, system prompt, user query, retrieved sources with versions, generated response, human modifications, and approval rationale.

Validation Paradigms for Probabilistic Systems

GAMP 5 classification places most LLM-based QMS applications in Category 5 (custom software), requiring risk-based validation.[10] The FDA's Computer Software Assurance (CSA) guidance advocates critical thinking approaches focusing validation on functions directly impacting product quality, patient safety, or data integrity.[11]

Traditional validation protocols encounter challenges with LLMs:

- **Output variability:** Probabilistic responses vary even with locked weights, making exact output matching impossible
- **Scope limits:** Infinite possible inputs require risk-based test scenario selection rather than exhaustive testing
- **Model opacity:** Billions of parameters make traditional code review impractical, requiring behavioral testing
- **Continuous learning incompatibility:** Some AI systems auto-update, conflicting with validated state concepts requiring change control

These challenges necessitate adapted validation approaches focusing on output quality metrics (accuracy, relevance, completeness), risk-based test coverage, and ongoing performance monitoring rather than deterministic verification.

III. ARCHITECTURAL PATTERNS FOR GXP COMPLIANCE

System Boundary Control: Open vs. Closed Architectures

The distinction between open- and closed-architecture systems fundamentally determines the feasibility of validation and data protection. Closed-architecture systems maintain exclusive organizational control over model hosting, data processing, and system access within validated boundaries. Open-architecture systems utilize external APIs or services where data transits outside organizational infrastructure.

For GxP applications involving confidential data or direct quality decisions, **closed architectures are strongly preferred**, providing:

- Complete control over model versions and configurations
- Full audit trails within organizational governance
- Enhanced data protection without third-party transmission risks
- Simplified validation without vendor dependency variables

Where open systems are necessary due to resource constraints, compensating controls include data encryption, contractual safeguards, vendor qualification audits, and enhanced monitoring.

Retrieval-Augmented Generation (RAG)

RAG architecture addresses the core challenge of constraining LLM outputs to approved organizational knowledge while maintaining traceability.[12] The RAG process involves:

1. **Knowledge base creation**: Approved documents (SOPs, regulations, batch records) are segmented into discrete units (commonly referred to as ‘chunks’)”, converted to vector embeddings capturing semantic meaning, and stored with metadata (version, effective date, approval status)
2. **Semantic retrieval**: User queries are embedded and vector similarity search identifies relevant knowledge chunks (typically 3-10 per query)

3. **Context augmentation:** Retrieved chunks are inserted into the LLM prompt with instructions to ground responses in provided context and cite sources
4. **Response generation:** The LLM synthesizes information from retrieved context with required citations
5. **Traceability logging:** Complete transaction logs include query, retrieved sources with relevance scores, response, and human review actions

For pharmaceutical applications, RAG provides critical advantages:

Source control limits information to validated, version-controlled documents, preventing generation based on uncontrolled internet sources or outdated training data.

Traceability enables each response to be traced to specific source documents with version identifiers, creating audit trails satisfying regulatory requirements.

Update management allows knowledge base updates by replacing documents without model retraining, supporting standard change control processes.

Validation feasibility focuses testing on retrieval accuracy and synthesis quality rather than validating entire model knowledge.

Implementation considerations include chunk size optimization, domain-specific embedding models for specialized terminology, retrieval threshold tuning (balancing context sufficiency with relevance), and access control integration ensuring users retrieve only authorized documents.

Human-in-the-Loop (HITL) Frameworks

The FDA-EMA guidance emphasizes "appropriate human oversight" commensurate with risk.[1] Effective HITL implementation requires:

Qualified reviewers: Personnel with formal training in relevant processes, demonstrated expertise, current AI system training, and organizational authority to approve quality decisions. Deviation narrative tools require reviewers qualified in deviation investigation; SOP review tools require QA document control expertise.

Structured review protocols: Checklists and rubrics providing consistent evaluation criteria. For deviation narratives: timeline completeness, evidence support, logical root cause, and appropriate CAPA. This prevents superficial "rubber-stamp" reviews.

Override mechanisms: Procedures for approval, modification, or rejection with documented rationale. Systems capture whether outputs were approved as-generated, modified (with specific changes), or rejected, plus reason codes and reviewer identity.

Escalation pathways: Automated flagging based on low confidence scores, edge case detection, high-consequence determinations, or unusual outputs, routing to senior reviewers or requiring multi-person review.

Performance feedback: Systematic aggregation of override rates, modification types, and trends. Baseline override rates established during validation trigger investigation when exceeded, providing objective performance monitoring.

Data Protection and Lifecycle Control

Data governance must address multiple data types with distinct requirements:

- **Training data** (fine-tuned models): Highest quality standards with full validation, ALCOA+ compliance, SME curation, and bias assessment
- **RAG/embedding data**: Version control, approval workflows, access restrictions, and metadata tracking
- **Inference data**: Comprehensive audit trails capturing user identity, timestamps, inputs, outputs, retrieved sources, and reviews

Security architecture includes encryption (TLS 1.2+, AES-256), multi-factor authentication, role-based access control (RBAC), network segmentation, and model file integrity verification via checksums.

Lifecycle strategies for GxP environments favor **locked-state models** with frozen weights and version control, updated only through formal change control. This provides:

- Clear validated state amenable to IQ/OQ/PQ protocols
- Predictable performance characteristics
- Straightforward revalidation criteria
- Simplified audit trails and inspection defense

Periodic retraining (e.g., quarterly/annually) balances maintaining validated state with incorporating operational learnings. Continuous unsupervised learning approaches are **not recommended** for GxP applications given incompatibility with validated state concepts and current regulatory frameworks.

IV. VALIDATION STRATEGY AND RISK-BASED TESTING

Three-Tier Validation Framework

The following three-tier framework was developed empirically from validation planning experience across the GxP LLM deployments described in this paper — including the GMP facility deviation management system and the SOP gap analysis tool — and is presented here as a practical organizing structure rather than a formally codified standard. It draws on GAMP 5 category concepts[10] and CSA risk-based principles[11] but represents the authors' synthesis of those frameworks applied specifically to LLM systems. The framework establishes three levels based on autonomy and design control:

Level A (Parallel AI CS): High autonomy, low control. Extensive validation equivalent to full IQ/OQ/PQ. Generally **inappropriate** for GxP applications (e.g., autonomous batch release decisions).

Level B (Classical Non-AI CS): Low autonomy, high control. Traditional validation for rule-based deterministic systems.

Level C (Piecewise Locked-State AI): Moderate autonomy, high control. **Recommended for most GxP LLM applications**. Features locked weights, controlled prompts, defined boundaries, and HITL verification. Risk-based testing focused on context of use.

Computer Software Assurance (CSA) Principles

CSA emphasizes "testing what matters"—focusing validation effort on high-impact functions while applying lighter oversight to low-risk components:[10]

High-risk functions (GxP-decisive): 50-100+ test scenarios, formal protocols, traceability matrices, full regression testing

Medium-risk functions (GxP-assistive): 20-50 test cases, documented procedures, SME evaluation, targeted regression

Low-risk functions (informational): 10-20 exploratory tests, user acceptance criteria, periodic spot-checking

Test Case Development and Acceptance Criteria

Effective test cases address functional coverage (normal scenarios, boundary conditions, edge cases, error conditions), domain coverage (representative product types, process areas, root cause categories), and output quality dimensions (factual accuracy, completeness, relevance, citation quality, logical coherence, absence of hallucinations).

Example acceptance criteria for deviation narrative generation:

- ≥90% require only minor edits (grammar/formatting)
- 100% factual accuracy (no hallucinations)
- ≥95% proper citation of regulatory requirements or cite procedures
- Average relevance score ≥4.0 on 5-point Likert scale as ranked by SMEs

For risk classification:

- Overall accuracy ≥85%
- High-risk sensitivity ≥95% (miss no more than 5% of truly high-risk deviations)
- High-risk precision ≥70%
- Operational override rate ≤25%
- AI determination of Criticality as Lower than SME classification 0%

The last of these bullet points is incredibly important when AI is used to classify impact or criticality, as the FDA requires that risk assessments are gauged between Influence and Decision Consequence. A model should never classify criticality, create impact assessments, at a lower classification than an SME. Doing so can lead the model to create a justification for no, or reduced, impact to a parameter that may negatively affect a SISPQ (Safety, Identity, Strength, Purity, Quality) attribute of a product. When used in such cases, the AI model/system's initial assessment should be more risk-averse than a human reviewer and require the human to provide information to justify a reduction in risk classification.

V. CASE EXAMPLES FROM PRACTICE

Case 1: AI-Driven Deviation Management and Site Remediation

Context and Business Challenge:

A multi-product GMP manufacturing facility faced critical operational and compliance challenges requiring urgent remediation. The site had developed a significant deviation backlog causing production delays, with classification and investigation processes slowed by resource constraints and workflow inefficiencies. Compounding these issues, the facility was preparing for FDA re-inspection following prior audit findings that identified procedural gaps due to missing site-level SOPs. An ongoing transition from Quality Document System A (QDS A) to Quality Document System B (QDS B) created additional documentation challenges, as legacy document references no longer aligned with new document numbering schemes.

The combination of deviation backlog, missing SOPs, outdated document cross-references, and pending regulatory inspection created substantial operational and compliance risk requiring rapid, systematic intervention.

System Design and Architecture:

- Classification: Multiple use cases ranging from advisory/assistive (deviation drafting) to decision-support (impact assessment), low to medium GxP impact.
- Architecture: Secure, air-gapped closed system deployed entirely within client infrastructure with no external API dependencies. This architecture addressed client concerns regarding confidential manufacturing data, product formulations, and investigation details requiring strict access control
- Deployment model: On-premises infrastructure with organizational data governance, eliminating data transmission risks inherent in cloud-based or vendor API services
- Multiple AI models deployed:
  - » Deviation narrative generation model: RAG-based system with knowledge base comprising site SOPs, regulatory guidance (FDA, ICH Q9), and historical high-quality deviation narratives (redacted of confidential information)
  - » Custom impact assessment model: Purpose-built to classify deviations by impact and urgency to prioritize resource allocation and bring timelines into SOP compliance
  - » SOP generation utility model: Leveraged global network SOPs from client's corporate library to draft missing site-level procedures, cross-referencing for conflicts, overlaps, and gaps against internal documents and industry guidelines
  - » Document reference update automation: Custom software to systematically update document cross-references following QDS transition

Human-in-the-Loop Controls:

All AI-generated outputs subject to qualified personnel review before use in quality records:

- Deviation narratives: Reviewed by deviation investigators with minimum qualifications (QA experience, training on organizational procedures) using structured evaluation criteria addressing timeline accuracy, investigation completeness, root cause logic, CAPA appropriateness, and factual accuracy

- Impact assessments: Reviewed by Tech-ops and Manufacturing staff, who provided data (ie. QC test results) to support reduction in product impact.
- SOP drafts: Reviewed by quality document control personnel and technical SMEs, with formal approval through established SOP approval workflows including quality unit review and management authorization
- Document reference updates: Sampling-based verification by document control to confirm reference accuracy

Operational Results (2-month intensive deployment):

The implementation achieved substantial measurable improvements across multiple quality system dimensions:

Deviation Timeline Performance:

- Overall deviation cycle time: 30% reduction, bringing site performance back within SOP-mandated timelines
- Investigation and drafting phase: 28% reduction (from 15 days to 11 days average)
- Deviation closures: 160 deviations closed in 2-month period, significantly reducing backlog and production holds

SOP Development Efficiency:

- Drafting time: 2-3 hours per SOP (compared to estimated 30-50 hours for manual drafting from templates)
- Quality improvement: All generated SOPs automatically cross-referenced for conflicts, overlaps, and gaps both internally across site documents and externally against FDA guidance, ICH guidelines, and industry standards
- Inspection readiness: All missing site-level SOPs identified in prior FDA inspection successfully drafted, reviewed, approved, and implemented within accelerated timeline

Workforce Augmentation:

- Performance leveling: Novice deviation writers using AI-assisted tools demonstrated output quality and efficiency exceeding experienced writers working without AI assistance
- Training acceleration: New investigators achieved acceptable performance levels more rapidly through structured AI interaction and example-based learning
- Resource optimization: Existing staff handled increased throughput without proportional headcount increases

Audit Readiness and Regulatory Defense:

Site achieved inspection-ready status ahead of anticipated FDA re-inspection timeline. Key elements supporting regulatory defensibility included:

- Closed-system architecture: Air-gapped deployment with no external data transmission addressed inspector concerns regarding data integrity and confidentiality
- Comprehensive audit trails: Complete documentation of AI-generated content, human review actions, modification tracking, and approval workflows satisfying electronic record requirements
- HITL controls: Demonstrated qualified personnel oversight with documented training, structured review procedures, and override mechanisms
- System documentation: Architecture

specifications, data flow diagrams, access controls, and change management procedures documented to GxP standards

Comparative Performance Analysis:

The most consequential finding from this deployment was not operational efficiency — it was a fundamental reversal of expected performance hierarchies between experienced and novice quality professionals:

- AI-assisted novice writers produced deviation narratives requiring less substantive revision than experienced writers without AI across multiple quality dimensions including timeline completeness, evidence citation, root cause logical support, and CAPA specificity
- This performance inversion suggests AI systems effectively encode and transfer expert knowledge patterns, enabling less-experienced personnel to produce higher-quality outputs through structured guidance and example-based learning

AI as a Knowledge Transfer and Workforce Development Tool:

The performance inversion observed in this deployment represents a finding with implications extending well beyond operational efficiency.

Pharmaceutical quality systems face a persistent structural challenge: critical investigation and documentation expertise is concentrated in senior personnel, creates organizational vulnerability when those individuals leave, and takes years to develop in new investigators through traditional mentorship and experience accumulation. This knowledge retention problem is widely recognized across the industry but has proven difficult to systematically address.

The results presented here suggest that well-architected AI systems may offer a practical solution. By encoding expert knowledge patterns into RAG knowledge bases populated with high-quality historical examples, organizations can make expert-level guidance available to every investigator at the point of need — effectively democratizing access to institutional expertise regardless of individual experience level.

This has direct implications for workforce development programs, succession planning, and quality system resilience. Organizations need not wait years for investigators to develop expertise organically; structured AI interaction may accelerate competency development while simultaneously reducing the quality variability that experience-level differences traditionally introduce into deviation investigations and quality records.

Implementation Challenges and Mitigations:

Several challenges emerged during rapid deployment requiring adaptive responses:

Technical challenge:

- Knowledge base curation: Site SOPs required preprocessing to extract structured information, particularly complex tables and flowcharts. Mitigation: Applied document parsing utilities and manual verification of critical sections

Organizational challenge:

- User adoption resistance: Some experienced investigators initially skeptical of AI assistance value. Mitigation: Demonstrated time savings through pilot testing, emphasized AI as augmentation not replacement, collected user feedback for system refinement

Quality challenge:

- Hallucination risk: Initial testing revealed occasional generation of unsupported technical details. Mitigation: Enhanced RAG retrieval thresholds, explicit uncertainty acknowledgment instructions, mandatory citation requirements, intensified human review training

Lessons Learned and Generalizable Insights:

This implementation provides several insights transferable to other pharmaceutical quality AI deployments:

1. Crisis scenarios accelerate adoption: The combination of deviation backlog, pending inspection, and resource constraints created organizational urgency that facilitated rapid decision-making and user acceptance—circumstances that paradoxically enabled faster deployment than greenfield projects lacking urgency
2. Closed-system architecture is feasible: Despite common assumptions that on-premises AI deployment requires prohibitive infrastructure investment, air-gapped systems can be established within typical pharmaceutical IT environments using available hardware and open-source models
3. Multiple use cases compound value: Deploying AI across related quality processes (deviation management, SOP development, document cross-referencing) creates synergistic benefits exceeding sum of individual applications, justifying initial investment and change management effort
4. The AI-Model + HITL outperforms BOTH Human AND Autonomous AI systems: Maintaining strong HITL controls with qualified review preserved compliance while enabling aggressive timeline compression—suggesting optimal pharmaceutical AI strategy emphasizes human-AI collaboration rather than autonomous AI operation or solely brute force human-only operation.
5. Performance monitoring enables continuous improvement: Systematic collection of override data, modification patterns, and user feedback supported iterative refinement of AI models and prompts, with measurable quality improvements over initial 8-week period
6. AI augmentation addresses knowledge retention challenges, accelerate training programs, and reduce quality variability introduced by experience level differences

Case 2: SOP Gap Analysis System

Context: Pharmaceutical quality document control requires systematic verification that SOPs address all applicable regulatory requirements — a process traditionally performed through manual review by experienced QA personnel. As regulatory frameworks grow in complexity and document libraries expand, the burden of comprehensive gap analysis scales accordingly, creating resource constraints that can delay document approval cycles and leave compliance gaps undetected.

System Design and Key Features:

A RAG-based gap analysis tool was developed and deployed as a web application, with a knowledge base comprising FDA regulations, ICH guidelines, and industry standards. The system analyzes submitted SOP drafts against this regulatory corpus, identifying potential gaps, inconsistencies, and areas requiring additional specificity. All flagged items require QA document control review before any action is taken, classifying this as a GxP-assistive rather than GxP-decisive application.

The system was designed specifically for SOP document types, with the intake pipeline incorporating a vision reasoning model capable of accurately processing complex document elements including tables and flowcharts, eliminating the preprocessing burden that typically constrains document ingestion in RAG-based systems.

Performance Results:

Analytical performance demonstrated a 90% detection rate for critical regulatory gaps. Importantly, the system's findings were consistently at or above the criticality level that a qualified human reviewer would assign — the system never understated the significance of a finding. This directional conservatism is a deliberate design outcome reflecting the quality-first development philosophy described in Section IV: in GxP applications, the cost of understating a compliance gap far exceeds the cost of flagging an item for human review that ultimately requires no action. The result is a system that functions as a risk-averse first reviewer, ensuring that no critical gap is minimized before reaching qualified human assessment, while the HITL review layer provides the judgment needed to contextualize and appropriately weight each finding.

Operational Impact — Review Time Reduction:

Client feedback from operational use reported a substantial reduction in time required for regulatory gap review. Manual regulatory gap review by qualified QA consultants typically requires approximately 3 hours per document — a figure that increases with document complexity and scales with corpus size, making large-scale gap analysis programs resource-intensive (based on professional consulting time estimates from the authors' practice for manual regulatory gap review engagements). Reviewers using the AI gap analysis tool reported HITL review times of approximately 5 to 20 minutes per document depending on the number and complexity of findings, with an average of approximately 15 minutes for a document with 10 flagged items.

The AI system does not exhibit the same time scaling with document complexity that manual review does. A corpus of 47 documents was processed in approximately 40 minutes — a throughput that would be impractical to replicate through manual review without significant resource allocation.

This throughput advantage addresses a structural limitation of traditional regulatory review practice. In conventional audit and inspection preparation, resource constraints typically necessitate spot-checking — reviewing a representative sample of SOPs rather than the complete document library. This means compliance gaps in unreviewed documents can persist undetected until an inspection finding surfaces them. The AI gap analysis system's ability to process an entire SOP corpus in a single pass fundamentally changes this dynamic, enabling comprehensive coverage rather than sampling-based coverage. Organizations can verify regulatory alignment across their complete document library on a routine basis rather than relying on periodic spot checks, meaningfully reducing the risk of undetected gaps surviving into an inspection.

False Positive Analysis:

The reported false positive rate is largely attributable to documents uploaded outside the system's designed scope — specifically, users uploading FMEAs alongside SOPs when the system was purpose-built for SOP analysis. When the system is used as designed with SOP documents as input, the effective false positive rate for substantive analytical findings is lower than the aggregate figure reported. This underscores the importance of user onboarding and scope adherence rather than reflecting any analytical limitation of the system itself.

VI. RISK MITIGATION AND PRACTICAL IMPLEMENTATION

LLM systems in pharmaceutical QMS applications exhibit characteristic failure patterns requiring specific mitigation strategies. The following approaches were implemented based on operational experience:

Hallucinations:

LLMs can generate text that appears credible but contains factually incorrect information, fabricated citations, or unsupported claims. This was mitigated through several mechanisms: RAG architecture retrieving exclusively from validated company SOPs rather than relying on model's general training knowledge; system prompt design that instructs the model to request additional information rather than generating responses when user input is insufficient; and parameter tuning limiting output token length to reduce "rambling" or unsupported elaboration beyond retrieved context.

Performance degradation over time:

Model performance can decline due to various factors including knowledge base staleness or emergent failure patterns. This was addressed through: daily SOP updates pushed to the RAG knowledge base (later adjusted to weekly cadence), ensuring models always accessed current approved procedures; locked model weights preventing uncontrolled drift; user feedback mechanisms including thumbs up/down buttons on generated outputs; systematic report generation and review of thumbs-down responses; and iterative model refinement to address identified issues revealed through user feedback analysis and internal model audit trails.

Output variability:

The probabilistic nature of LLM text generation can produce inconsistent outputs for similar inputs. This was controlled through use-case-specific parameter tuning, recognizing that different applications require different generation characteristics. Temperature and other sampling parameters were independently optimized for each deployed model (e.g., impact assessment models configured differently than QMS authorship models or SOP-related applications), allowing precise control over output determinism appropriate to each use case's requirements.

Context window limitations:

When relevant information exceeds what can be effectively retrieved and processed, response quality degrades. This was mitigated through comprehensive RAG system tuning addressing multiple dimensions: chunk size optimization determining how source documents are segmented; top-K value tuning controlling how many chunks are retrieved per query; K-nearest neighbor (KNN) configuration with differentiated embedding temperature parameters applied to retrieved chunks based on relevance ranking; and cosine similarity threshold tuning to balance retrieving sufficient context against including marginally relevant content (e.g., retrieving highly relevant chunks with stricter similarity requirements alongside additional chunks with relaxed thresholds).

Systematic performance variability across use cases:

A single general-purpose configuration cannot adequately serve diverse pharmaceutical quality applications. This was addressed through: tightly constrained system prompts that prevent use outside defined Context of Use (COU), reducing risk of applying models to scenarios they were not validated for; and development of custom models tailored to specific departmental needs and use case requirements rather than deploying one-size-fits-all solutions, allowing optimization of architecture, parameters, and knowledge bases for distinct quality functions.

Implementation Roadmap

Organizations should follow a phased approach:

1. Pilot selection: Choose low-risk, high-value use case (advisory tools, informational queries) to build capability before tackling GxP-decisive applications
2. Architecture decisions: Establish closed system infrastructure, select appropriate base models, design RAG knowledge base, define HITL workflows
3. Validation planning: Develop context of use document, model explainability, perform risk assessment, define acceptance criteria, create test scenarios
4. Controlled deployment: Validate system, train users, implement monitoring, collect override data
5. Expansion: Apply learnings to additional use cases, scale infrastructure, develop organizational AI expertise

Critical success factors include executive sponsorship, cross-functional teams (QA, IT, process SMEs), adequate resource allocation (typically 3-6 months for first system), and realistic expectations about performance — when properly built, the majority of AI-generated outputs should require only minor refinement, but human reviewers must remain genuinely engaged rather than treating review as a formality.

VII. CONCLUSION

The integration of Large Language Models into pharmaceutical Quality Management Systems is achievable when GxP principles and regulatory expectations are established as foundational design criteria rather than post-hoc adaptations. The quality-first development approach produces fundamentally different system architectures than retrofitting compliance onto general-purpose AI.

Five generalizable principles support credible LLM deployment: (1) controlled system boundaries (closed architectures for GxP applications); (2) retrieval-augmented structures constraining outputs to approved sources; (3) robust human-in-the-loop oversight by qualified personnel; (4) lifecycle control strategies enabling validation and change management (locked-state models); and (5) comprehensive data protection addressing privacy, security, and traceability.

The case studies presented demonstrate that AI in pharmaceutical quality is not theoretical but operational reality, with validated systems functioning in GMP environments today. These implementations are not perfect, human reviewers will occasionally modify or reject AI-generated outputs, and robust HITL controls must exist to capture these corrections, but when systems are properly architected, the majority of outputs require only minor refinement rather than substantive rewriting, representing meaningful progress toward augmenting quality professionals with tools enhancing consistency, speed, and analytical depth while preserving human judgment and accountability.

As regulatory guidance continues evolving and technology advances, several trends will shape LLM integration: continued regulatory clarity through agency experience with AI inspections, standardization of validation approaches via industry working groups, technological advances improving model transparency, and growing organizational expertise in AI quality assurance.

Organizations that invest in quality-first AI development now; building technical capability, regulatory expertise, and validated operational experience, will be better prepared to adapt as guidance matures and industry best practices consolidate around validated implementation standards. Near-term regulatory developments will continue shaping implementation expectations. The FDA-EMA joint guidance issued in January 2026[1] establishes foundational principles, and the ongoing EudraLex Annex 22 consultation process[5] is expected to produce more prescriptive requirements for AI in GMP manufacturing once finalized. As FDA inspectors encounter AI systems with increasing frequency, inspection observations and enforcement actions will progressively clarify agency priorities in ways that formal guidance alone cannot. Organizations with operational validated systems today will be better positioned to contribute practical experience to these developing standards, and to adapt quickly as expectations become more defined.

The path forward requires balancing innovation with patient safety, maintaining rigorous quality standards while embracing efficiency gains, and collaborating across industry to develop shared validation frameworks and regulatory expectations. The principles and examples in this paper provide practical guidance for organizations beginning this journey.

REFERENCES

1. U.S. Food and Drug Administration & European Medicines Agency. Joint Guidance on Artificial Intelligence in Pharmaceutical Development and Manufacturing. 2026. Available from: https://www.fda.gov/media/189581/download and https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf

2. Code of Federal Regulations, Title 21, Parts 210-211: Current Good Manufacturing Practice. Washington DC: U.S. Government Publishing Office.

3. Code of Federal Regulations, Title 21, Part 11: Electronic Records; Electronic Signatures. Washington DC: U.S. Government Publishing Office.

4. European Medicines Agency. EudraLex Volume 4, Annex 11: Computerised Systems. 2011. Available from: https://health.ec.europa.eu/system/files/2016-11/annex11_01-2011_en_0.pdf

5. European Commission. EudraLex Volume 4, New Annex 22: Artificial Intelligence [Draft consultation guideline]. 2025. Available from: https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf

6. International Council for Harmonisation. ICH Q9(R1): Quality Risk Management. 2023. Available from: https://database.ich.org/sites/default/files/ICH_Q9(R1)_Guideline_Step4_2022_1219.pdf

7. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act). Official Journal of the European Union. 2024. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

8. U.S. Food and Drug Administration. Data Integrity and Compliance With Drug CGMP: Questions and Answers. Guidance for Industry. 2018. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-integrity-and-compliance-drug-cgmp-questions-and-answers

9. Medicines and Healthcare products Regulatory Agency. Guidance on GxP Data Integrity. 2018. Available from: https://www.gov.uk/government/publications/guidance-on-gxp-data-integrity

10. International Society for Pharmaceutical Engineering. GAMP® 5: A Risk-Based Approach to Compliant GxP Computerized Systems. 2nd ed. Tampa: ISPE; 2022.

11. U.S. Food and Drug Administration. Computer Software Assurance for Production and Quality System Software. Guidance for Industry and FDA Staff. 2025. Available from: https://www.fda.gov/media/188844/download

12. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020;33. Available from: https://arxiv.org/abs/2005.11401

Evaluating AI-Enabled Emr Analytics to Address Claims Gaps in Rare Disease: Insights From A Hereditary Angioedema (HAE) Pilot

Rob Sederman,[i] Casey Sederman,[i] Birnur Ozbas-Erdem,[i] Kathryn Sands,[ii] Jakob Lapsley,[ii] Xinkun Nie[ii]

ABSTRACT:

Objective: Across rare diseases, claims-based analytics suffer from structural limitations, including diagnostic ambiguity, incomplete therapy visibility, and a lack of clinically documented disease burden. These issues are magnified in conditions characterized by non-specific coding, heterogeneous presentations, and treatment pathways that are not reliably recorded in transactional claims systems. Hereditary Angioedema (HAE) exemplifies these challenges. HAE lacks a specific ICD-10 code, faces substantial therapy undercapture due to manufacturer and specialty-pharmacy blocks, exhibits attack rates that are poorly reflected in office-visit–dependent claims, and shows frequent treatment switching that cannot be explained through claims alone. These characteristics make HAE an ideal test case to evaluate whether AI-powered curation of electronic medical record (EMR) data, particularly unstructured physician notes, can improve diagnostic clarity, treatment visibility, and clinical characterization beyond what is achievable using claims data alone.

Approach: We apply an AI-powered EMR-analytics framework built on clinical and temporal reasoning to curate structured data from the EMR records of 4,368 patients with potential HAE indicators. Moving beyond conventional keyword-based NLP, the framework applies clinical reasoning, temporal sequencing, and conflict resolution to holistically reason throughout physician documentation and curate structured, clinically validated variables. These structured outputs include diagnostic confirmation, treatment history, therapy transitions, specialty attribution, and physician-documented disease burden. This methodology is compared against multi-year claims data to evaluate alignment and divergence across diagnostic accuracy, therapy visibility, and real-world disease dynamics.

Key Insights: Across all major analytic dimensions, EMR-curated patient journeys provided clarity where claims showed blind spots. EMR content differentiated true HAE from mimicking conditions frequently coded under non-specific ICD-10 categories. It revealed substantially higher therapy capture than what claims can provide under specialty-pharmacy and manufacturer blocks and uncovered treatment-switching drivers, such as efficacy concerns, not visible in claims. EMR narratives also documented attack frequency and symptom patterns that are largely absent from claims. These findings demonstrate that EMR-derived clinical documentation provides critical diagnostic and treatment insights that are not reliably observable in claims data alone.

Relevance: This work demonstrates that AI-enabled EMR analysis substantially improves the accuracy and completeness of rare disease analytics where claims alone are insufficient. By integrating the breadth of claims with clinically validated EMR-derived insights, life sciences teams can strengthen patient identification, treatment evaluation, and real-world evidence generation. The HAE pilot provides a proof of concept for applying this integrated, AI-driven framework across other under-coded or clinically complex rare diseases.

Key words: Artificial Intelligence (AI), Real-World Data (RWD), Real-World Evidence (RWE), Data Integration, Commercial Analytics

[i] Ambit RD Inc., https://ambitinc.com/
[ii] Lighten Platforms, Inc., https://www.lighten-ai.com/

INTRODUCTION

Real-world data (RWD) have become central to modern healthcare analytics, enabling evaluation of disease epidemiology, treatment patterns, and clinical outcomes outside controlled trial settings. Administrative claims data, in particular, provide large-scale longitudinal visibility across healthcare interactions, making them a foundational resource for life sciences research, real-world evidence generation, and commercial analytics.

Despite their utility, claims data present well-recognized structural limitations. Claims are designed primarily for billing and reimbursement rather than clinical documentation, resulting in incomplete clinical context and limited visibility into diagnostic confirmation, treatment rationale, and therapeutic response. These limitations are particularly consequential in rare diseases, where accurate patient identification, treatment visibility, and longitudinal disease characterization are essential for epidemiologic analysis, clinical trial recruitment, treatment evaluation, and evidence-based clinical and commercial decision-making.

Hereditary angioedema (HAE) illustrates these challenges. HAE is a rare genetic disorder characterized by recurrent, unpredictable swelling episodes affecting various anatomical locations, including the extremities, gastrointestinal tract, and airway. Accurate diagnosis is critical, as untreated attacks may result in significant morbidity and, in severe cases, mortality. However, HAE lacks a unique ICD-10 code, with many patients coded under broader categories such as complement system disorders. This nonspecific coding contributes to both false positive and false negative identification in claims-based analytics. Additionally, many HAE therapies are distributed through specialty pharmacies or administered in settings not consistently captured in claims, resulting in incomplete treatment visibility.

Electronic medical records (EMR), particularly physician-authored clinical notes, contain rich clinical information, including diagnostic confirmation, symptom severity, treatment rationale, and longitudinal disease progression. Beyond diagnostic confirmation, unstructured EMR documentation provides clinically meaningful insight into disease severity, progression, treatment response, and therapy utilization that cannot be reliably inferred from billing records alone. Physician-authored notes often contain detailed descriptions of symptom frequency, treatment decisions, therapeutic response, and reasons for discontinuation or switching therapies. These clinically validated longitudinal narratives enable more accurate reconstruction of patient journeys and provide critical context for evaluating real-world treatment effectiveness, disease burden, and therapy adoption patterns. Extracting structured, analysis-ready insights from unstructured EMR records has historically required labor-intensive manual abstraction by trained clinicians, limiting scalability and timeliness.

Traditional natural language processing (NLP) approaches have attempted to address this challenge but have often been limited by their reliance on keyword matching and pattern recognition, which do not fully capture clinical reasoning, diagnostic confirmation, or temporal relationships between clinical events. Rare diseases such as hereditary angioedema present additional complexity

due to nonspecific symptom descriptions, overlapping clinical presentations, and variable documentation practices across providers. These factors increase the risk of both incomplete patient identification and inaccurate disease characterization when relying solely on structured data fields or claims-based algorithms.

Recent advancements in clinical AI models have introduced the ability to perform contextual interpretation of medical narratives, enabling extraction of structured insights through clinical reasoning rather than simple keyword detection. These models can evaluate diagnostic evidence, reconcile conflicting information across records, and reconstruct longitudinal patient journeys, providing a more accurate and complete representation of disease progression, treatment exposure, and therapy sequencing. This capability is particularly valuable in rare diseases, where accurate patient identification, treatment visibility, and longitudinal disease characterization are essential for real-world evidence generation, evaluation of treatment effectiveness, and evidence-based clinical and commercial strategy.

To evaluate this potential at scale, a collaborative pilot was conducted between Ambit, a healthcare analytics firm with experience in rare disease strategy and real-world data analysis, and Lighten, a clinical AI technology company specializing in curation of unstructured EMR data. The partnership combined domain expertise in rare disease analytics with AI-enabled clinical reasoning capabilities to develop and apply a scalable framework for extracting structured clinical insights from longitudinal EMR records.

Using hereditary angioedema (HAE) as a test case, this study compares claims-derived and EMR-derived insights across key dimensions of rare disease analytics, including diagnostic accuracy, therapy capture, disease burden assessment, and treatment dynamics. By systematically evaluating differences between claims-based and EMR-derived patient journeys, this analysis quantifies the incremental value of AI-enabled EMR curation in improving diagnostic accuracy, treatment visibility, and completeness of real-world rare disease characterization.

METHODS

Data Sources

Two real-world data sources were used for analysis:

Administrative Claims Data:
Multi-year administrative claims data were analyzed, spanning June 1, 2020, through May 31, 2025. These data included medical and pharmacy claims across a nationally representative patient population, capturing diagnoses, procedures, and dispensed therapies.

Electronic Medical Record Data:
Longitudinal EMR data were analyzed across the complete available patient history. EMR records included structured fields and unstructured physician notes, containing clinical documentation related to diagnosis, symptoms, treatment decisions, and outcomes.

EMR data were derived from Loopback Health, a multisource EHR-based real-world database of US patients. All rights reserved by Loopback.

Patient-level linkage between datasets was performed using privacy-preserving tokenization methods, enabling comparison of claims-derived and EMR-derived insights while maintaining data privacy.

Study Population

Patients were identified based on the presence of potential hereditary angioedema (HAE) indicators in administrative claims and EMR data. In claims data, candidate patients were identified using the ICD-10 diagnostic code D84.1 ("Defects in the complement system"), which is the primary diagnostic code used in clinical practice to capture hereditary angioedema but is not specific to HAE. Patients were included if they had at least one medical claim containing ICD-10 code D84.1 during the observation period.

To improve specificity, additional claims-based identification criteria included the presence of HAE-specific therapies. These therapies included plasma-derived C1 esterase inhibitors (Berinert, Cinryze, Haegarda), recombinant C1 esterase inhibitor (Ruconest), kallikrein inhibitors (Kalbitor, Takhzyro, Andembry), bradykinin receptor antagonists (Firazyr, Icatibant), and oral prophylactic therapy (Orladeyo). Patients with evidence of one or more claims for these therapies were considered to have treatment patterns consistent with potential HAE diagnosis. A subset of patients with multiple diagnostic claims or treatment claims was also evaluated to reflect conventional claims-based patient identification methodologies commonly used in rare disease analytics.

Within EMR data, patients were initially identified using the same diagnostic and treatment indicators to ensure consistency with claims-based identification. Confirmed HAE diagnosis was determined through physician-documented clinical confirmation extracted from unstructured EMR records using AI-enabled curation. Diagnostic confirmation was based on explicit physician diagnosis statements, specialist documentation, laboratory evidence where available, and treatment patterns consistent with hereditary angioedema management. The abstraction framework evaluated longitudinal clinical documentation to distinguish confirmed HAE from related or nonspecific complement disorders.

Claims-based and EMR-derived patient cohorts were linked using privacy-preserving tokenization methods, enabling comparison of diagnostic accuracy, treatment capture, and disease burden across data sources while maintaining patient privacy.

AI-Enabled EMR Curation Framework

Framework Overview

An AI-enabled curation framework was applied to unstructured EMR records to curate structured clinical information. The framework incorporated multiple clinical reasoning capabilities to holistically reason through physician-authored documentation, reconstruct longitudinal patient histories, and generate analysis-ready structured datasets. A clinical curation expert was engaged at every stage of the workflow to ensure credible, accurate, and coherent data curation.

Core framework capabilities included:

Clinical Reasoning

The system evaluated physician notes to identify diagnostic confirmation, treatment patterns, symptom documentation, and clinical rationale.

Temporal Sequencing

The framework reconstructed longitudinal patient journeys, identifying diagnostic timing, treatment initiation, switching events, and disease progression.

Conflict Resolution

When inconsistent clinical documentation existed across records, the framework applied clinical reasoning to reconcile discrepancies and determine the most clinically plausible interpretation.

Clinical Abstraction and Diagnostic Confirmation

Diagnostic confirmation was determined through systematic evaluation of physician-authored clinical documentation, including explicit diagnostic statements, specialist assessments, laboratory findings, and treatment history consistent with hereditary angioedema. The abstraction framework evaluated multiple clinical indicators to distinguish confirmed HAE from conditions with overlapping symptoms or nonspecific coding.

When documentation contained conflicting or ambiguous information, the system applied clinical reasoning to evaluate the overall evidence across the longitudinal patient record, prioritizing specialist-confirmed diagnoses, objective clinical findings, and treatment patterns consistent with HAE management.

Longitudinal Clinical Reconstruction

Longitudinal reconstruction enabled identification of key clinical events, including diagnosis timing, treatment initiation, therapy escalation, switching, and discontinuation. This temporal framework provided clinically validated sequencing of disease progression and treatment exposure, addressing limitations in claims data caused by incomplete capture or fragmented observation.

Treatment Capture and Exposure Assessment

Treatment capture within EMR records included therapies administered across all care settings, including specialty pharmacy distribution, infusion centers, and physician-administered treatments that may not be consistently captured in claims.

Physician documentation provided direct evidence of treatment initiation, continuation, discontinuation, and therapeutic response. This enabled comprehensive reconstruction of therapy exposure and treatment sequencing, overcoming structural limitations in claims visibility.

Disease Burden Assessment

Disease burden assessment was performed using physician-documented attack frequency, and treatment response. Clinical notes provided detailed descriptions of swelling episodes, attack frequency, and symptom progression, enabling evaluation of disease severity and treatment effectiveness.

These clinically derived measures provided longitudinal insight into disease burden not reliably available in claims data, which capture only healthcare encounters associated with billable services.

Data Structuring and Variable Generation

Extracted clinical insights were transformed into structured datasets suitable for quantitative analysis, including diagnostic status, treatment history, attack frequency, treatment sequencing, and treatment rationale. Structured outputs enabled systematic comparison of EMR-derived clinical variables with claims-derived measures.

Data Linkage and Privacy Protection

Privacy-preserving tokenization methods were used to enable patient-level linkage between EMR and claims datasets while maintaining compliance with data protection requirements. Tokenization enabled integration of complementary datasets without exposing personally identifiable information, allowing comparative analysis of diagnostic accuracy, treatment patterns, and disease burden.

Analytical Framework

Comparative analysis between administrative claims data and EMR-derived structured data was conducted to evaluate differences in patient identification, treatment capture, disease burden characterization, and treatment dynamics. The unit of analysis was the individual patient, with claims-derived and EMR-derived clinical profiles compared for each linked patient.

Comparative evaluation was conducted across the following analytical dimensions:

1. **Diagnostic accuracy and patient identification:** Claims-based identification was compared with EMR-confirmed diagnosis, which served as the clinical reference standard. Sensitivity and specificity were calculated to assess the ability of claims-based methodologies to correctly identify confirmed HAE patients.
2. **Therapy capture and treatment patterns:** Treatment exposure was assessed using claims-derived pharmacy and medical claims and EMR-derived physician-documented treatment history. Differences in treatment capture rates were evaluated to quantify incomplete therapy visibility in claims data.
3. **Disease burden assessment:** Disease burden was evaluated using EMR-derived physician-documented attack frequency, symptom progression, and treatment response. These measures were compared with claims-derived proxies, such as healthcare utilization and treatment claims.
4. **Treatment initiation timing:** Time from documented diagnosis to treatment initiation was evaluated using longitudinal EMR records and compared with treatment timing observable in claims data.
5. **Treatment switching and discontinuation:** Treatment switching patterns and discontinuation events were identified using EMR-derived longitudinal treatment histories and compared with treatment claims.
6. **Clinical rationale for treatment decisions:** Physician-documented clinical rationale for treatment initiation, switching, or discontinuation was evaluated using EMR abstraction to provide clinical context not available in claims data.

This comparative framework enabled systematic evaluation of differences in analytic completeness, diagnostic accuracy, and treatment visibility between claims-derived and EMR-derived datasets.

Statistical Analysis

Descriptive statistics were used to characterize patient populations, diagnostic classification, and treatment patterns. Continuous variables were summarized using means, medians, and ranges, while categorical variables were summarized using counts and proportions.

Diagnostic sensitivity and specificity of claims-based patient identification were calculated using EMR-confirmed diagnosis as the reference standard. Sensitivity was defined as the proportion of EMR-confirmed HAE patients correctly identified using claims-based identification criteria. Specificity was defined as the proportion of patients without EMR-confirmed HAE correctly excluded using claims-based identification.

Treatment capture rates were calculated as the proportion of confirmed patients with evidence of treatment exposure identified in claims data compared with EMR-derived treatment histories. Differences in treatment capture rates between claims and EMR data were evaluated descriptively.

Longitudinal analysis was conducted to evaluate treatment initiation timing, switching patterns, and discontinuation events using EMR-derived structured clinical data. Temporal relationships between diagnosis, treatment initiation, and treatment changes were evaluated using longitudinal patient timelines.

All analyses were conducted using de-identified datasets at the patient level.

RESULTS

Diagnostic Accuracy and Patient Identification

Among 4,368 patients with potential HAE indicators in EMR data, 1,593 patients (36%) had physician-confirmed HAE diagnosis.

Claims-based identification demonstrated limited sensitivity relative to EMR confirmation. Claims-based algorithms identified only 40% of confirmed HAE patients, indicating that approximately 60% of patients with clinically confirmed HAE in EMR data would be missed using claims alone, reflecting low diagnostic sensitivity of claims-based approaches. Claims specificity was 89%, indicating some degree of false positive identification due to nonspecific diagnostic coding.

These findings reflect structural limitations in claims-based identification driven by nonspecific diagnostic coding and incomplete clinical context. In hereditary angioedema, the absence of a disease-specific ICD-10 code contributes to both missed confirmed patients and misclassification of patients without true disease.

EMR-derived diagnostic confirmation provided direct clinical validation, reducing both false positive and false negative identification. Physician documentation included explicit diagnostic confirmation, laboratory findings, and specialist assessments, providing a reliable reference standard for disease classification. The improved diagnostic accuracy afforded by EMR data has important implications for epidemiologic estimation, patient identification, and real-world evidence generation.

Claims and EMR data linkage demonstrated that among 4,368 patients with potential HAE indicators identified in EMR data, 4,162 (95%) were also present in claims data, enabling direct comparison between datasets. Among these linked patients, claims-based identification produced 892 false negatives, representing confirmed HAE patients identified in EMR but not captured using claims-based criteria, and 291 false positives, representing patients classified as potential HAE in claims but lacking clinical confirmation in EMR documentation. These findings demonstrate that reliance on claims-based identification alone results in both underreporting of true HAE patients and misclassification of patients without confirmed disease.

Additional EMR-derived clinical variables further demonstrated improved diagnostic precision. For patients with available diagnostic age information, EMR-derived age at diagnosis was more consistent with established epidemiologic expectations, whereas claims-derived diagnostic timing appeared substantially delayed due to incomplete longitudinal patient history and reliance on initial claim appearance rather than true diagnostic confirmation. EMR documentation also enabled accurate identification of diagnosing provider specialty, restoring clinical attribution that is frequently misclassified or incomplete in claims-based analysis.

The substantial difference in therapy capture between EMR and claims further illustrates structural limitations in claims-based analysis. Specialty-distributed therapies, physician-administered treatments, and infusion-based therapies may not generate traditional pharmacy claims, resulting in incomplete therapy visibility. EMR documentation, by contrast, reflects physician treatment decisions and clinical management regardless of billing pathway, providing a more complete representation of therapy exposure.

Enhanced visibility into treatment initiation timing and switching patterns provided additional clinical insight. EMR documentation enabled identification of treatment initiation delays, escalation patterns, and reasons for therapy discontinuation. This longitudinal clinical context is essential for understanding real-world treatment effectiveness and disease management strategies.

Therapy Capture and Treatment Patterns

Therapy capture differed substantially between claims and EMR data.

In claims data, treatment was observed in only 6% of confirmed patients, reflecting structural limitations in therapy visibility, particularly for specialty-distributed and infusion-based therapies.

In contrast, AI-enabled EMR curation identified treatment in 73% of confirmed patients. EMR records provided comprehensive treatment documentation, including therapies administered outside traditional claims capture pathways.

EMR data also revealed more accurate representation of treatment strategies, including the use of both prophylactic and on-demand therapies. Claims data disproportionately reflected on-demand therapies due to incomplete capture of prophylactic treatments.

Disease Burden Assessment

EMR data provided substantially greater visibility into disease burden compared with claims. Attack frequency, a key indicator of disease severity, was inconsistently captured in claims data, as attacks are only reflected when associated with billable healthcare encounters.

EMR documentation enabled longitudinal tracking of attack frequency based on physician notes and patient reported attacks, providing more accurate assessment of disease burden and progression.

Among prophylactically treated patients, claims data captured only a limited subset of disease activity. Specifically, among 491 patients receiving prophylactic therapy, claims data identified 246 attack-related claims, reflecting only attacks associated with billable healthcare encounters. In contrast, EMR-derived clinical documentation provided longitudinal capture of attack frequency regardless of healthcare utilization, enabling more comprehensive assessment of disease activity and treatment response.

EMR-derived longitudinal tracking further enabled identification of patients experiencing persistent disease activity despite therapy. For example, among 429 patients receiving a commonly prescribed prophylactic therapy, 107 patients experienced two or more documented attacks per year while on treatment. This level of clinical detail is not reliably observable in claims data and provides critical insight into real-world treatment effectiveness and disease control durability.

This enhanced visibility enables more precise evaluation of treatment effectiveness and patient outcomes.

Treatment Initiation and Switching Patterns

EMR data enabled reconstruction of longitudinal treatment journeys, including time from diagnosis to treatment initiation. EMR-derived clinical histories demonstrated that treatment initiation frequently occurred substantially later than suggested by claims data. Among 211 patients with confirmed diagnosis and documented treatment initiation dates, approximately 50% initiated prophylactic therapy at least three years after diagnosis. Claims-based analysis underestimated this delay due to incomplete longitudinal history and diagnostic ambiguity, highlighting the importance of EMR-derived clinical confirmation for accurate evaluation of treatment timing.

EMR analysis also provided visibility into treatment switching patterns. EMR-derived treatment histories also enabled detailed evaluation of treatment sequencing and switching patterns. Among 54 patients initiating a commonly prescribed prophylactic therapy as a second-line or later treatment, prior therapy exposure and switching patterns were clearly observable, enabling accurate reconstruction of treatment progression and source of therapy initiation. These insights are not reliably derivable from claims data alone due to incomplete therapy capture and lack of clinical context.

Among patients receiving a commonly prescribed prophylactic therapy, 43% of treatment discontinuations were associated with physician-documented lack of efficacy. These findings demonstrate that EMR-derived clinical documentation provides direct insight into treatment effectiveness and drivers of discontinuation, enabling evaluation of real-world therapeutic response and persistence patterns that are not observable using claims-based analysis alone.

EMR-derived longitudinal clinical data also enabled evaluation of treatment response durability and disease progression over time. Continuous clinical documentation provided visibility into attack frequency trends, treatment adjustments, and persistence patterns, enabling more comprehensive characterization of real-world treatment effectiveness than is achievable using claims data alone.

DISCUSSION

This study demonstrates that AI-enabled EMR curation substantially improves diagnostic accuracy, treatment visibility, and analytic completeness compared with claims-based analysis alone. Using hereditary angioedema as a test case, EMR-derived clinical confirmation revealed that the majority of confirmed patients would not have been reliably identified using claims-based criteria. These findings reflect fundamental structural differences between the two data sources. Claims provide population-scale longitudinal visibility but lack clinical validation and complete treatment capture, while EMR documentation provides direct evidence of diagnosis, treatment decisions, and disease progression. AI-enabled data curation enables extraction of these clinically validated insights while overcoming the scalability limitations of manual chart review and enabling reconstruction of longitudinal patient journeys with greater accuracy.

The complementary strengths of claims and EMR data highlight the importance of integrated real-world data analysis. Claims provide population-level insights and longitudinal continuity, while EMRs provide diagnostic validation, comprehensive treatment visibility, and detailed clinical context. Integration of these datasets enables more accurate identification of rare disease populations and more reliable evaluation of treatment patterns and outcomes.

Improved diagnostic confirmation and treatment visibility have important implications for real-world evidence generation, clinical research, and therapeutic strategy. More accurate patient identification enables reliable epidemiologic estimation and clinical trial feasibility assessment, while enhanced visibility into treatment sequencing and response enables evaluation of therapy adoption, persistence, and effectiveness. These capabilities support more informed clinical, research, and commercial decision-making.

From a real-world evidence and therapeutic strategy perspective, improved diagnostic confirmation and treatment visibility enable more accurate characterization of therapy adoption patterns, persistence, and treatment sequencing. This enables more reliable evaluation of treatment effectiveness in real-world populations and supports evidence-based clinical and commercial decision-making. Incomplete visibility into diagnosis and treatment may lead to underestimation of disease prevalence, mischaracterization of treatment utilization, and incomplete understanding of treatment effectiveness.

The methodological framework evaluated in this study demonstrates scalability to other rare diseases and clinical conditions where claims-based identification is limited by nonspecific coding, fragmented treatment pathways, and incomplete clinical context. Many rare diseases lack unique diagnostic codes or involve complex treatment pathways that are not fully observable in administrative claims. AI-enabled EMR curation provides a scalable approach to overcome these limitations and improve the accuracy and completeness of rare disease analytics.

These findings demonstrate that AI-enabled curation of EMR data represents a scalable and clinically grounded approach to enhancing real-world rare disease analytics. By enabling more accurate patient identification, comprehensive treatment visibility, and clinically validated longitudinal disease characterization, this approach strengthens the foundation for real-world evidence generation and supports more informed clinical, research, and therapeutic decision-making.

LIMITATIONS

EMR data coverage may not fully represent the national patient population, and documentation completeness may vary across providers.

AI-enabled curation accuracy depends on the quality and completeness of clinical documentation.

Claims and EMR datasets may differ in population composition and observation periods.

Despite these limitations, the findings demonstrate consistent and substantial improvements in analytic completeness using AOI-enabled EMR curation.

CONCLUSIONS

AI-enabled curation of EMR data substantially enhances rare disease analytics by improving diagnostic accuracy, treatment visibility, and longitudinal disease characterization beyond what is achievable using claims data alone. Integration of clinically validated EMR-derived insights with claims-based population breadth provides a more complete and reliable foundation for real-world evidence generation.

This approach enables more accurate patient identification, reconstruction of treatment journeys, and evaluation of disease burden and therapeutic response in real-world populations. The findings demonstrate that AI-enabled EMR curation represents a scalable and clinically grounded framework for improving analytic completeness across rare diseases and complex clinical conditions where claims data alone are insufficient.

REFERENCES

Core Real-World Data and Claims Limitations

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323–37.

2. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Ann Intern Med. 1993 Oct 15;119(8):844–50.

3. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013 Aug;51(8 Suppl 3):S30–7.

EMR Data and Real-World Evidence

4. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence—what is it and what can it tell us? N Engl J Med. 2016 Dec 8;375(23):2293–7.

5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 Jun;13(6):395–405.

Rare Disease Identification Challenges

6. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, et al. Rare disease terminology and definitions—a systematic global review. Value Health. 2015 Sep;18(6):906–14.

7. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases. Eur J Hum Genet. 2020 Feb;28(2):165–73.

8. Facey KM, Hansen HP, Single ANV, editors. Rare Disease Registries: Key Considerations and Best Practice. Copenhagen: Pan European Networks; 2014.

Hereditary Angioedema Epidemiology and Clinical Management

9. Busse PJ, Christiansen SC. Hereditary angioedema. N Engl J Med. 2020 Feb 27;382(12):1136–48.

10. Maurer M, Magerl M, Betschel S, Aberer W, Ansotegui IJ, Aygören-Pürsün E, et al. The international WAO/EAACI guideline for the management of hereditary angioedema. Allergy. 2022 Jul;77(7):1961–90.

AI, NLP, and EMR Abstraction

11. Carlisle MN, et al. Development and validation of a generative artificial intelligence system for automated extraction of clinical data from electronic health records. JMIR Med Inform. 2026;14:e70708.

12. Dagli MM, et al. Development and validation of a large language model–integrated NLP framework for automated extraction of surgical data from electronic health records. Sci Rep. 2024;14:77535.

13. Alkhalaf M, et al. Retrieval-augmented generative AI for automated extraction and structured summarization of electronic health record data. J Biomed Inform. 2024;152:104667.

14. Clay B, et al. Natural language processing techniques applied to electronic health records: applications and pipeline considerations. Comput Methods Programs Biomed. 2025;244:108009.

15. Perkins SW, et al. Improving clinical documentation with artificial intelligence: a systematic review. J Am Med Inform Assoc. 2024;31(9):1802-1816.

GenAI-Powered Conversational BI Assistant: Accelerating Pharma Dashboard Insights

Vinay Vihari Lakamsani, Senior Manager, Data & AI, Trinity Life Sciences;
Mohamed Arruman Shalik, Consultant, Data Science, Trinity Lifesciences;
Pulkit Sharma, Principal, Data & AI, Trinity Life Sciences

1. ABSTRACT

Pharmaceutical commercial teams often struggle to extract precise insights from dashboards with limited views, time consuming navigation, and misinterpretation of visualizations, slow decision making, and increasing cognitive load.

Our objective was to develop a GenAI-powered BI Assistant that transforms static dashboards into dynamic, conversational experiences. The solution enables natural language interaction with Power BI dashboards, delivering instant summaries, reasoning, and actionable insights.

We designed and implemented a conversational analytics platform integrated with BI dashboards and underlying datasets. The system uses GenAI models to interpret user queries, generate SQL dynamically, and return responses enriched with insights, assumptions, executed queries, reasoning steps, visualizations, and follow-up questions.

Key components include an automated onboarding interface for integrating multiple dashboards and data sources with minimal effort, supported by AI-driven semantic layer creation and analyst review. A conversational interface provides real-time answers to metric, KPI, and trend-related questions. An admin and governance panel tracks usage, manages onboarding, and ensures reliability while capturing user feedback for continuous improvement.

The platform achieved over 90% accuracy in answering user questions after minimal manual cleanup of raw tables. Compared to manual dashboard navigation, the solution reduced average query resolution time by 30%, saving an average of ~5 minutes per query. This improvement was benchmarked against traditional workflows where users manually navigate multiple dashboard views. The assistant also enabled rapid scenario analysis for commercial brand performance dashboards, providing dynamic responses to questions not addressed by static views. Users reported improved trust and transparency due to explanations of assumptions and underlying queries.

Challenges included preparing raw data tables for GenAI consumption, such as column cleanup and naming standardization; designing a seamless onboarding module with an intuitive interface for schema validation and review; ensuring compatibility across diverse dashboard views and data sources; and handling incomplete or ambiguous user questions while generating accurate visualizations. Strategies to address these challenges will be detailed in the full manuscript.

This innovation introduces automated semantic layer creation and multi-dashboard onboarding, enabling scalability across enterprise analytics. The chatbot interface is deployable as a standalone application or browser plugin, ensuring flexibility and ease of adoption. By combining natural language processing with BI integration, the platform sets out a new benchmark for user centric analytics.

By bridging the gap between complex BI tools and business decision makers, this solution accelerates insight-to-action cycles in pharmaceutical analytics. It supports data-driven decisions across commercial analytics, forecasting, and market access, demonstrating how GenAI can operationalize advanced analytics for broader enterprise adoption.

Keywords: Generative AI, Business Intelligence, Conversational Analytics, Dashboard Insights, Pharmaceutical Analytics

2. BACKGROUND

Business intelligence (BI) dashboards serve as a foundational pillar of pharmaceutical analytics, enabling data-driven decision making across commercial performance tracking, brand monitoring, campaign effectiveness, forecasting, and market access analyses. Over time, organizations have developed significant institutional maturity in leveraging dashboards as structured decision-support tools. Platforms such as Microsoft Power BI and Tableau have evolved into highly sophisticated ecosystems, offering rich visualization capabilities, interactive filtering, drill-down functionality, role-based access, and scalable enterprise integration. As a result, dashboards hold an established and trusted position within commercial and analytics operating models.

This maturity has enabled dashboards to become the primary interface through which marketing teams, business stakeholders, and analytics functions consume performance insights. They provide standardized reporting frameworks, ensure metric consistency, and support governance and traceability requirements that are critical in pharmaceutical environments.

However, as analytical needs continue to expand in complexity and speed, certain structural characteristics of traditional dashboards introduce friction in day-to-day insight generation. Information is often distributed across multiple pages, tabs, and visual components, requiring users to navigate between views and mentally synthesize trends across disparate charts, tables, and KPIs. While interactive features mitigate some of this fragmentation, cross-metric or cross-view analysis frequently demands manual interpretation, increasing cognitive effort and extending the time required to derive actionable conclusions.

Usability considerations also play a role. Effective interpretation of dashboards often assumes familiarity with layout conventions, metric definitions, business logic, and the analytical intent embedded within each visualization. For new or infrequent users, developing this contextual understanding can require time and domain exposure, potentially limiting broader accessibility of insights across stakeholder groups.

Most importantly, dashboards are inherently designed around predefined analytical perspectives. Even with advanced filtering and drill capabilities, users remain bounded by the metrics, aggregations, and visual constructs configured during development. When business questions extend beyond these predefined views such as exploratory combinations of metrics, dynamic re-aggregation, or deeper root-cause analysis users must often engage analytics teams for additional support.

This dependency does not diminish the value of dashboards; rather, it reflects the distinction between structured reporting and dynamic analytical exploration. In fast-moving commercial environments, the time required to translate new business questions into updated analyses can affect insight velocity and responsiveness.

Collectively, these considerations highlight an opportunity not to replace dashboards, but to augment them. While dashboards remain indispensable for standardized reporting, governance, and executive tracking, there is scope to enhance the analytical experience by enabling more flexible, intuitive, and conversational interaction with underlying data. Such augmentation can preserve the strengths of established BI systems while expanding their ability to support rapid exploration, cross-dimensional reasoning, and evolving business inquiries within pharmaceutical organizations.

3. OBJECTIVE AND VISION

The objective of this work is to enhance how pharmaceutical stakeholders interact with dashboard-driven analytics by introducing a conversational interface that complements existing BI ecosystems while supporting evolving decision-making needs. The intent is not to replace dashboards, but to extend their value by enabling more flexible and intuitive interaction with governed analytical assets.

At the core of this vision is a natural language interface through which business users can express analytical questions in plain language and receive structured, context-aware responses grounded in enterprise-approved metrics. Structured navigation within dashboards will continue to be important, particularly in broad analytical domains. The conversational layer serves as an additional interaction modality that supports exploratory, cross-metric, and iterative questioning when business intent extends beyond predefined visual pathways.

From an architectural perspective, the conversational layer operates within the same governed analytical framework as existing dashboards. It leverages curated semantic models, validated metric definitions, and approved data views that are already established within enterprise BI platforms such as Power BI or Tableau. Any access to underlying data occurs within controlled, user-initiated workflows aligned with existing governance structures, ensuring consistency with organizational standards.

Deployment is envisioned in a complementary manner. The conversational interface may be embedded within existing dashboards or deployed as a parallel application integrated with the same semantic layer. In practice, dashboards continue to provide standardized KPI monitoring and structured reporting, while the conversational layer supports dynamic follow-up questions, rapid drill exploration, and iterative reasoning.

The broader vision includes formalizing the semantic context derived from dashboards, such as metric definitions and business rules, into a reusable enterprise knowledge foundation. This foundation can support conversational BI as well as future generative AI applications across analytics and decision support functions.

In summary, this work proposes a complementary operating model in which traditional dashboards and conversational analytics function side by side.

Dashboards continue to anchor standardized reporting, while the conversational layer enhances flexibility and accelerates insight generation, strengthening analytical responsiveness within pharmaceutical organizations.

4. METHODOLOGY

Following the identification of key limitations in static dashboard-based analytics and the vision for a conversational BI assistant, the proposed solution was developed using a structured methodology centered on three foundational capabilities: (1) automated dashboard onboarding, (2) semantic layer creation and validation, and (3) a hybrid semantic and Agentic AI framework that combines large language models with specialized SQL-generation agents to support reliable query execution, insight synthesis, and conversational decision support.

Together, these components enable scalable integration of enterprise dashboards with natural language analytics.

4. 1 Dashboard Onboarding Layer

The first methodological step involves onboarding new dashboards into the conversational analytics platform. Dashboard onboarding is designed to be primarily automated and is initiated by an application or product owner with appropriate administrative access.

During onboarding, the platform is granted controlled access through an application service account, allowing an automated pipeline to traverse all dashboard pages and visual elements. Key metrics and dashboard content are captured through systematic screenshot extraction. These visual artifacts are subsequently processed using multimodal large language models (LLMs) to convert chart and card information into structured textual representations.

In parallel, the platform connects to the underlying data sources associated with the dashboard, such as relational database tables. An automated context creation manager scans available tables, column structures, and representative values to generate an initial schema-level understanding of the dataset landscape.

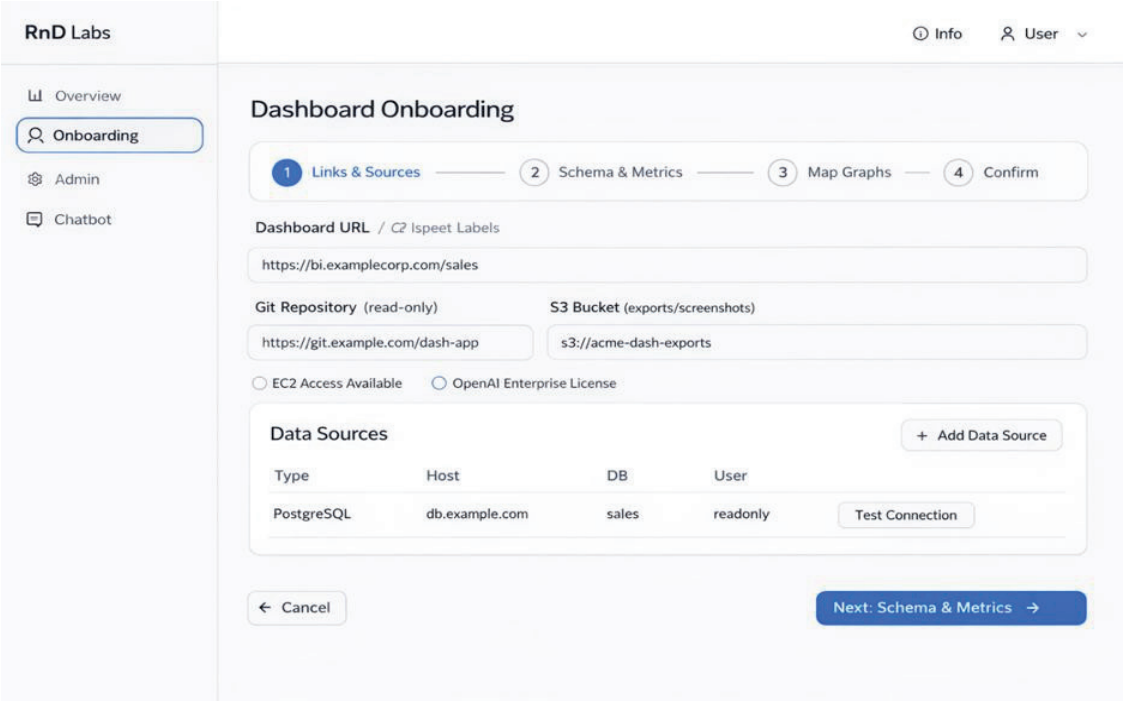


Figure 4.1. *Dashboarding Onboarding and Semantic Validation Workflow*

The outputs from dashboard visual extraction and backend schema discovery are combined to produce an initial inventory of metrics, KPI definitions, and dataset context. This automated onboarding approach minimizes manual effort while accelerating dashboard readiness for conversational interaction.

4.2 Semantic Layer Creation and Validation

Once onboarding is completed, the platform establishes a semantic layer that serves as the foundation for reliable natural language querying. This semantic layer integrates three key elements: validated schema definitions, metric-level business context, and subject matter expertise.

Although initial semantic context is generated automatically by AI agents, human validation is incorporated as a critical governance step. The dashboard owner or subject matter expert reviews the inferred schema, confirms metric definitions, and enriches the context with additional business rules, including primary key-foreign key relationships, domain-specific terminology, and organizational metric logic.

To ensure factual consistency with the source dashboard, the platform provides a validation module in which each metric definition is paired with an AI-generated SQL query. These queries can be executed directly through the interface against the connected backend database.

Retrieved results are then passed to a visualization reconstruction agent, which regenerates charts comparable to those displayed in the original dashboard.

This iterative validation workflow ensures that dashboard-reported values align with system-derived outputs. Any discrepancies can be corrected through refinement of schema context, metric definitions, or query logic. Once validated, the enriched semantic layer is stored as a reusable knowledge foundation for future conversational analytics.

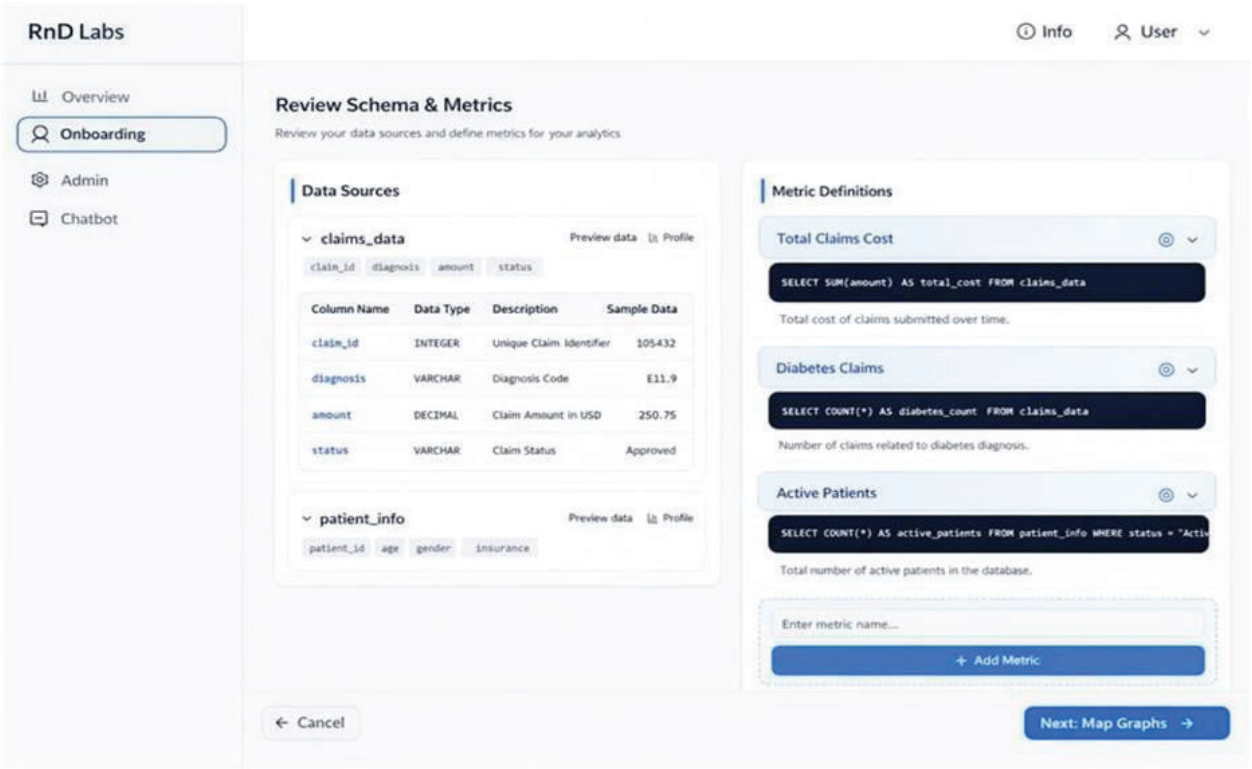


Figure 4.2. *Enterprise Semantic Layer construction and Knowledge Formalization Framework*

After user validation, the platform consolidates and stores all information across the business context, semantic layer, and AI data layer. This structured, interconnected data foundation enhances consistency, improves context retrieval, and enables the system to answer a wide range of user questions with greater precision.

4.3 Conversational Text-to-SQL Reasoning Engine

The core analytical capability of the platform is a multi-agent conversational text-to-SQL engine that translates user questions into executable database queries and interpretable insights.

The workflow begins with an analyst orchestration agent that evaluates whether a user query is relevant, well-formed, and answerable within the scope of the onboarded dashboard and semantic context. If queries are incomplete or ambiguous, the assistant prompts the user for clarification, improving robustness in real-world usage.

For valid analytical requests, a specialized SQL generation agent produces expert-level SQL aligned with both the validated semantic layer and database schema. Guardrails are incorporated to prevent unsafe operations, such as unauthorized data access, schema modification, or destructive queries. In addition to these guardrails, dedicated quality-check agents validate the generated SQL against schema constraints, semantic consistency, and logical correctness before execution.

The generated SQL is executed through a database manager module responsible for controlled query execution and data retrieval. If execution errors occur due to syntax issues, performance inefficiencies, or logical inconsistencies, an automated retry mechanism is triggered. Error traces and execution feedback are routed back to the SQL generation agent and quality-check layer for refinement. This iterative correction cycle may repeat in a controlled manner until an optimal and executable query is derived, ensuring reliability while avoiding excessive query loops.

Once results are successfully retrieved, an insights agent synthesizes the output into business-friendly summaries, highlighting key figures, trends, and assumptions used in interpretation. When appropriate, a visualization agent determines whether chart-based representation is suitable and generates interactive visualization specifications (for example, Highcharts-renderable HTML) for direct embedding into the user interface.

Through this layered architecture, the assistant delivers not only numerical outputs, but also reasoning-driven insights and contextual visual explanations grounded in validated enterprise data.

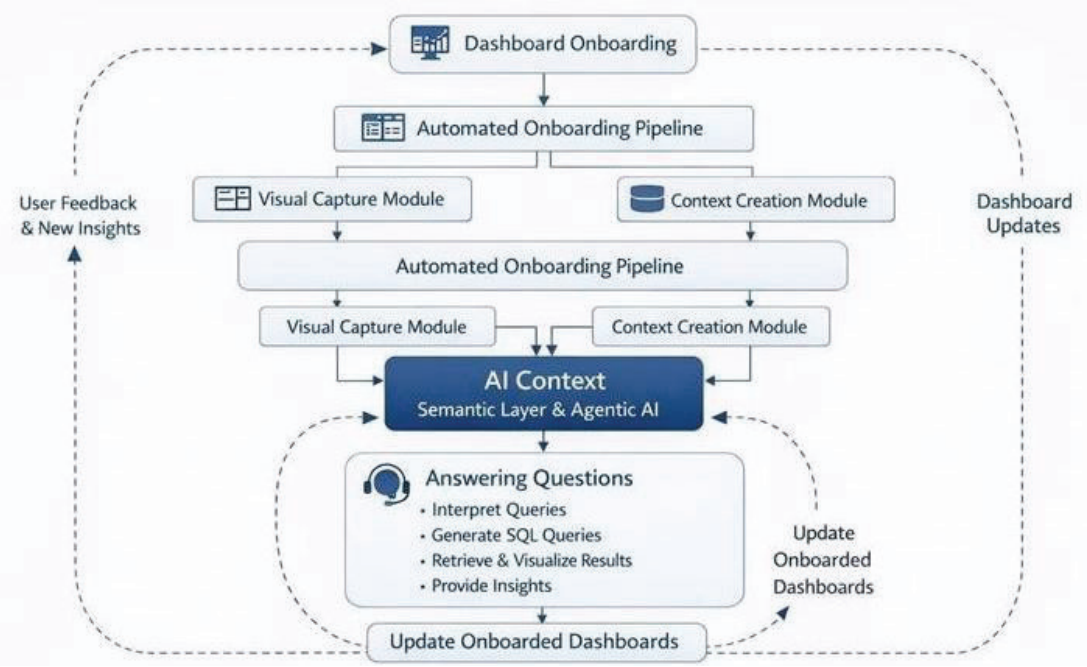


Figure 4.3. *End-to-End Methodology*

5. SYSTEM ARCHITECTURE OVERVIEW

The platform architecture distinguishes between two primary user roles: administrative owners and business end users. Application owners manage dashboard onboarding, semantic layer validation, and governance workflows, while end users interact exclusively through the conversational interface.

Once a dashboard has been onboarded and its semantic context validated, end users may submit natural language questions related to dashboard performance, KPIs, trends, or exploratory analytical needs.

If a query corresponds to an existing validated metric, the system leverages the stored semantic SQL logic for efficient execution and consistent interpretation. If the query extends beyond predefined dashboard metrics, the assistant dynamically generates new SQL using its schema and business context, enabling ad hoc exploration without requiring manual analyst intervention.

The platform also supports continuous learning through user-driven caching. When end users identify particularly useful derived insights or queries, these interactions can be reviewed and incorporated into the semantic layer, progressively expanding the assistant's reusable knowledge base.

To further enhance usability, a recommendation agent generates related follow-up questions based on the user's analytical intent and the available semantic context. This enables multi-turn conversational workflows and supports iterative investigation and deeper analytical reasoning.

Importantly, the agentic architecture is designed with modular responsiveness in mind. While the current implementation centers on text-based interaction, the underlying orchestration, semantic grounding, and SQL reasoning components are adaptable to additional interaction channels. The system can operate at varying levels of response speed and analytical granularity, enabling future integration into more responsive environments, including live or voice-enabled workflows, without requiring a fundamental redesign of the analytical core.

Overall, the architecture operationalizes conversational BI as an adaptive layer on top of enterprise dashboards, enabling standardized reporting consistency, flexible exploratory analytics, and extensibility for evolving interaction models.

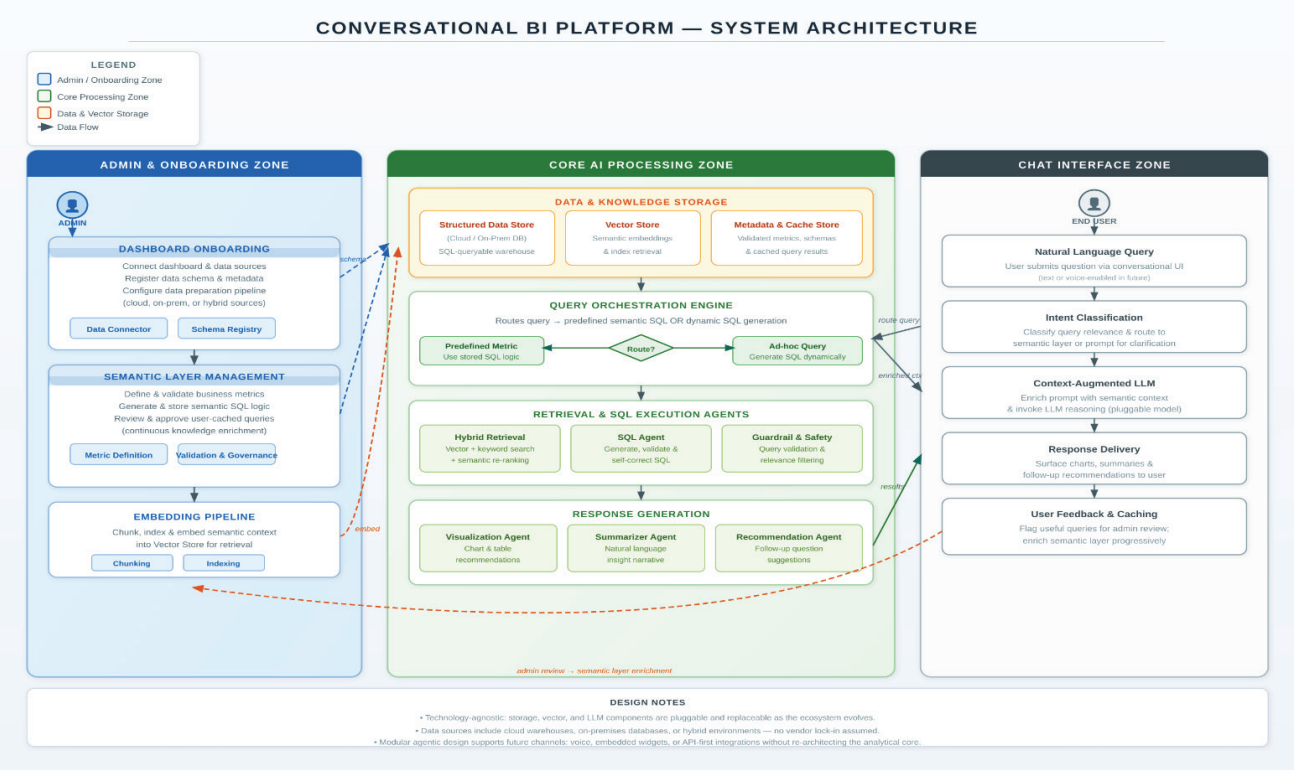


Figure 5. *High Level Agentic Conversational BI Functional Architecture*

6. EVALUATION AND LLM OPERATIONS MONITORING

Evaluation of conversational analytics systems in enterprise settings presents challenges beyond traditional machine learning metrics such as precision and recall.

Given the applied nature of this solution and its emphasis on usability and decision support, performance was assessed through a structured feedback and observability framework.

A user feedback mechanism was embedded directly into the interface, allowing end users to provide both binary satisfaction indicators and qualitative comments on response usefulness, accuracy, and interpretability. This feedback was persistently stored and used to assess real-world assistant reliability.

In addition, all interactions across system modules, including agent inputs and outputs, SQL execution traces, token usage, retry cycles, and feedback metadata were captured within an LLMOps monitoring tool. This end-to-end tracing enabled debugging, governance oversight, and continuous improvement of agent behavior.

Across ~1,200 traced interactions, aggregated feedback indicated an overall response accuracy and usefulness score of ~93%. This monitoring-driven evaluation approach provided both transparency and operational control, supporting scalable deployment in enterprise pharmaceutical analytics environments.

7. GOVERNANCE, COMPLIANCE, AND ENTERPRISE CONTROLS

The deployment of Generative AI systems within pharmaceutical analytics environments requires strong governance to ensure trust, reliability, and compliance with organizational and regulatory expectations. Given that commercial and clinical decision-making often depends on sensitive, high-impact data, the platform was designed with multiple layers of governance and operational safeguards.

A central governance principle of the solution is the separation of responsibilities between administrative owners and business end users. Application owners maintain control over dashboard onboarding, semantic layer validation, and metric certification workflows, ensuring that the assistant's knowledge base is grounded in verified business definitions rather than unreviewed automated interpretations. This validation-first approach reduces the risk of misaligned metric logic and promotes consistency with enterprise reporting standards.

To further strengthen reliability, the platform incorporates semantic approval gates before metrics become available for conversational querying. Only validated schema context, approved metric definitions, and reconciled SQL logic are promoted into the production semantic layer. This ensures that high-frequency business questions are answered using certified analytical foundations rather than ad hoc model inference alone.

From a compliance and operational control perspective, multiple guardrails were implemented within the text-to-SQL execution workflow. The SQL generation agent operates under constrained permissions and is restricted from producing unsafe operations such as data deletion, schema modification, or unauthorized table access.

Query execution is mediated through a controlled database manager layer, enabling enforcement of access controls aligned with enterprise service accounts and existing governance structures.

In addition to control mechanisms, explainability was treated as a core design requirement. Each conversational response is accompanied by user-friendly trace elements, including the underlying metric logic, executed SQL (where appropriate), applied filters, and summarized assumptions used in interpretation. This layered transparency allows business users to understand not only the answer, but how the answer was derived, thereby increasing confidence in the system's outputs.

At the operational level, transparency and auditability are further supported through comprehensive logging. All agent-level interactions are captured through the LLMOps monitoring framework, providing lineage across prompt inputs, SQL generation steps, retry corrections, quality-check validations, and user feedback. This observability layer enables governance teams to audit assistant behavior, diagnose failure modes, monitor model performance, and maintain accountability over time.

Finally, the platform supports continuous improvement through structured user feedback capture and semantic layer evolution. End-user validation of helpful insights can be cached and promoted into the knowledge layer, while negative feedback can trigger structured review cycles by dashboard owners. This feedback-driven governance model ensures that the assistant improves iteratively while maintaining enterprise oversight and traceability.

Overall, the governance and compliance framework enables the assistant to function as a scalable and explainable decision-support layer within pharmaceutical analytics environments, balancing innovation with the operational rigor required for real-world adoption.

8. RESULTS AND BUSINESS IMPACT

The implemented conversational BI assistant demonstrated measurable improvements in both analytical efficiency and user experience when compared with traditional dashboard-driven workflows.

Following onboarding and minimal manual refinement of raw table structures, the platform achieved over 90% response accuracy in handling user analytical questions. Accuracy was assessed through structured trace-based evaluation within the LLMOps monitoring framework. A random sample of 1,000 synchronous interaction traces was reviewed, with each trace containing the original user query, agent-to-agent orchestration steps, generated SQL, execution outputs, and recorded user feedback.

Traces were classified as successful when responses were semantically correct, contextually aligned, and positively validated by end users. Based on this evaluation methodology, over 90% of sampled traces were observed to produce satisfactory and correct outcomes. This indicates that the combination of semantic validation, guarded SQL generation, iterative correction cycles, and conversational reasoning can deliver reliable insights across commercial dashboard use cases.

In terms of operational efficiency, user interviews and structured feedback sessions indicated a significant perceived reduction in average query resolution time when compared with conventional dashboard navigation. While no instrumented time-motion study was conducted, end users consistently reported that analytical questions were resolved approximately 30% faster on average, corresponding to an estimated savings of around five minutes per request. These improvements were most evident in workflows that previously required switching across multiple dashboard tabs, manually comparing visuals, or escalating requests to analytics teams.

Beyond speed, the assistant expanded analytical coverage beyond the static limits of predefined dashboard views. Users were able to perform dynamic exploration, including ad hoc aggregations and follow-up investigative questioning, without waiting for backend analyst intervention. This capability enabled more rapid scenario assessment and supported business questions that were not originally anticipated during dashboard design.

Importantly, users also reported increased trust and interpretability due to the assistant's transparency mechanisms. Responses were accompanied by clear summaries, underlying query logic, and stated assumptions, reducing ambiguity and improving confidence in decision-making. The ability to generate contextually appropriate visualizations further enhanced usability by presenting results in familiar analytical formats.

Collectively, these results suggest that conversational augmentation of dashboards can meaningfully accelerate insight-to-action cycles in pharmaceutical analytics. By reducing cognitive effort, minimizing dependency on specialized analyst teams, and enabling flexible real-time exploration, the solution supports faster, more accessible, and more scalable decision-making across commercial organizations.

9. CONCLUSION

Based on the results, the solution reduced average query resolution time by 30%, saving an average of ~5 minutes per request. These gains reinforce the broader finding that adding a conversational layer to existing BI dashboards can make day-to-day analysis faster and more intuitive for both HQ and field teams. By combining automated onboarding, a validated semantic layer, and a text-to-SQL engine, the BI assistant helps users reach the insights they need without digging through multiple dashboard views or relying on analyst support. The improvements seen in accuracy, response speed, and the ability to explore questions beyond predefined visuals highlight the practical value of this approach. Just as importantly, the transparency introduced through semantic validation strengthens user trust and makes advanced analytics more approachable. This approach ultimately offers teams a more straightforward and dependable way to interact with their data and uncover insights when they need them most.

10. NEXT STEPS

- Expand onboarding coverage to support multiple dashboard ecosystems, including Tableau and Power BI, along with structured onboarding of underlying data sources and enterprise reports such as PowerPoint, Word, and PDF documents to enrich contextual understanding.
- Develop a standardized API service layer to expose the validated semantic context for reuse across broader Generative AI applications within the organization, enabling consistent metric interpretation across use cases.
- Introduce a persona-aware “Smart View” that dynamically surfaces the most relevant KPIs and analytical summaries based on the logged-in user’s role, function, and historical interaction patterns.
- Enable dynamic dashboard generation tailored to individual user interests, business priorities, and frequently accessed metrics, allowing personalized analytical workspaces to be assembled automatically from the governed semantic layer.
- Extend the interaction model to support voice-enabled input and output, including real-time speech-to-text query capture and synthesized voice responses. The underlying agentic architecture is designed to accommodate voice-to-voice workflows, enabling future integration into live conversational environments while preserving the same semantic grounding and governance framework.

REFERENCES

1. Using large language models for semantic interoperability: A systematic literature review (https://www.sciencedirect.com/science/article/pii/S24 0595952500092X)
2. An LLM-Based Approach for Insight Generation in Data Analysis (https://arxiv.org/abs/2503.11664)
3. Large Language Model Enhanced Text-to-SQL Generation (https://arxiv.org/pdf/2410.06011)

AUTHORS

Vinay Vihari Lakamsani is a senior data science and GenAI leader with extensive experience driving innovation across pharmaceutical analytics, AI engineering, and commercial data science. With a strong foundation in machine learning, predictive modeling, and LLM-powered solutions, he has led global teams in developing enterprise-scale GenAI platforms, commercial analytics engines, and NBE data science pipelines. Vinay's work spans advanced AI adoption, conversational analytics, MLOps, and autonomous agent frameworks, all focused on transforming decision-making in life sciences.

Mohamed Arruman Shalik is a Senior Data Scientist and GenAI Engineer at Trinity Life Sciences, based in Bengaluru. With a client-facing role, he specializes in designing GenAI architectures and solutions that address complex business problems. Shalik bridges the gap between stakeholders and technology, crafting data-driven solutions that drive impact and innovation across industries

Pulkit Sharma is a Principal Data Scientist experienced in driving Gen (AI)-enabled pharma transformations by bridging business needs with AI solutions.

Mining the Invisible: Overcoming Diagnostic Ambiguity and Prevalence Inflation in Ultra-Rare Disease Patient Identification

Nikhil Jain, Partner, ProcDNA; Rohit Madaan, Director, ProcDNA; Nikhil Borker, Director, ProcDNA; Kanika Maheshwari, Lead Data Scientist, ProcDNA; Tushar Verma, Data Scientist, ProcDNA; Achintye Kamra, Associate Data Scientist, ProcDNA

Abstract: Patient identification in ultra-rare disease through claims poses a fundamental data challenge. With over 5 million probable candidate records and fewer than 1,000 confirmed diagnoses, and acknowledging potential misclassification even within the confirmed pool due to absence of consistent genetic confirmation, standard identification approaches consistently break down. Traditional approaches like rule-based filters miss patients whose journeys are coded inconsistently, while traditional ML inflates patient volumes way over epidemiological evidence. This paper presents a calibrated iterative Positive-Unlabeled (PU) learning framework designed specifically for the low-label, high-noise conditions that typically characterize ultra-rare disease identification problems. The approach combines targeted cohort filtering with an iterative self-learning model that progressively surfaces clinically identical patients from the unlabeled pool, and prevalence calibration to anchor final outputs to real-world epidemiologic expectations. Applied across multiple ultra-rare therapeutic areas, this paper presents findings from one such genetic indication, Autosomal Dominant Hypocalcemia Type 1 (ADH1), where the framework achieved an AUPRC of 72.2% and reduced over-identification by more than 80% thus optimizing business efforts for effective business targeting. This finally yielded a high-confidence cohort of ~12.5K patients, in line with published prevalence estimates. The methodology offers a replicable template for rare-disease signal detection under low-label, high-noise conditions.

Keywords: Ultra-Rare Disease Patient Identification, Positive-Unlabeled Learning, Real-World Evidence, Bayesian Prevalence Calibration, Diagnostic Ambiguity, Claims-Based Patient Detection

1. BACKGROUND

1.1 Rare disease patient identification

Over the past decade, rare disease drug development has undergone a structural shift. What started as a niche segment of biotech startups tracking regulatory incentives has now become a core strategic priority for large pharma. By the end of 2024, an estimated two-thirds of the R&D portfolios at top pharma manufacturers were targeting rare diseases.[1]

Further, by 2026, orphan drugs are projected to account for roughly 20% of total prescription drug sales, growing at nearly twice the rate of non-orphan drugs. The commercial rationale of premium pricing, limited competition, and favorable regulation under the Orphan Drug Act, is well established. But what remains less discussed is that the entire setup depends on something that the industry has not yet fully solved: '*Actually finding these patients*.'

There are an estimated 7,000 to 10,000 distinct rare diseases collectively affecting roughly 30 million Americans.[2] While most don't have an approved therapy, even for conditions that have one, it takes ~4.3 years to confirmed diagnosis from their first medical contact.[3] This time period is lengthy enough to lead to disease progression, narrowed treatment opportunities, and higher healthcare costs without any productive direction. The underlying problem is that the healthcare system in the US generates extensive longitudinal data through claims, prescriptions, and procedure codes. And yet it does not surface the signals that point towards a rare disease patient, given they are typically fragmented, inconsistently coded, and buried beneath far more prevalent comorbidities.

The practical result is a population of patients who are, in a sense, hiding in plain sight. These are individuals who are already in the system, seeing specialists, generating claims, filling prescriptions, but remain unidentified because their symptoms don't add up to a clear diagnosis. Rule-based filters anchored to ICD-10 codes miss them when specific codes are absent or inconsistently applied. Traditional machine learning tends to overcorrect for the highly imbalanced data and inflates its estimated numbers well beyond what epidemiology supports. Neither of the approaches closes the gap between the commercial potential of a rare disease therapy and the practical ability to connect with the right patients at the right time.

The framework described in this paper addresses that gap directly. The work described here develops and validates a calibrated, iterative Positive-Unlabeled (PU) learning framework for rare disease patient identification, one designed specifically for conditions where labeled data is scarce and the unlabeled population is inherently noisy. Model outputs are subsequently evaluated against published prevalence estimates as an external validation measure. The goal is not simply a high-performing model producing a ranked patient list; it is one anchored to epidemiologic priors, producing volumes that hold under clinical scrutiny.

1.2 Exploring unmet needs in existing approaches

As we started to solve this problem, we initially explored various conventional approaches in depth. While each one of them is a reasonable starting point for a classification problem of this type, they had their own limitations, specifically under ultra-rare, low-label conditions.

	Rule-Based / ICD-10 Filters	One-Class Classification	Standard Classification model (all unlabeled = class 0)
How is it Executed	Prevalence assessment of all diagnostic codes and clinical markers to identify the most frequently occurring ones for confirmed patients	One-class SVM to model the impact of various clinical factors in defining the positive class only	A tree-based classification algorithm trained on the patient history of all the patients to predict the likelihood of any incremental patient being positively diagnosed
Core Assumption	Confirmed patients are identifiable through diagnostic codes and clinical markers	Only the positive class needs to be modeled; anomalies from it are suspects	Unlabeled records are true negatives; binary classification applies
Strengths	Transparent, auditable, clinically interpretable	Does not require negative labels; useful when positives are well-defined	Handles class imbalance with minority class weighting; strong predictive performance on standard tasks
How it breaks down in ultra-rare setting	Rare disease codes are often absent, delayed, or substituted with broader unspecified codes. Patients coded inconsistently are systematically missed	The positive class is too small and heterogeneous, while also being homogenous to unknown patients, to reliably define a boundary. Anomaly scores become unstable and clinically uninterpretable	Assigning class 0 to millions of unlabeled records contaminates the negative class with hidden positives. The model learns a distorted decision boundary, inflating predicted volumes well past prevalence expectations
Output failure mode	Under-identification: suspect lists are too small and incomplete to be commercially actionable	Unstable ranking with poor precision at any actionable threshold	Over-identification: suspect pools of 50K–100K+ that destroy clinical trust in the output
Key Challenge	How do you find patients that the coding system itself has not captured?	How do you model a class that is too rare and minimally variable to define cleanly?	How do you train a binary classifier when you cannot trust the negative class?

Table 1: *Comparison of conventional methods for patient identification in rare diseases*

2. OUR SUGGESTED APPROACH

Given the structural constraints outlined above, a conventional formulation is insufficient. In such settings, the core issue is not simply limited data, but the absence of reliably labeled negatives. Positive-Unlabeled learning has been widely applied in domains such as credit risk anomaly detection within the BFSI industry and intrusion or malware detection in cybersecurity where only confirmed positive instances are available and the remaining population cannot be assumed negative.

Accordingly, we pivoted toward a Positive-Unlabeled (PU) learning framework, which is explicitly designed for scenarios where only confirmed positives are available and the unlabeled pool contains a mixture of hidden positives and true negatives. This shift reframes the patient identification problem to align with the true structure of rare disease data, rather than forcing it into a traditional binary classification paradigm.

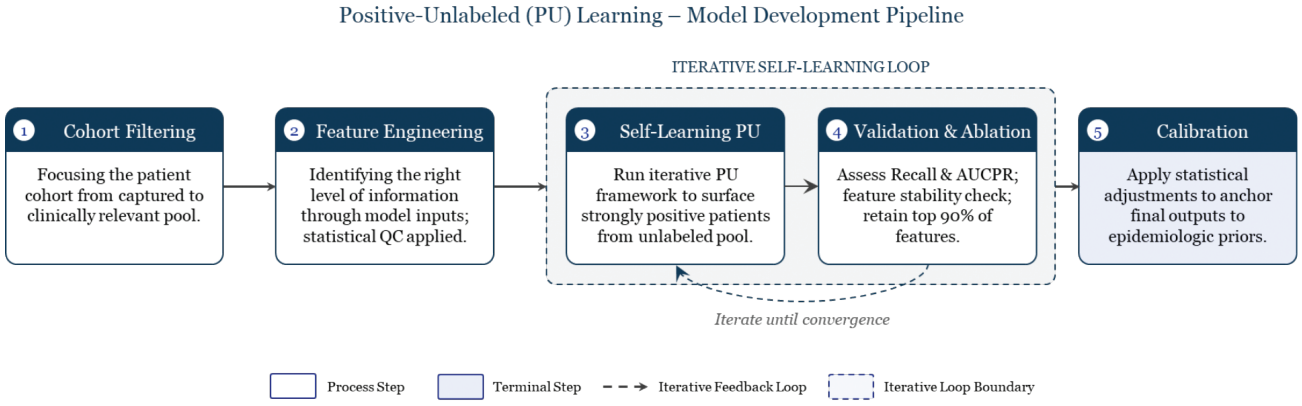


Figure 1. *Positive-Unlabeled (PU) learning framework*

The proposed framework consists of five sequential steps guiding model development from cohort definition to final calibration. The main methodological enhancements occur in feature engineering, framework selection, and prevalence calibration. These stages involve structured experimentation to improve predictive performance, robustness, and alignment with real-world outcome prevalence.

2.1 Step 1 - Cohort filtering:

While a national real-world claims database may capture over 30 million patients, the clinically relevant universe in this context typically represents only a subset of approximately 5 million individuals. The substantial gap between the captured population and the target universe introduces noise, coding variability, and inclusion of clinically irrelevant records, making a single-step modeling approach insufficient to establish a reliable baseline.

By leveraging our clinical understanding of the market, as a first step, we apply clinically relevant business rules to filter specialty codes, procedure markers, and diagnostic adjacencies. This helps reduce the captured data to modeling-ready candidate pool while retaining sufficient breadth to capture undiagnosed patients whose records may be coded inconsistently.

2.2 Step 2 - Feature Engineering:

This is one of the first points where standard practices require revisiting. Within the confirmed patient cohort, we operationalized three complementary representations to systematically derive predictive signals from longitudinal claims data:

- **Recency-based features** to quantify the temporal proximity of specific clinical events to a defined reference point, capturing shifts in activity closer to diagnostic progression
- **Frequency-based features** to measure the accumulation of events within an observation window (±2 years around a clinical anchor), reflecting sustained or repeated clinical engagement
- **Binary presence flags to** record whether a given code or event occurred at any point within the longitudinal window surrounding the first documented signs of relevant concomitant conditions, ensuring signal retention even in sparse or inconsistently coded records

Together, these representations capture the timing, intensity, and occurrence of clinical events within longitudinal claims data, enabling a multidimensional characterization of patient journeys.

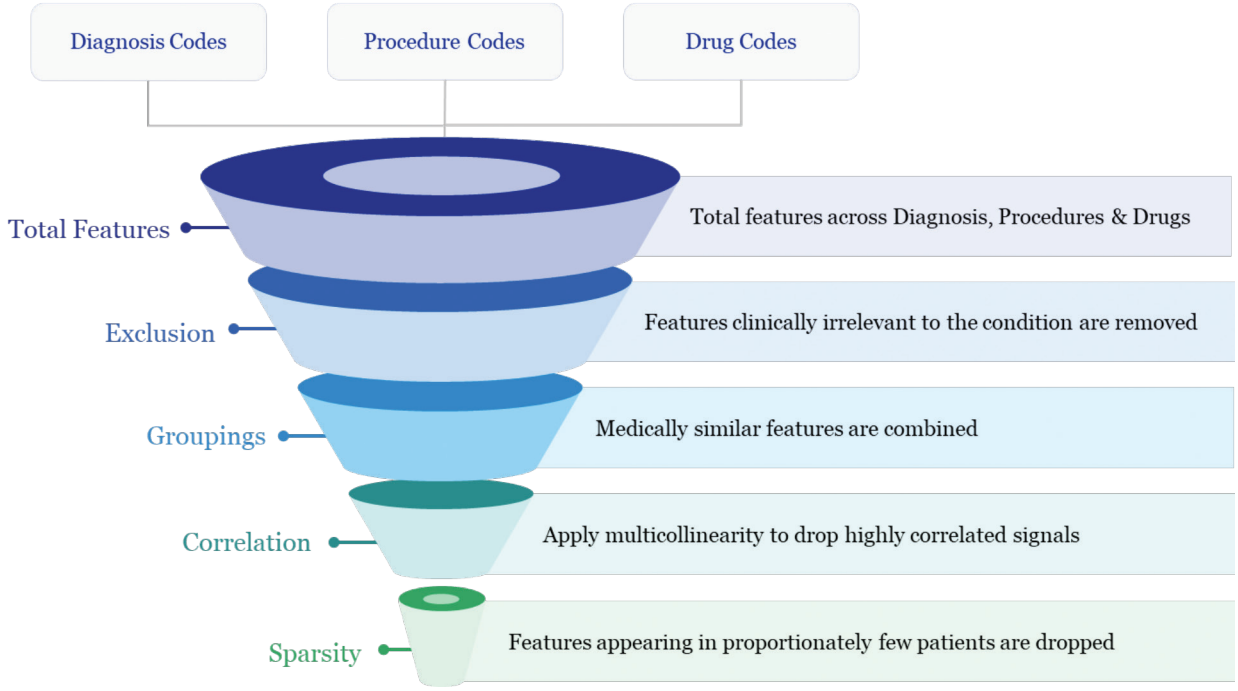


Figure 2. *Feature selection funnel*

At this stage, standard statistical validation checks are performed to ensure data reliability prior to modeling. Fill-rate assessment quantifies feature completeness and identifies sparsity that may introduce instability or bias. Multicollinearity diagnostics detect highly correlated predictors, reduce redundancy and prevent coefficient distortion / variance inflation, thus helping in surface true drivers of diagnosis. Collinearity with the target variable is also evaluated to flag potential data leakage and confirm that predictors represent true pre-outcome signals rather than artifacts of labeling. These checks collectively enhance model stability, validity, and interpretability.

2.3 Step 3 - Multi-stage Positive-Unlabeled (PU) learning framework:

PU learning is a semi-supervised approach designed for exactly the scenario described above - where confirmed positive labels exist, but verified negatives do not. Rather than treating the unlabeled pool as the negative class, PU learning treats it as a mixture: some records are true negative, others are positives that simply have not been labeled yet. The model's job is to progressively separate them.[4]

PU learning is not a single algorithm, but a family of approaches with materially different assumptions. In ultra-rare disease settings characterized by extreme class imbalance, label uncertainty, and limited confirmed positives, these assumptions become critical.

- **Traditional two-step methods** first attempt to isolate a set of reliable negatives from the unlabeled pool using clustering or heuristic separation techniques (e.g., K-means).[5] A standard binary classifier is then trained on confirmed positives and these inferred negatives. This framework assumes a reasonably distinct boundary between positive and negative classes - an assumption often violated when rare disease phenotypes overlap substantially with the broader population.
- **Class-prior estimation and re-weighting** methods treat unlabeled records as weighted negatives. Techniques such as the use of spy instances are employed to estimate the proportion of true positives within the unlabeled pool, after which the classifier's loss function is adjusted using the estimated class prior.[6] While statistically principled, this approach depends heavily on accurate prior estimation and stable weight tuning, which becomes challenging when confirmed positives are few and true negatives are unverified.
- Other more rigorous methods like **Generative/Adversarial Models and Auto-PU** use a generative model to generate synthetic negatives or search through combinations of algorithms and parameters, respectively, require substantial data volume to learn stable representations, limiting their applicability in ultra-rare disease contexts.[7]
- The best-suited solution, hence, is an **iterative self-training framework** which makes no initial assumptions about the composition of the unlabeled pool. It progressively separates positives from the unlabeled cohort by promoting high-confidence predictions into the positive class across successive retraining cycles. This makes it more robust to label noise, absence of validated negatives, and small positive sample sizes - conditions inherent to ultra-rare disease identification.

	Dependence on Clear P–N Separation	Need for Validated Negatives	Sensitivity to Class Prior Estimation	Data Volume Requirement	Suitability for Ultra-Rare Disease Context
Two-Step (Reliable Negative Identification + Classifier)	High	Moderate	Low	Moderate	Limited - unstable when class boundaries are poorly defined
Class-Prior Estimation & Re-weighting	Moderate	Low	High	Moderate	Moderate -sensitive to prior estimation with specially with few confirmed positives
Generative / Adversarial / Auto-PU Methods	Low	Low	Moderate	High	Low - requires substantial data to learn stable distributions
Iterative Self-Training Framework	Low	None	Low	Low–Moderate	High – robust to label noise, imbalance, and limited positives

Table 2: *Comparative Assessment of PU Learning Approaches for Ultra-Rare Disease Identification*

Execution of iterative self-training PU-framework using XGBoost:

- Iteration 0 — Baseline model: A 3-fold cross-validated XGBoost classifier trained on a subset (80% of the sample retained for training and validation, 20% holdout for testing) confirmed cases against a random draw of unlabeled records. This initial pass is deliberately conservative as its output is a rough probability ranking, not a final score
- Iterations 1 to N — Probable positive promotion: Unlabeled records scoring at or above 98% probability are reclassified as probable positives and folded into the positive training set for the next cycle. This strict promotion threshold acts as a natural control on how aggressively the positive set grows and ensures we do not end up adding noisy labels. With each iteration, the unlabeled sample is re-drawn randomly rather than reused, which prevents the model from learning patterns tied to specific institutional or regional coding practices rather than the underlying disease

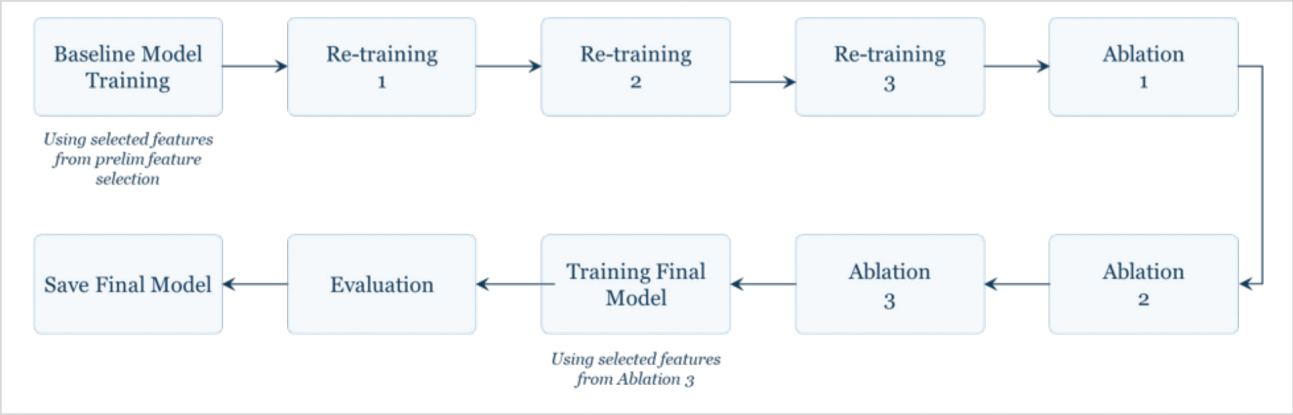


Figure 3. *Flowchart representing iterative self-training process*

How to determine N: number of iterations?

- Validation approach: Standard accuracy metrics are misleading in a PU setting because the true composition of the negative class is unknown. Two alternative validation strategies are recommended to be used: Recall and AUPRC on a held-out set of confirmed patients never shown during training, with recovery rate from the unlabeled pool serving as the primary success criterion. Secondly, feature importance rankings are tracked across iterations and reviewed with clinical input to ensure consistency with established disease biology. This helps confirm that identified drivers reflect meaningful clinical patterns rather than statistical artifacts.

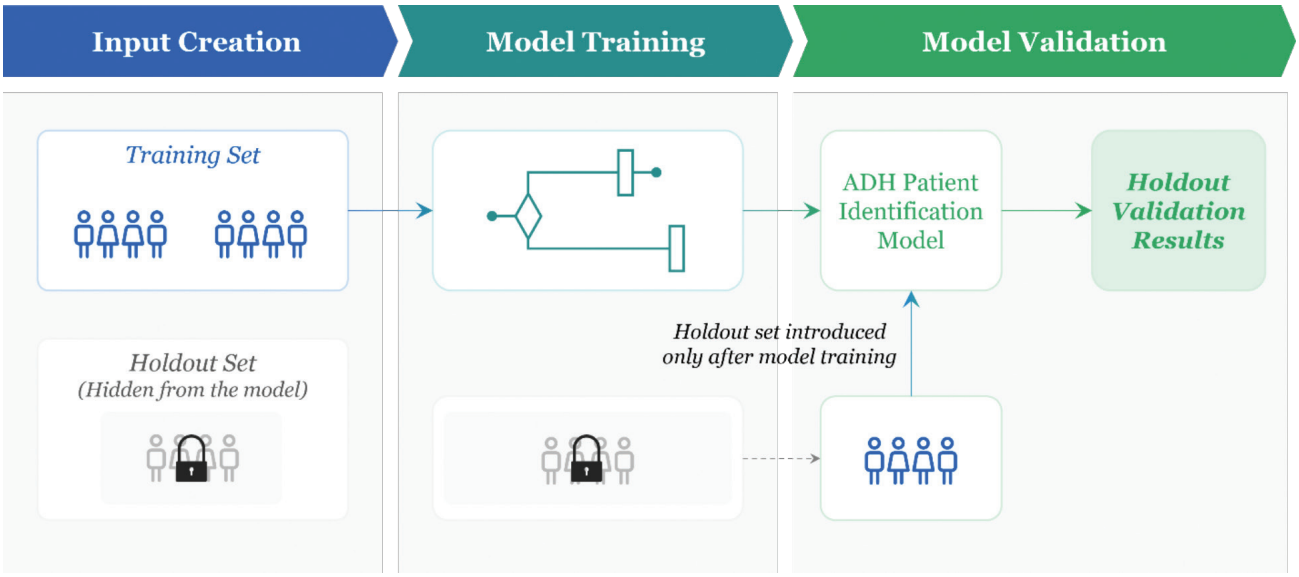


Figure 4. *Model Holdout Validation Framework*

- Stability Monitoring: These two signals were tracked across iterations to confirm the model was deepening its clinical understanding rather than drifting toward noise. When both the **Recall & AUPRC** on a held-out validation set, and the rank-order of the top feature drivers stabilize or improves, the cycle is considered successful. Once stability was confirmed across iterations, the loop was stopped there

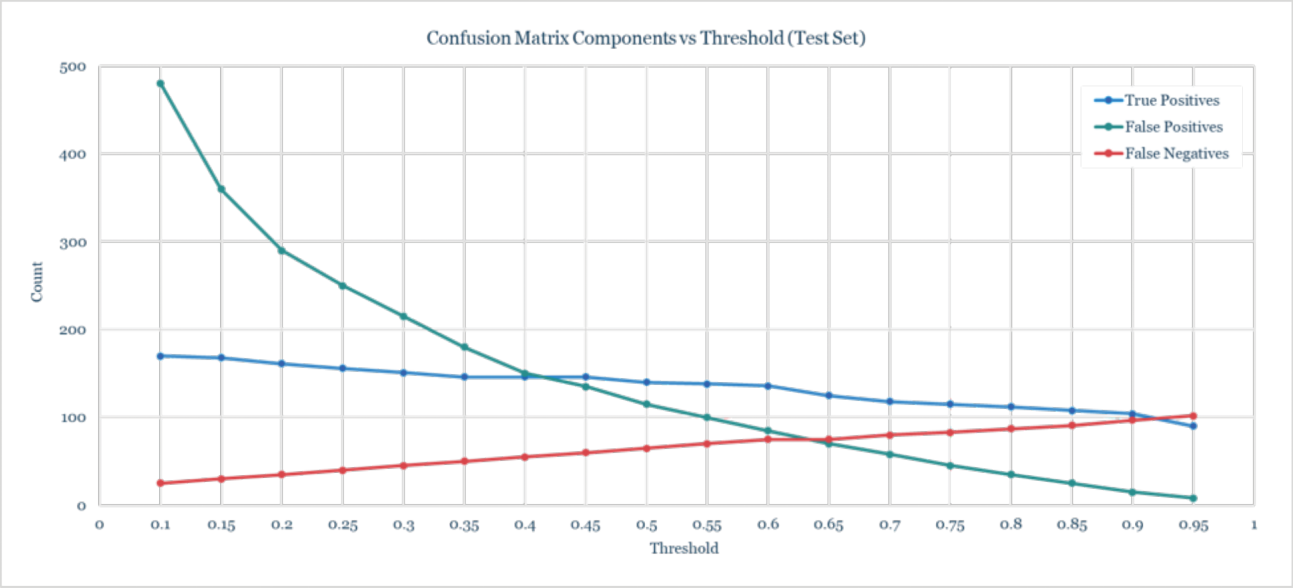


Figure 5. *Analysis of confusion matrix components across probability thresholds*

Collectively, the proposed PU learning framework is designed to address the structural limitations of conventional patient identification approaches. By relaxing strict labeling assumptions and iteratively refining the positive class, it provides a more stable and inclusive identification strategy. Table 3 summarizes how this framework directly mitigates the key challenges outlined earlier in Table 1.

Challenge from prior approaches	How PU Learning addresses it
Rule-based filters miss patients with inconsistent or absent coding	PU framework learns from clinical patterns across the full longitudinal record, not from the presence of a specific code. It also incrementally adds patients who were misdiagnosed or miscoded but show all clinical similarities with known patients
One-class boundary is unstable with very few positives who closely resemble negatives	Given an extent of homogeneity with the negative class, the boundary is refined iteratively as more probable positives are surfaced
Standard classification model contaminates the negative class with hidden positives	Unlabeled records are never assumed to be negative. They remain candidates until the model assigns them a low enough probability across iterations

Table 3. *Mitigation of conventional modeling challenges via PU learning*

2.4 Step 4 & 5 - Validation and Calibration

2.4.1 Feature Ablation:

Running ablation on an early-stage model risks discarding features that only become meaningful once the training set has been enriched through other uncaptured patients of interest. Ablation was therefore reserved until feature importance rankings had stabilized across iterations, and model performance metrics showed consistency. The final ablation retained the top 90% of cumulative feature importance, producing a compact and interpretable final feature set.

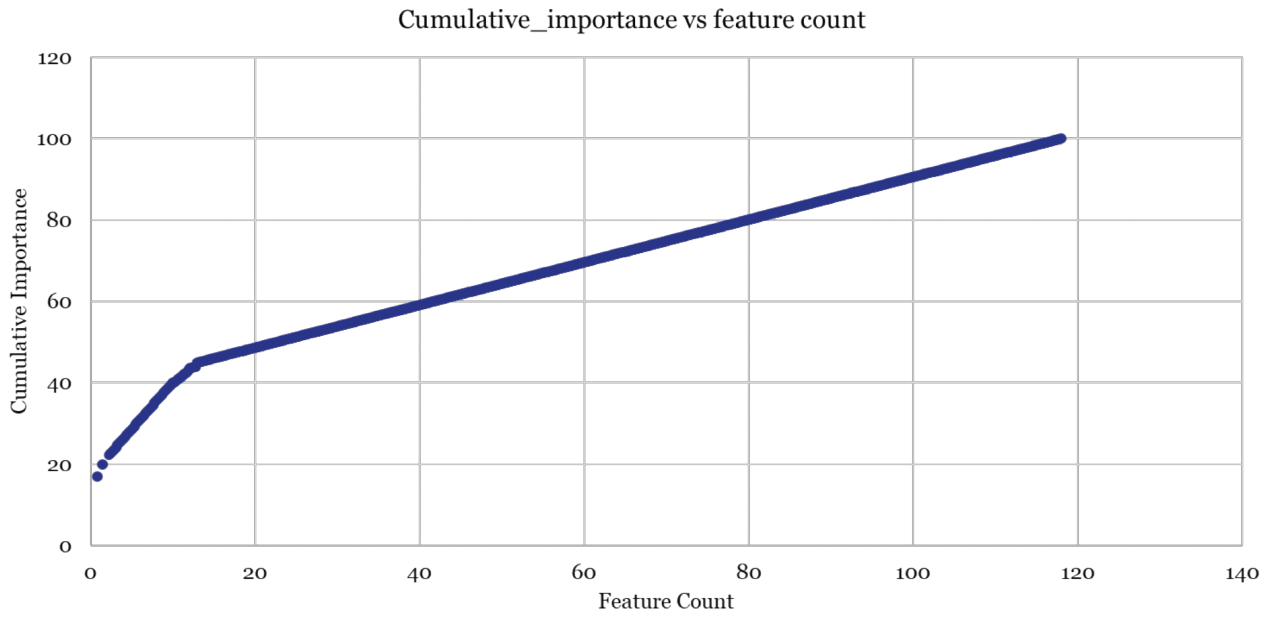


Figure 6. *Tracking cumulative feature importance*

2.4.2 Probability Calibration

Raw XGBoost probability scores in this setting are systematically overconfident. The model is trained on an enriched sample where the positive-to-negative ratio is far higher than in the real world. Hence, the model scores do not inherently reflect actual population-level likelihood. Left uncorrected, even a well-performing model produces a predicted patient pool that is epidemiologically implausible. Three correction approaches were evaluated:

	XGBoost minority class weighting	Elkan-Noto Calibration	Bayesian Prevalence Correction
Mechanism	Scales the minority i.e. positive class weight during training to adjust the decision boundary	Divides raw scores by c, the estimated probability that a true positive is labeled, derived under the "selected at random" assumption	Applies Bayes' Theorem to update raw posterior scores using an external epidemiologic prior on disease prevalence
What it corrects	Class imbalance during training, not post-hoc score distribution	Post-hoc score inflation under PU assumptions	Post-hoc score inflation anchored to real-world prevalence
Key limitation	Adjusts the model's sensitivity but does not constrain output volume to prevalence expectations. Still produced suspect pools well above epidemiological estimates	Relies on the "selected at random" assumption - that the labeled positives are a random draw from all true positives. In rare diseases, diagnosed patients are often the most severe or clinically obvious cases, violating this assumption and biasing the constant c	Requires a reliable external prevalence estimate. If literature underestimates true prevalence, correction may be too aggressive
Output in this study	Over-identification persisted; not viable as a standalone calibration	Partial correction but unstable across threshold ranges due to assumption violation; unsuitable for clinical data	Produces a final cohort of patients more consistent with published prevalence estimates

Table 4. *Comparison of different calibration methods*

Bayesian correction is recommended for the final framework. For clinical data where sufficient published epidemiologic literature is available to establish a defensible prevalence prior, this method's primary limitation is manageable.[8] The correction mathematically penalizes raw scores by factoring in the ratio of known epidemiologic prevalence to the prevalence observed in the training sample, pulling the score distribution down until the resulting patient volume falls within a clinically plausible range.

$$p_{adjusted} = \frac{p_{model} \times \frac{p_{true}}{p_{train}}}{p_{model} \times \frac{p_{true}}{p_{train}} + (1 - p_{model}) \times \frac{1 - p_{true}}{1 - p_{train}}}$$

Figure 7. *Bayesian prevalence correction equation*

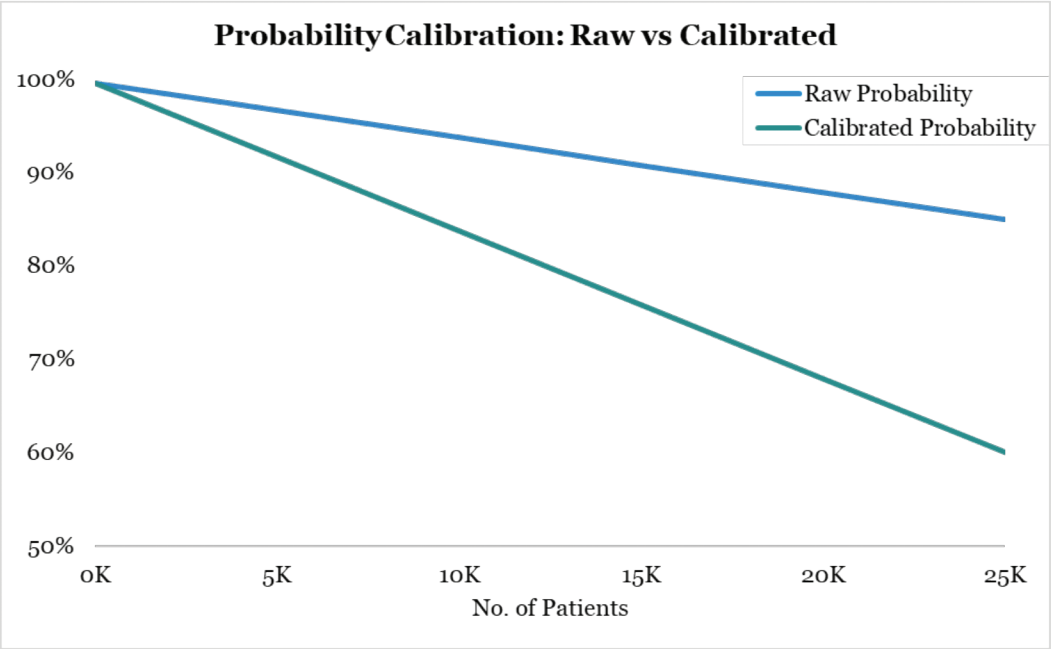


Figure 8. *Distribution of predicted patients using raw probabilities v/s calibrated probabilities*

3. CASE STUDY: PATIENT IDENTIFICATION IN AN ULTRA-RARE GENETIC DEFICIENCY

Autosomal Dominant Hypocalcemia Type 1 (ADH1) is an ultra-rare genetic disorder caused by activating mutations in the *CASR* gene, characterized by hypocalcemia with inappropriately low parathyroid hormone levels. Published prevalence estimates range from approximately 1 per 70,000 to 3.9 per 100,000 individuals,[9] placing it firmly within the ultra-rare category. In large real-world claims databases, this translates into extremely sparse confirmed diagnoses, with many affected patients potentially misdiagnosed or coded under broader hypocalcemia-related conditions.

Patient identification in ADH1 is particularly challenging because its clinical presentation overlaps with more common calcium metabolism disorders, and definitive genetic confirmation is rarely captured in administrative claims data. Genetic testing may not be performed routinely, may not be reimbursed or coded explicitly, and laboratory results are often not directly available within claims databases. As a result, reliance on diagnostic codes alone can lead to both under-identification (missing true cases) and over-identification (capturing clinically similar but biologically different conditions).

The framework described in the preceding sections was applied in the context of ADH1, an ultra-rare genetic deficiency characterized by low prevalence, diagnostic ambiguity, and heterogeneous coding patterns. This setting provides a stringent evaluation of whether calibrated PU learning can recover clinically plausible patient cohorts where conventional supervised approaches have proven inadequate. We partnered with a pharmaceutical manufacturer to identify patients with ADH1 in large-scale claims data, applying the framework in a real-world ultra-rare disease setting.

3.1 Data Architecture

- The candidate universe was constructed from a national real-world claims database, starting at roughly 5.2 million records. To get to this reduced pool that was ready for modelling, cohort filtering employed clinical business rules anchored to specialty codes, pertinent procedure markers, and diagnostic adjacencies. Since the first documented diagnosis or treatment of a related nutrient deficiency was usually the earliest detectable signal in the administrative record for patients on this disease trajectory, the longitudinal window for each patient was constructed as a four-year clinical history around that anchor point.
- Rather than using frequency or recency, feature construction used in the binary flags format outlined in Section 2.2: presence or absence of clinical markers during the four-year period. Features that appeared in less than 2% of records were eliminated by fill-rate filtering, and Spearman correlation screening at a 0.75 threshold

3.2 Modeling Execution

- Starting from a baseline of more than 950 confirmed cases, the iterative PU loop further ran three cycles. In each cycle, unlabeled records that met the 98% probability threshold were moved to the positive training set, and the unlabeled sample was randomly redrawn for each iteration. To ensure that majority class dominance did not distort performance metrics, a custom hyperparameter tuning function maintained a fixed 20:80 positive-to-negative ratio across all validation folds.
- By 3rd iteration on top of the baseline, feature importance rankings stabilized, and three rounds of ablation pruned the feature set down to the top cumulative 90% of drivers. The candidate pool was reduced from 5.2 million to roughly 12,500 high-confidence patients by applying Bayesian prevalence correction to the final model scores.

3.3 Clinical Alignment of Results

- This is where the case study gets interesting. Going in, the clinical team's expectation was straightforward: calcium deficiency-related codes would dominate feature importance, given that hypocalcemia is the most visibly shared characteristic between confirmed patients and the broader look-alike population. That turned out to be an incomplete picture.
- The model did surface calcium-related markers, but they ranked lower than anticipated precisely because they were too common across the unlabeled pool to provide meaningful separation. What drove the model's discrimination more effectively were markers associated with chronic kidney disease (CKD) and a cluster of secondary metabolic indicators that reflect the downstream systemic consequences of the genetic deficiency rather than its most obvious surface presentation. These features made clinical sense once surfaced as CKD is a well-documented comorbidity pathway in this condition, but they would not have been the starting point for any rule-based system built around domain intuition alone. This is the practical value of letting the model find the signal rather than encoding what clinicians expect the signal to be.

3.4 Results

At a deliberately stringent classification threshold of 0.75, the model produced the following results:

Metric	Value
AUPRC	72%
Precision	74%
Recall	60%
Final high-confidence cohort	~12,500 patients
Over-identification reduction vs. uncalibrated output	>80%

Table 5. *Model Performance Metrics & Outputs*

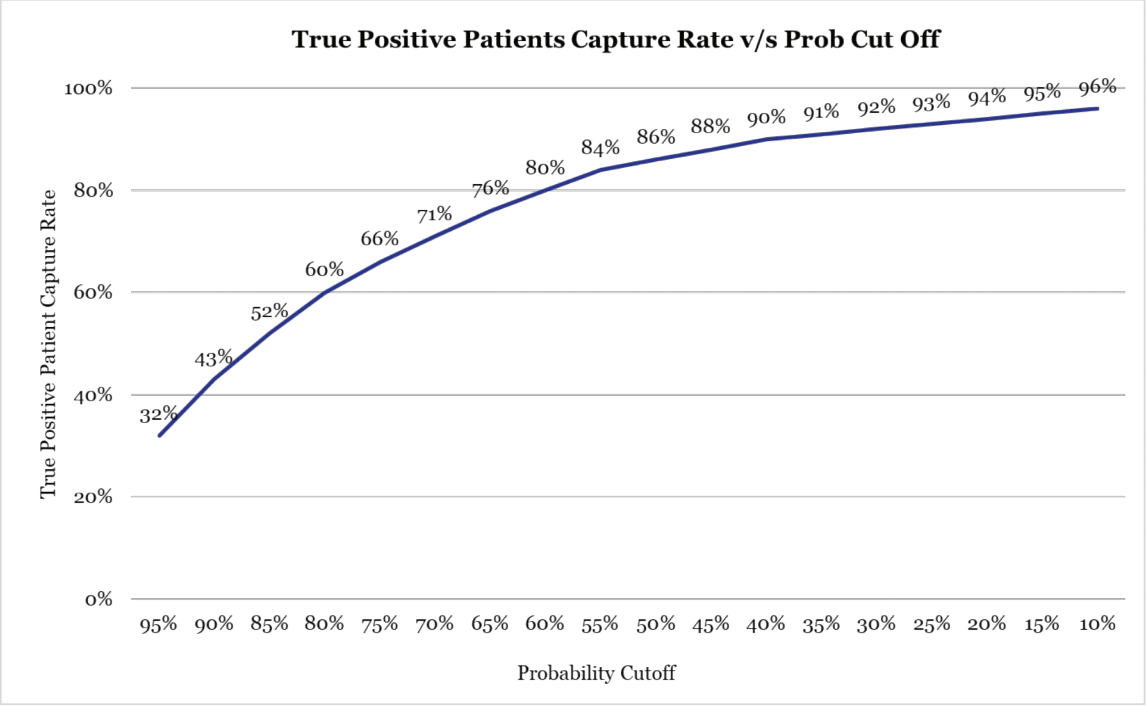


Figure 9. *Percentage of True Positive patients captured at different probability thresholds*

3.5 Insights

- Holding a 0.75 threshold and still recovering 60% of known patients makes the result highly reliable. In a standard supervised setting with this degree of class imbalance, that combination of precision and recall at a high threshold typically does not hold. The designed framework's baseline data preparation, iterative positive expansion with the right calibration is what makes it possible
- Across other therapeutic areas where this framework has been applied, recall and AUPRC have stayed within roughly ±3% of these figures at comparable thresholds. The consistency is not coincidental, it reflects the fact that the underlying data problem of extreme label scarcity, contaminated unlabeled pool, prevalence constraints, is structurally similar across ultra-rare conditions, even when the clinical context differs substantially.

---

4. CONCLUSION AND FUTURE SCOPE

- Patient identification in ultra-rare diseases has historically been treated as a data access problem with core idea being to get better data, find more patients. What this work suggests is that the bottleneck is less about access and more about how the available data is used.
- The framework described here closes a meaningful portion of that gap. By rejecting the assumption that unlabeled records are true negatives, iteratively surfacing probable positives from a noisy and heterogeneous pool, and anchoring final outputs to epidemiologic priors rather than raw model scores, the approach produces patient cohorts that are both analytically defensible and clinically actionable.
- On the question of where it goes next, the most immediate extension is moving from identification to prioritization. The current framework produces a ranked patient list; it does not yet model which patients are most likely to benefit from early intervention, or which are at highest risk of further diagnostic delay. Integrating treatment pathway data and patient journey sequencing into the feature space is a natural next step and one already underway in follow-on work. While in longer term, the framework's structure is also well suited to settings beyond rare disease identification, including Rx switching prediction and early treatment discontinuation modeling, where labeled data is limited but the unlabeled population carries a recoverable signal.
- Beyond prioritization, a further extension lies in strengthening causal attribution and clinically grounded feature interpretation. While the current framework identifies probable patients, deeper causal effect analysis combined with medically validated feature importance assessments can help distinguish correlation from actionable clinical drivers. This would enable clearer understanding of the determinants of delayed diagnosis and potential intervention points. For pharmaceutical manufacturers, such insights extend beyond identification toward evidence-based patient support strategies and more informed deployment of educational and engagement efforts.

REFERENCES

1. Getz K. Applied Clinical Trials. The hard truth about rare disease and gene therapy drug development: https://www.appliedclinicaltrialsonline.com/view/the-hard-truth-about-rare-disease-and-gene-therapy-drug-development

2. US Food and Drug Administration. Rare diseases at FDA: https://www.fda.gov/patients/rare-diseases-fda

3. Faye, F, Crocione, C, Anido de Peña R, *et al.* Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a Rare Barometer retrospective patient survey. *Eur J Hum Genet* 32, 1116–1126 (2024).

4. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. Mach Learn. 2020;109(4):719-760.

5. Liu B, Dai Y, Li X, Lee WS, Yu PS. Partially supervised classification of text documents. In: Proc 19th Int Conf Mach Learn. 2002. p. 387-394.

6. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proc 14th ACM SIGKDD Int Conf Knowl Discov Data Min. 2008. p. 213-221.

7. Saunders JD, Freitas AA. Ga-auto-PU: a genetic algorithm-based automated machine learning system for positive-unlabeled learning. In: Proc Genet Evol Comput Conf Companion. 2022. p. 288-291.

8. Saerens M, Latinne P, Decaestecker C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Comput. 2002;14(1):21-41. doi: 10.1162/089976602753284446.

9. Roszko KL, Stapleton Smith LM, Sridhar AV, et al. Autosomal Dominant Hypocalcemia Type 1: A Systematic Review. J Bone Miner Res. 2022;37(10):1926-1935. doi:10.1002/jbmr.4659

ABOUT THE AUTHORS

Nikhil Jain is a Partner with ProcDNA where he leads the Analytics & Data Science Excellence Center, focusing on solving challenging healthcare problems based on a variety of datasets using advanced analytics and machine learning methods. He has 14+ years of experience working with the Pharmaceutical, Bio-Tech, and Medical Devices industries. He has multifaceted expertise in areas including product launch preparedness, sales force effectiveness, commercial excellence, training, and strategic planning. He received his MBA from the Indian School of Business.

Rohit Madaan, Nikhil Borker, Kanika Maheshwari, Tushar Verma and Achintye Kamra represent the Analytics & Data Science Excellence Center in ProcDNA. The team is an expert in designing and delivering AI/ML-enabled innovative healthcare solutions such as Patient Identification, KOL Identification, Determining Referral Pathways, Patient Event Prediction, Next-Best-Action (NBA), and Identifying Early Adopters.

The Next Decade of Forecasting: A Learning System Integrating Commercial, Medical, and Access Analytics

Anindya Roy | Vipul Vaid | Vikrant Kumar | Himanshu Gurnani

VISCADIA

Abstract: Pharmaceutical forecasting has grown substantially more comprehensive over the past decade, yet its organizational architecture has not kept pace. Medical, access, and commercial analytics functions continue to operate in relative isolation, producing outputs that are too often disconnected in ways that compound into material commercial underperformance.

This article argues that the most consequential forecasting improvements over the next decade will come not from upgrading core models, but from transforming the architecture within which those models operate. Specifically, pharmaceutical forecasting must evolve from a periodic planning deliverable into a learning system — an integrated decision architecture defined by shared data foundations, structured assumption governance, and cross-functionally coherent scenarios.

Drawing on patterns observed across oncology, specialty, and rare disease product launches, the article diagnoses three recurring silo failures and introduces a framework — organized around medical intelligence, access analytics, and commercial analytics — for addressing them. It proposes that Agentic AI and LLM-based orchestration serve as enabling infrastructure within this architecture, not as a substitute for it. Four forward-looking proposals define the path forward: continuous forecasting, cross-functional ownership, AI-native assumption governance, and early asset-level convergence.

Organizations that build this architecture will not just improve singular forecast cycles but will compound their decision intelligence over the lifecycle of every asset they commercialize.

1. THE FORECASTING PARADOX

Pharmaceutical forecasting has never been more sophisticated — more granular models, more abundant data, more specialized talent. And yet, roughly half of new pharmaceutical product launches continue to miss their pre-launch commercial forecasts, with actual peak sales deviating from projections by as much as 71% and forecasts remaining materially off target even six years post-launch.[1]

This is not primarily a model failure. The core analytical machinery — patient-based epidemiological forecasting, segment-level uptake modeling, gross-to-net revenue adjustment — is well-established. The failure is architectural: as the commercialization environment has grown more complex, the organizational structures within which forecasting operates have not evolved at commensurate pace.

The science has accelerated: biomarker-driven approvals, label restrictions defined by diagnostic tests, accelerated pathways, and cell and gene therapy modalities have compounded the clinical complexity forecasters must navigate. Market access has grown more consequential and fragmented: payer formulary restrictions, indication-specific contracting, prior authorization requirements, and specialty pharmacy dynamics now shape time-to-therapy and net revenue in ways traditional launch analogs cannot predict. The commercial landscape has become more competitive, data-rich, and real-time.

Each development has driven investment within specific analytical functions — real-world evidence capabilities, payer intelligence, adoption and segmentation models. The investments are real, and the capabilities have improved.

The problem is that these investments have been made within silos. The decisions that forecasting must support — launch sequencing, indication prioritization, resource allocation, contracting strategy — require intelligence no single silo can provide. The forecasting gap is not a question of intra-function quality. It is a question of inter-function integration.

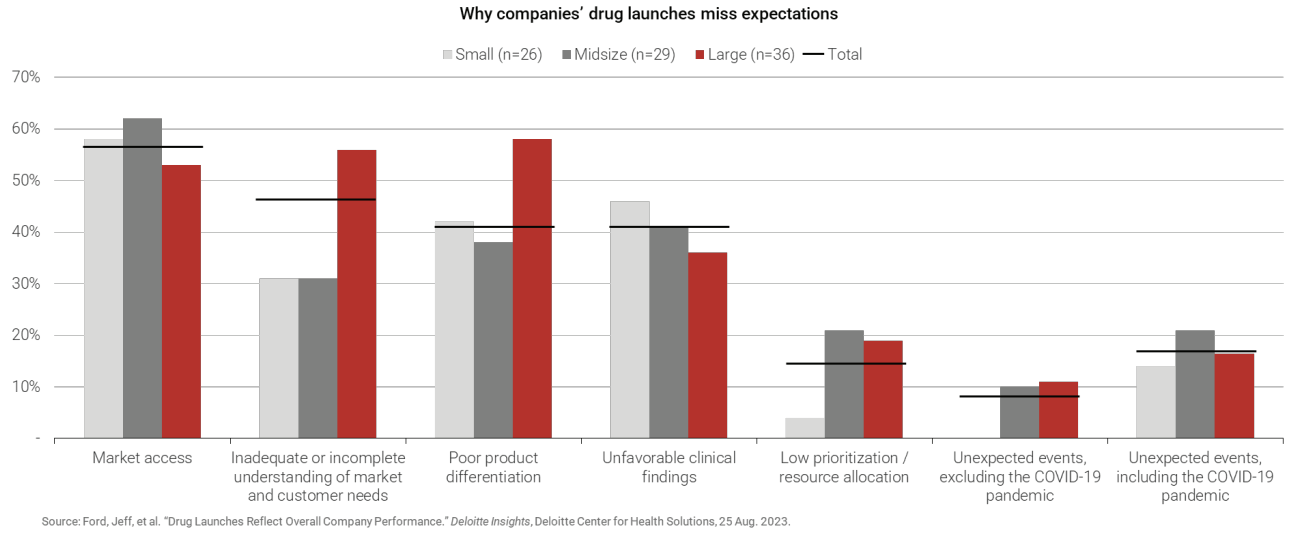


Figure 1. *Attributable Reasons for Missed Expectations*

> *The most consequential forecasting improvements over the next decade will not come from replacing or upgrading core models. They will come from transforming the architecture within which those models operate - specifically, from replacing siloed analytical functions with an integrated decision system that learns continuously.*

2. ONE ASSET, THREE BLIND SPOTS: THE COST OF FRAGMENTATION

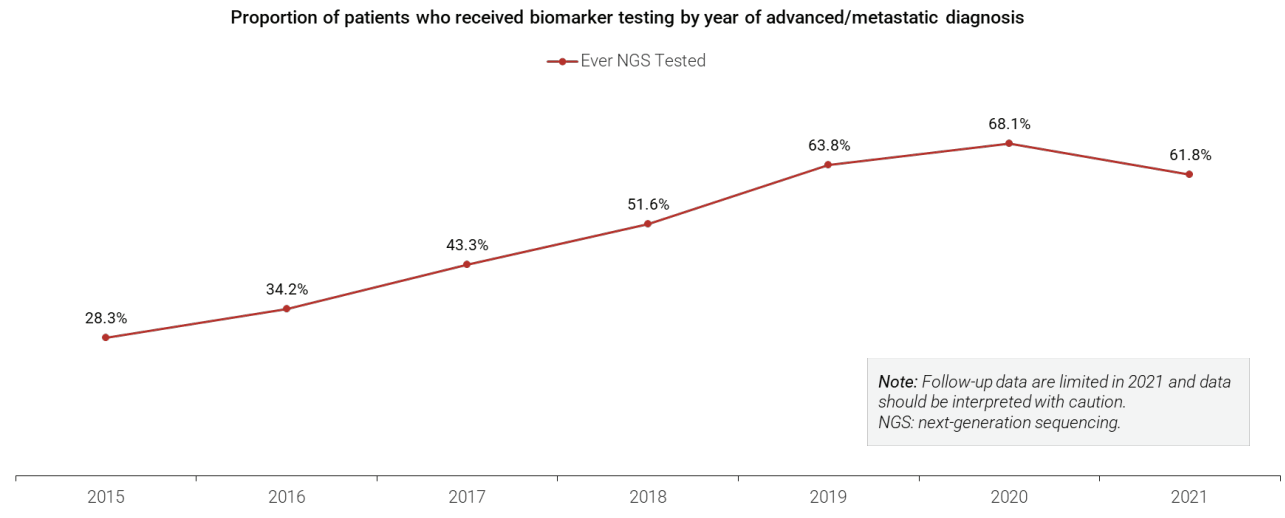
To ground this argument in commercial reality, let us consider a representative pipeline asset — a biomarker-selected therapy entering a competitive second-line specialty indication through a specialty pharmacy distribution model. The scenario is illustrative but constructed around well-documented patterns. Let us call it Asset X.

Asset X is a PD-L1/biomarker-selected agent in a crowded NSCLC-adjacent indication, with a restricted label, specialty pharmacy channel, and active step-edit requirements across major PBMs. The commercial team has built a rigorous market model. Medical analytics has generated real-world evidence on the eligible population. The access team has developed detailed coverage assumptions. Each team's work is internally sound — yet the collective picture is fundamentally incomplete in three predictable ways.

2.1 Gap One: Medical and Commercial Analytics Are Reasoning from Different Denominators

The commercial model begins with an eligible patient pool derived from epidemiology estimates and historical claims. Meanwhile, medical analytics is tracking something the model has not incorporated: biomarker testing penetration rates in real time, across practice settings.

This gap is well-documented. A 2022 Flatiron Health analysis of 17,513 patients found that NGS-based testing grew from 28% in 2015 to 68% in 2020, with community practices lagging academic centers and dependent on external commercial laboratories. The MYLUNG Consortium's prospective study of 12 community practices found that only 37% of metastatic NSCLC patients had results for all nine recommended biomarkers at the time of treatment initiation[2,3] settings were not tested for core actionable markers, with substantial interphysician variability that correlated directly with survival outcomes.[4]



Source: Hess, Lisa M., et al. "Biomarker Testing for Patients With Advanced/Metastatic Nonsquamous NSCLC in the United States of America, 2015 to 2021." *JTO Clinical and Research Reports*, vol. 3, no. 6, 2022, 100336.

Figure 2. *Growth in Biomarker Testing Rate in NSCLC Patients*

These are not marginal variations. For a biomarker-selected asset, the tested-and-positive population is the addressable market. A forecast built on stable epidemiological counts that does not account for variable testing penetration is structurally misaligned with the actual denominator. The divergence surfaces at a brand review, by which time field deployment and resource allocation have been calibrated to an incorrect patient count.

2.2 Gap Two: Access Analytics and Commercial Analytics Are Not Speaking to Each Other

Asset X's commercial forecast projects uptake based on historical launch analogs — a standard methodology. What it does not reflect is the access environment the access team is observing in real time.

The access team is watching two of three major PBMs place Asset X on a non-preferred tier with an active step-edit requirement — adding three to six weeks to time-to-therapy for a significant proportion of patients and suppressing early refill rates in the initial quarters. A 2024 ASCO analysis documented that prior authorization for cancer treatments introduced clinically relevant delays, with most payer response times exceeding five days and nearly half of adult oncology patients facing more than a week's delay in medication approvals. The same analysis found that a significant majority of denied requests were ultimately approved on appeal — confirming that the burden represents access friction, not clinical inappropriateness.[5]

The gross-to-net implications are equally material. Drug Channels Institute estimates the total brand-name gross-to-net bubble at $356 billion in 2024. For specialty oncology products, the gap between list and net price is routinely structured by formulary tier placement, rebate commitments, and coverage restrictions — variables the access team is tracking and the commercial model is not structurally reflecting.[6]

2.3 Gap Three: Medical and Access Analytics Are Compounding Each Other's Blind Spots

The third failure is the most structurally important. In a rare disease variant of the Asset X scenario, the medical and access assumptions that drive realized opportunity interact in ways neither team can model independently.

Medical analytics tracks the patient-finding journey: diagnosis rates, testing rates, time from symptom onset to confirmed eligibility. Published evidence is clear: a 2024 EveryLife Foundation study estimated an average diagnostic journey of more than six years across rare diseases, with 17 clinical encounters between onset and diagnosis. A European rare disease survey of 6,507 patients in 41 countries found an average total diagnosis time of 4.7 years, with 56% diagnosed more than six months after first medical contact.[7,8]

Access analytics, meanwhile, tracks the coverage journey: formulary timelines, prior authorization approval rates, specialty pharmacy fill and abandonment rates. These are not independent of the patient-finding journey — they compound it. A patient who has waited years for diagnosis reaches treatment initiation precisely when access friction is highest.

The result is a realized opportunity constrained by two compounding delays that produce an outcome worse than either alone. Only a model integrating medical patient-journey logic with access friction data can anticipate and quantify it.

2.4 The Structural Pattern

Asset X does not fail. But it underperforms in ways retrospectively traceable to structural fragmentation — to the absence of an architecture that would have made these blind spots visible before they became commercial deficits. The pattern repeats across biomarker-driven oncology launches, specialty launches with unmodeled formulary dynamics, and rare disease launches where diagnostic and coverage delays compound unobserved.dynamics are unmodeled in the commercial forecast; across rare disease launches where the diagnostic journey and the coverage journey compound unobserved.

These failures are the predictable consequence of an architecture designed for functional optimization rather than integrated decision support.

> ***On Adjacent Signal Sources***
>
> *Two additional functions merit explicit recognition as underrepresented in most forecasting architectures: Health Economics and Outcomes Research (HEOR), whose value evidence directly shapes formulary positioning and payer negotiations; and Patient Services, whose co-pay assistance utilization, hub performance, and adherence program dynamics directly affect persistence and net realized volume. An optimal learning system must integrate these adjacencies, not treat them as downstream adjustments. Structurally though, these could be integrated with Medical, Access, and Commercial Analytics for the sake of simplicity (i.e., limiting it to 3 functional silos).*

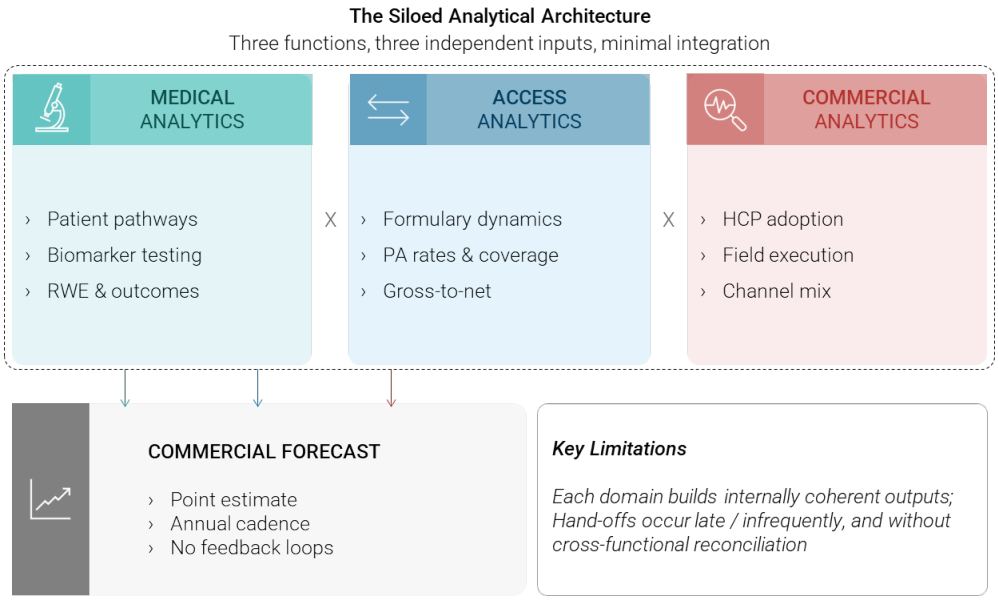


Figure 3. *The siloed analytical architecture: three independent tracks converging only at the forecast node, without shared inputs or feedback loops.*

3. THE LEARNING SYSTEM: A DECISION ARCHITECTURE FOR INTEGRATION

The response to structural fragmentation is not a new model — it is a new decision architecture. The diagnosis in Section 2 points to a specific organizational failure: analytical intelligence flows within functions but not between them. The remedy is equally specific: building infrastructure to enable intelligence to flow across all three domains simultaneously, continuously, and in usable form.

Forecasting as a learning system is the operating model that achieves this. A learning system is not a technology platform or AI deployment. It is a set of organizational and process design choices governing how evidence, assumptions, and scenarios are updated and shared across functions. Three structural components define it.

3.1 Multi-Source Signal Integration

The foundation is a shared data layer — a common analytical substrate that all functions access and contribute to. This does not require a single unified database. It requires aligned definitions: common patient eligibility criteria, consistent treatment pattern sources, agreed denominators, and shared biomarker and diagnostic data.

Without this substrate, each function reasons from a different denominator. The commercial team's 'eligible patient' and the medical team's 'tested-and-positive patient' are not the same number. A shared definition layer resolves this by ensuring the population entering each model is counted the same way, from the same sources.ensuring that the population entering each team's model is the same population, counted the same way, updated from the same sources.

3.2 Structured Assumption Governance

The second component — arguably the most operationally important — is a shared assumption library. Assumptions about eligibility, time-to-therapy, formulary coverage, biomarker adoption, and uptake velocity are currently dispersed across function-specific models and planning documents, with no single source of truth.

A shared library makes key assumptions explicit, cross-functionally owned, version-controlled, and accompanied by documented rationale. When medical analytics updates biomarker testing assumptions, the commercial team sees it. When access revises formulary timelines, the commercial model reflects it without waiting for the next planning cycle.

This governance model also enables conflict resolution. When commercial assumes 40% share in a setting where access assumes 25% covered lives penetration, the inconsistency is visible and subject to explicit resolution rather than hiding across separate models.

3.3 Continuous Scenario Refinement

The third component is a shift from calendar-driven to decision-driven scenario management. In a traditional model, scenarios are built once per planning cycle and remain static. In a learning system, scenarios are continuously refinable — updated when evidence thresholds are crossed, not when the planning calendar permits.

This requires scenarios designed with cross-functional coherence from the start. A scenario is not a single-variable sensitivity — it is a named, internally consistent representation of a real-world future in which commercial, medical, and access assumptions all reflect plausibly co-occurring conditions.

3.4 The Role of Agentic AI

Agentic AI and LLM-based orchestration are increasingly capable of supporting this infrastructure: mining documents for assumption-relevant evidence, flagging cross-functional inconsistencies, automating scenario documentation, and accelerating the data profiling work that underpins shared signal integration. These capabilities are real and growing, and their application to forecasting workflows is an important emerging practice.

However, AI is infrastructure within the learning system — not the learning system itself. An organization deploying AI onto a fragmented assumption architecture will automate fragmentation rather than resolve it. The governance model must be in place first; AI then makes it scalable and auditable at a pace human-only processes cannot sustain.

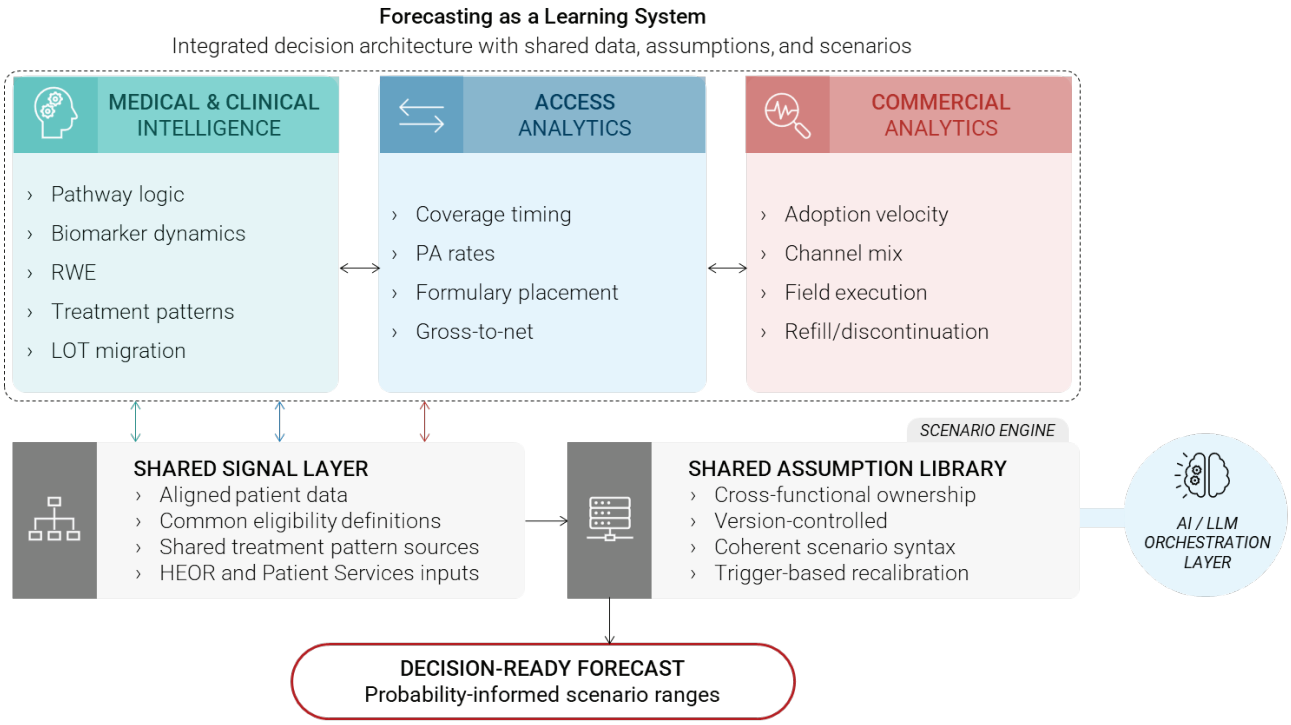


Figure 4. *The Learning System architecture: integrated decision infrastructure with shared signal layer, assumption library, scenario engine, and AI orchestration.*

4. THREE DOMAINS, ONE SYSTEM: ROLES AND CONTRIBUTIONS

Integration does not eliminate the distinctive value of each analytical domain. It defines what each contributes to a shared system and what it receives in return.

4.1 Medical and Clinical Intelligence: The Pathway Logic

Medical and clinical intelligence — drawing from medical affairs, clinical development, HEOR, and the published literature — contributes the evidence logic defining who is eligible, how they arrive at eligibility, and how the clinical pathway evolves. This includes biomarker dynamics, line-of-therapy migration, real-world treatment sequences, and standard-of-care evolution.

The terminology is intentionally broader than any single function. The evidence base draws from HEOR-generated value evidence shaping payer positioning, clinical development data informing label assumptions, and published literature defining standards of care. Making this evidence base shared and explicit is the first act of this sequence.

In return, medical intelligence receives commercial signal on adoption velocity and access intelligence on coverage friction — helping prioritize which evidence gaps are most consequential for the forecast. A medical team that knows which biomarker gaps are creating the largest commercial shortfall can direct real-world evidence generation more precisely.

4.2 Access Analytics: The Realization Layer

Access analytics contributes intelligence determining what proportion of the eligible population actually reaches and sustains therapy — and at what net revenue. This encompasses coverage timing, formulary dynamics, prior authorization rates, specialty pharmacy fill and abandonment rates, gross-to-net structure, and patient support program performance.

The most important conceptual contribution is the access-adjusted patient funnel: a representation applying access friction at each stage — eligible to tested, tested-and-positive to prescribed, prescribed to filled, filled to persistently adherent — rather than as a single aggregate adjustment at the revenue line. This stage-by-stage application keeps the forecast structurally honest about when and how many patients will actually be reached.

In return, access analytics receives medical pathway logic and commercial execution signal, allowing it to model friction in the context of actual provider behavior.

4.3 Commercial Analytics: The Adoption Signal

Commercial analytics contributes execution-level intelligence: prescriber adoption velocity, segment responsiveness to promotional investment, channel mix dynamics, field execution signal, and early warning from refill and discontinuation data.

In a learning system, these signals are calibration inputs, not merely backward-looking metrics. When early adoption tracks below assumptions in community oncology, that signal should trigger examination of whether biomarker testing is also below assumptions. When refill rates suppress net revenue, that should connect to specialty pharmacy abandonment and prior authorization friction.

In return, commercial analytics receives access-adjusted funnel logic making its adoption denominator more accurate — it is not modeling the theoretically addressable population but the actually reachable one — and medical pathway intelligence grounding its timing assumptions in the clinical realities of when patients present, are tested, and are confirmed eligible.

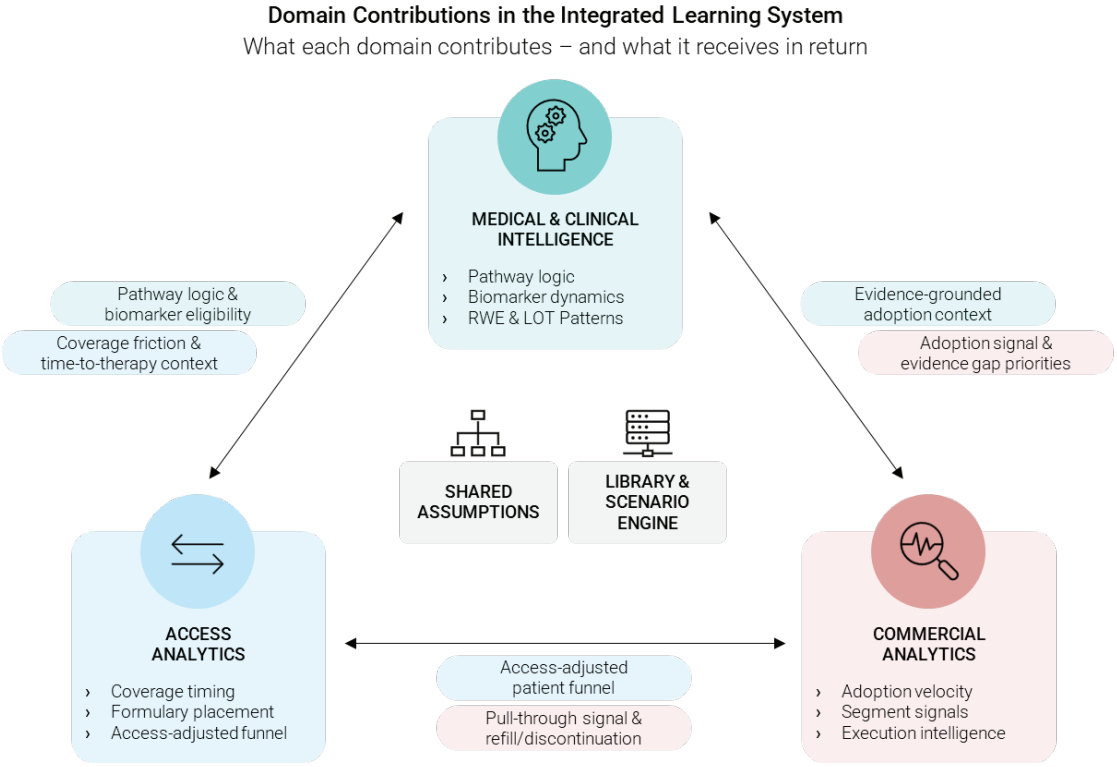


Figure 5. *Bidirectional domain contributions in the Learning System: what each domain contributes and what it receives in return, organized around the shared assumption library.*

5. FROM POINT ESTIMATES TO DECISION-READY INTELLIGENCE

The most consequential output of the learning system is not a more accurate single-point forecast. It is the shift from a single number to a set of named, probability-informed, decision-relevant scenario ranges - ranges that are directly actionable for the decisions pharmaceutical leaders actually face.

5.1 Why Point Estimates Are Structurally Inadequate

A point estimate is a deterministic model output: one value for each assumption, one output per time point. It hides uncertainty. For inherently probabilistic decisions — launch sequencing, indication prioritization, resource allocation, contracting strategy — a single forecast number is inadequate and potentially misleading.

The commercial environment for any complex asset is characterized by genuine uncertainty across interacting dimensions: testing penetration, payer coverage, competitive entries, standard-of-care evolution. None can be confidently predicted to a single value. Presenting them as such erodes forecast credibility.

5.2 The Mechanics of the Shift: Three Approaches

Three approaches, in increasing order of sophistication, enable the shift. All are made more tractable by the learning system's cross-functional assumption governance.

Scenario-based banding is the most accessible starting point. Three to five discrete, coherent scenarios — each representing cross-functionally aligned assumption values — produce individual estimates whose range becomes the decision band. The critical requirement is plausibility: a pessimistic access scenario must reflect medical and commercial assumptions that would plausibly co-occur with delayed coverage.

Monte Carlo simulation adds statistical rigor by replacing fixed assumptions with probability distributions informed by the learning system's shared evidence. The model runs thousands of times, producing a full probability distribution of revenue outcomes with formal confidence intervals.

Bayesian updating is the most conceptually aligned approach. The forecast begins with a prior distribution based on analogs and expert judgment, then continuously updates as real-world signals arrive — formulary confirmations, early prescription trends, specialty pharmacy fill rates. The result is a posterior distribution that tightens over time as uncertainty resolves — explicitly reflecting accumulated evidence. This is what it means, in the most rigorous sense, for a forecast to learn.

5.3 Why Named Scenarios Are More Useful Than Confidence Intervals

Statistical confidence intervals are modeling artifacts; named scenarios are decision tools. The distinction is fundamental.

A confidence interval tells a decision-maker that revenue will likely fall within a band. This is technically correct but practically limited — decision-makers cannot act on a statistical property.

A named scenario — 'Delayed Access, Strong Clinical Differentiation' — describes a real-world future with specific co-determined values

across all three domains. It tells a coherent story the leadership team can recognize, and maps directly to commercial responses: adjust contracting to accelerate formulary placement, redeploy patient support toward prior authorization management, recalibrate field expectations for the first two quarters. The scenario is more useful and more trustworthy than a confidence interval, because its coherence can be validated by each domain expert and its assumptions traced to specific evidence sources.

The learning system produces named ranges naturally, because cross-functional governance enforces the coherence that makes them credible.

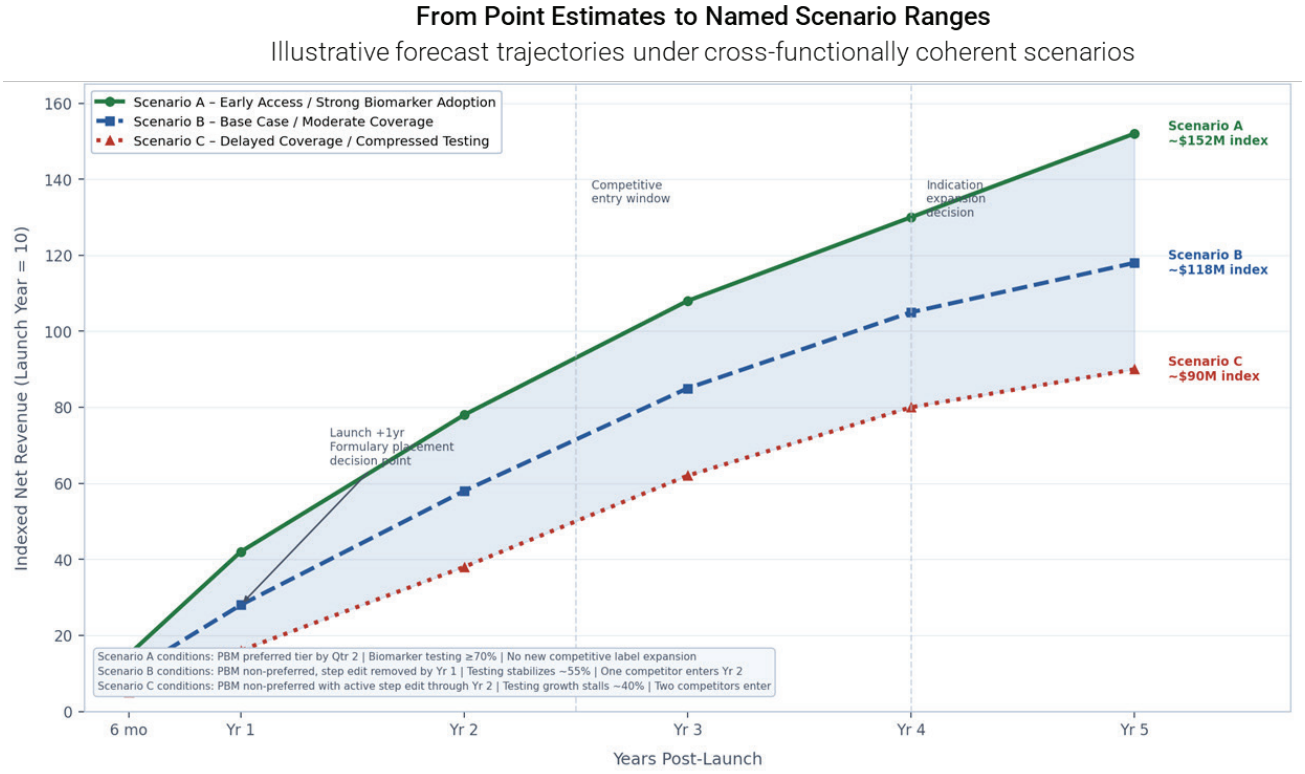


Figure 6. *From point estimates to named scenario ranges: three cross-functionally coherent forecast trajectories tied to specific, decision-relevant assumption conditions.*

> *The shift from single-point estimates to named, probability-informed scenario ranges is not primarily a modeling decision. It is an organizational one. It requires that the assumptions underlying each scenario be cross-functionally owned, evidence-grounded, and coherent across all three analytical domains. This is what the Learning System's assumption governance infrastructure makes possible.*

6. THE ROAD AHEAD: FOUR PROPOSALS FOR THE NEXT DECADE

Based on direct experience with the fragmentation described in this article, four proposals define the forecasting agenda for the next decade. These are actionable directions organizations can begin pursuing now.

Proposal 1: Forecasting as a Continuous Function

The most immediate shift is from calendar-driven to trigger-driven forecasting. Currently, forecasts update on fixed planning cycles. Evidence emerging between cycles either waits or is handled informally.

In a learning system, key signals — biomarker testing rates, formulary confirmations, early specialty pharmacy fill rates, refill and discontinuation trends, competitive label updates — are monitored continuously, with assumption updates triggered when evidence crosses defined thresholds rather than when the planning cycle permits. The model is always current. The planning cycle remains as the formal review cadence, but it is reviewing a forecast that has been learning in real time rather than one that has been static since the last update. This does not mean the model is constantly changing; it means the model is always current. The planning cycle remains as the formal review cadence, but it is reviewing a forecast that has been learning in real time rather than one that has been static since the last cycle.

The practical infrastructure is within reach: signal monitoring dashboards on claims and specialty pharmacy data, assumption alert protocols, and governance for fast-tracking evidence-threshold updates.

Proposal 2: Cross-Functional Forecast Ownership

The most important organizational shift is in forecast ownership. Currently, the forecast is owned by commercial or finance, with medical and access contributing inputs at defined handoff points.

In a learning system, the forecast is owned jointly. Medical, access, and commercial leaders co-own assumptions within their domains, share accountability for scenario coherence, and collectively resolve assumption conflicts.

The *Forecast Integration Lead* is a new organizational role: owning the shared assumption library, facilitating cross-functional alignment on key assumptions, maintaining scenario coherence, and managing the trigger-based calibration process. It is less a technical position than a governance and facilitation function — the organizational embodiment of the learning system's integration mandate. Organizations that create and invest in this role will find forecast quality and credibility improve materially, independent of any model changes.

Proposal 3: AI-Native Assumption Governance

The third proposal applies Agentic AI and LLM-based orchestration to assumption management. Manually maintaining a shared assumption library — mining research reports, tracking evidence, flagging inconsistencies, managing version history — is labor-intensive at scale. AI flips this limitation.

LLM-based agents that continuously mine sources for assumption-relevant evidence, flag divergences from current assumptions, surface cross-functional inconsistencies, and maintain living documentation are not speculative —

the underlying capabilities are demonstrably available in current deployments.The application of this capability to pharmaceutical forecasting is an immediate practical opportunity, not a future state.

The critical requirement is that AI-managed updates are traceable, auditable, and subject to human approval. AI provides scale and consistency, human oversight provides judgment. The combination is more robust than either alone.

Proposal 4: Early Convergence at the Asset Strategy Level

The fourth proposal integrates access and medical analytics into the forecast architecture before launch — specifically at the asset strategy level during late-stage development.

By the time the commercial team builds the launch forecast, the most consequential assumptions — label, biomarker restrictions, payer evidence standards, diagnostic infrastructure, patient support architecture — have already been determined during development.

As payer evidence requirements increasingly shape clinical development decisions — trial design, endpoint selection, patient population definitions — the separation between pipeline and launch forecasting has become inconsequential. Organizations integrating access and medical analytics at the asset strategy stage will enter the commercial period with more realistic forecasts.

This is not a call for earlier commercial involvement in development for its own sake. It is a specific, evidence-driven necessity. Because payer evidence requirements and patient identification dynamics are now principal determinants of commercial outcome, they belong in the forecast architecture from the point at which they become knowable — which is during Phase III, not after approval.

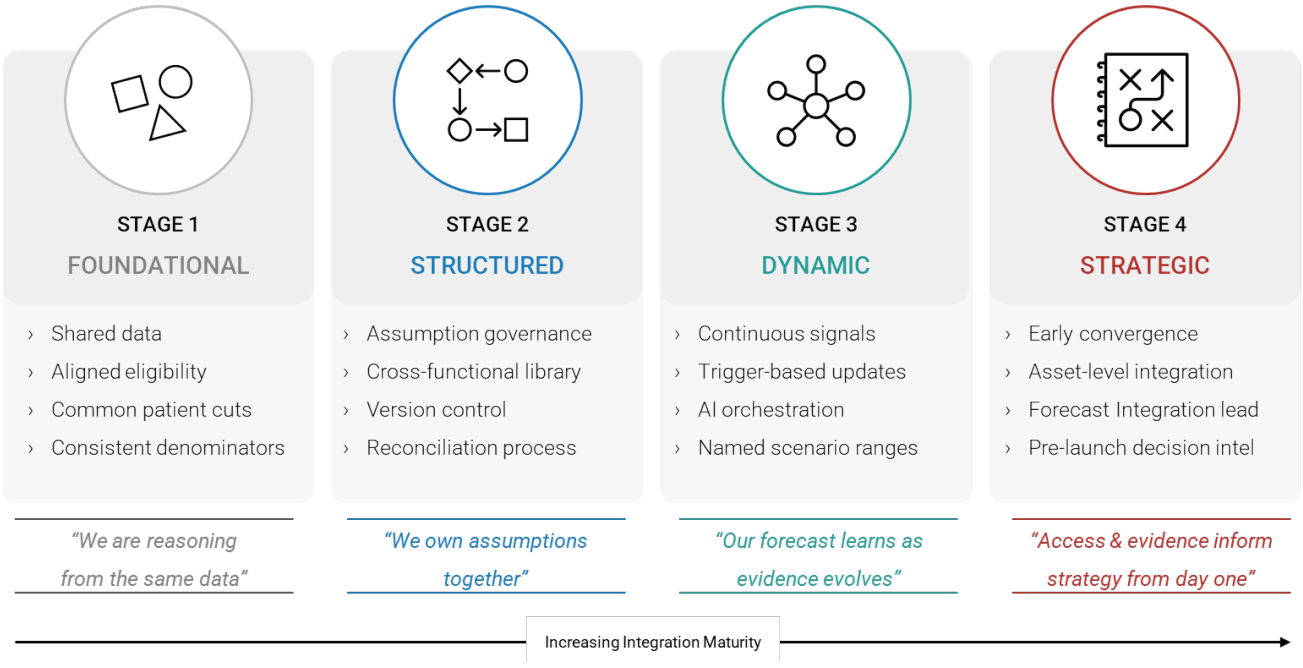


Figure 7. *The Forecasting Maturity Roadmap: a four-stage pathway from shared data foundations to strategic asset-level integration.*

7. CONCLUSION: THE ARCHITECTURE OF LEARNING

The forecasting gap in the pharmaceutical industry is not a testimony to analytical limits. The models, data, and talent are present. The gap reflects the structural disconnection between analytical domains that each produce high-quality outputs but have not been organized to produce an integrated picture.

The learning system addresses this through organizational design and process architecture. Its three components — shared signal integration, structured assumption governance, and continuous scenario refinement — are achievable with current data, technology, and talent, given the right governance design. AI can accelerate the integration; it cannot replace the design.

The four proposals — continuous forecasting, cross-functional ownership, AI-native governance, and early asset-level convergence — represent a sequential progression from foundational integration to strategic differentiation.

The deeper payoff is a more useful forecast: one that speaks to decisions rather than plans, that reflects the full commercial reality of medical, access, and commercial dynamics rather than a functional slice of it, and that learns continuously from the evidence that emerges after every launch milestone and every payer negotiation.

Organizations that build this architecture will compound their decision intelligence over the lifecycle of every asset they commercialize. In a commercial environment of increasing complexity and compressing timelines, that compounding advantage is among the most durable competitive advantages accessible.

The next decade of forecasting belongs to organizations that treat it as a living system — and to the leaders with the vision and conviction to build it that way.

ABOUT THE AUTHORS

All authors are with Viscadia, a strategic forecasting team of experts serving biopharmaceutical clients across therapeutic areas and asset lifecycle stages.

- **Anindya Roy**, Principal – Forecasting, Market Research, Cross-functional Commercial Strategy, AI in Forecasting, BD Strategy
- **Vipul Vaid**, Principal – Oncology and Rare Disease Forecasting, Market Access Analytics, Strategic Scenario Planning
- **Vikrant Kumar**, Manager – Forecasting Model Development, Data Strategy, Cross-Functional Analytics Integration
- **Himanshu Gurnani**, Manager – Commercial Analytics, Assumption Governance, AI-enabled Forecasting Infrastructure

REFERENCES

1. DrugPatentWatch. How clinical trial data supports accurate pharma forecasting. [Internet]. 2025 [cited 2026 Feb]. Available from: https://www.drugpatentwatch.com/blog/how-clinical-trial-data-supports-accurate-pharma-forecasting/

2. Waqar SN, et al. Biomarker testing for patients with advanced/metastatic nonsquamous NSCLC in the United States of America, 2015 to 2021. JTO Clin Res Rep. 2022;3(6). doi:10.1016/j.jtocrr.2022.100330

3. Butrynski JE, Paschold JC, Ward PJ, et al. Contemporary biomarker testing rates in both early and advanced NSCLC: Results from the MYLUNG pragmatic study. J Clin Oncol. 2023;41(16 Suppl):9109. doi:10.1200/JCO.2023.41.16_suppl.9109

4. Lipsyc-Sharf MD, et al. Real-world biomarker test ordering practices in non-small cell lung cancer: Interphysician variation and association with clinical outcomes. JCO Precis Oncol. 2024. doi:10.1200/PO.24.00039

5. Doherty M, et al. Impact of prior authorization on patient access to cancer care. Am Soc Clin Oncol Educ Book. 2024. doi:10.1200/EDBK_100036

6. Fein AJ. Gross-to-net bubble hits $356B in 2024 - but growth slows to 10-year low. Drug Channels Institute. [Internet]. 2025 [cited 2026 Feb]. Available from: https://www.drugchannels.net

7. Yang C, Cintina I, Pariser A. Cost of delayed diagnosis in rare disease. EveryLife Foundation for Rare Diseases. [Internet]. 2023 [cited 2026 Feb]. Available from: https://everylifefoundation.org/delayed-diagnosis-study/

8. Dubief J, Faye F. The diagnostic odyssey of people living with a rare disease: Key findings from a Rare Barometer survey. EURORDIS-Rare Diseases Europe. 2024.

9. Waterhouse DM, et al. Understanding contemporary molecular biomarker testing rates and trends for metastatic NSCLC among community oncologists. Clin Lung Cancer. 2021;22(6):e901-e910.

10. American Cancer Society National Lung Cancer Roundtable. Strategic plan: Advancing comprehensive biomarker testing in non-small cell lung cancer. CA Cancer J Clin. 2021;71(3):207-228. doi:10.3322/caac.21658

11. IntegriChain. Navigating the complexities of biopharma/pharma net revenue forecasting. [Internet]. 2025 [cited 2026 Feb]. Available from: https://www.integrichain.com

12. PwC. Future of pharma - Breakthroughs at scale. [Internet]. 2025 [cited 2026 Feb]. Available from: https://www.pwc.com/us/en/industries/pharma-life-sciences/pharmaceutical-industry-trends.html

Accelerating Market Research Insights through an Ontology-Driven, Agentic Data Digitization Framework

Yuan Niu,[1] Junying Zheng,[i] Mathew Ratnam,[i] PKS Prakash,[ii] Shivam Shivam,[ii] Swaraj Prakash,[ii] Rishav Mishra[ii]

Abstract: Pharmaceutical primary market research generates vast volumes of heterogeneous, multimodal data spanning reports, presentations, transcripts, and multimedia recordings. Analysts face significant challenges extracting reliable, decision-ready insights manually, a process that is slow, inconsistent, and poorly suited to the complexity of modern pharmaceutical intelligence workflows. Standard Retrieval-Augmented Generation systems fall short, lacking pharmaceutical terminology awareness, relational entity understanding, and granular source traceability demanded by regulatory and commercial stakeholders. We present GRAM (Graph-based Retrieval for Augmented Market research), integrating ontology-driven knowledge graphs with semantic vector search in a hybrid retrieval pipeline over AWS Neptune and AWS OpenSearch, unified through a Learning-to-Rank model optimized for pharmaceutical relevance. A multimodal ingestion pipeline processes text, tables, charts, audio, and video while preserving domain relationships. Expert-led validation across oncology and women's health demonstrates improved retrieval precision, near-elimination of hallucinations, and exact slide-level source attribution compared to standard RAG baselines, confirming GRAM as an enterprise-ready foundation for trustworthy pharmaceutical market research intelligence.

Index Terms: Retrieval-Augmented Generation, Knowledge Graph, Pharmaceutical Market Research, Learning-to-Rank, Multimodal Processing, Ontology-Driven Search

I. INTRODUCTION

Pharmaceutical companies rely substantially on market research to guide their strategic commercial decisions. As brands progress through different lifecycle stages, the volume of research continues to grow, posing a challenge for stakeholders to quickly locate actionable insights. This research—spanning primary sources such as physician interviews and surveys, and secondary sources such as market reports and published literature—exists across vast repositories of presentations, transcripts, and multimedia content, making efficient retrieval difficult. Traditional insight retrieval approaches are manual, time-intensive, and error-prone, limiting timely, data-driven decision-making. Recent advancements in Artificial Intelligence (AI), particularly in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG)[1] systems, offer promising solutions for automating information retrieval and synthesis.

Research artifacts span PDFs, presentations, spreadsheets, scanned documents, and audio/video recordings, each requiring distinct processing before retrieval or synthesis can occur. Extracting structured information from

[i] Bayer Pharmaceuticals, USA
[ii] ZS Associates, India

charts, tables, and transcribed speech introduces complexity that generic pipelines are ill-equipped to handle, with downstream business impacts including delayed strategic decisions, inconsistent analyst outputs, and compliance risks from overlooked regulatory information.

LLMs such as GPT-4o[2] and LLaMA[3] show strong NLP performance[4] but are limited by pre-trained knowledge, resulting in hallucinations, domain gaps, and high finetuning costs.[5] RAG addresses this via dynamic external retrieval,[6] but conventional RAG overlooks relational entity dependencies—a critical gap for understanding drug relationships, therapeutic areas, and competitive landscapes.

To overcome these limitations, we propose GRAM (Graphbased Retrieval for Augmented Market research), a RAG framework tailored for pharmaceutical market research, encompassing both primary research (physician interviews, patient surveys, conference proceedings) and secondary research (market reports, competitive intelligence, published literature). GRAM employs a hybrid retrieval mechanism combining AWS OpenSearch and AWS Neptune, enhanced by domain-specific ontologies and semantic chunking.[7] Audio and video are transcribed via OpenAI Whisper Large-v3 with GPT-4o terminology refinement,[2] and a customized LTR model optimizes result prioritization. The solution is therapeutically agnostic, enabling extension across pharmaceutical portfolios with minimal customization.

Our main contributions are as follows:

- **Document Processing Pipeline:** A multimodal pipeline handling text, tables, charts, audio, and video, with Whisper-based transcription, semantic chunking, and hybrid AWS OpenSearch and Neptune integration, capturing domain process knowledge through ontology-aligned extraction and structured provenance preservation across all ingested artifacts.
- **Domain-Specific RAG Application:** An agentic retrieval system orchestrating query decomposition, graph traversal, embedding search, and ontology-driven re-ranking via LTR models, encoding pharmaceutical process knowledge — including drug-indication relationships, competitive positioning, and regulatory workflows — to deliver analyst-quality market research responses.
- **PPT/Image Generation:** A Slide and Image Generator Accelerator that automatically transforms retrieved insights and captured process knowledge into structured presentations and visual summaries, enabling consistent, audit-ready stakeholder communication with full source traceability.

The remainder of this paper is structured as follows: Section 2 reviews background on RAG methodologies and market research challenges; Section 3 details the document processing pipeline; Section 4 presents the search and retrieval mechanism; Section 5 describes the technical architecture; and Section 6 reports experimental validation and results.

II. BACKGROUND

Pharmaceutical commercialization depends on timely, accurate intelligence drawn from heterogeneous research sources. This section reviews the core challenges of pharmaceutical market research, the role of RAG-based systems in addressing them, and the limitations that motivate GRAM's graph-enhanced design.

A. Market Research Challenges in Pharmaceuticals

Pharmaceutical market research involves analyzing diverse primary sources (physician interviews, patient surveys, conference proceedings) and secondary sources (market reports, competitive analyses, published literature) to inform strategic decisions. Key challenges include:

Heterogeneous Data Sources: Research data spans across quantitative reports, qualitative interviews, competitive intelligence, and regulatory documents, each requiring distinct processing while preserving entity relationships across drugs, competitors, and indications.

Domain-Specific Terminology: Pharmaceutical terminology encompasses drug names (brand, generic, chemical), therapeutic areas, and market dynamics (NBRx, TRx, market share). Standard RAG systems struggle with these terms, misinterpreting abbreviations (VMS, HRT, HCP) and losing drug-indication relationships.

Query and Precision Requirements: Queries frequently require multi-document synthesis, temporal reasoning, and quantitative analysis. Unlike general Q&A systems, pharmaceutical market research additionally demands exact slide-and page-level source attribution, conflict handling, and full provenance tracking for regulatory compliance.

B. RAG and Its Role in Market Research

LLMs such as GPT-4o and LLaMA show strong NLP capabilities[4] but are prone to hallucinations and outdated responses due to static pre-trained knowledge. RAG addresses this by enabling LLMs to dynamically fetch relevant documents,[6] improving factual accuracy. In pharmaceutical contexts, RAG-augmented LLMs have shown promise in clinical decision support, biomedical literature analysis, and drug discovery,[8] with external knowledge integration ensuring accuracy and factual consistency.[9,10] However, standard RAG lacks slide-level provenance, fails to capture pharmaceutical entity relationships through generic embeddings, and cannot model hierarchical structures essential for multi-entity queries, motivating GRAM's graph-enhanced design.

C. Graph-Based Traversal and Ontology-Driven Search

Graph-based retrieval encodes entity relationships in knowledge graphs traversed at query time, enabling more precise retrieval than text similarity alone,[11] structuring data around drugs, competitors, and regulatory agencies. Ontology-driven search extends beyond classification by providing an explicit domain map.

A well-defined pharmaceutical ontology encodes entity types, relationships, and hierarchies, enabling the agent to decompose complex questions into ontology-guided sub-queries. During retrieval, it acts as a semantic filter preventing category errors, while synonym resolution across brand, generic, and chemical identifiers prevents terminological fragmentation. Ranked results inherit ontological relevance signals, weighting matches satisfying both semantic similarity and structural proximity more highly.

Our system combines AWS OpenSearch for vector search with AWS Neptune for graph-based retrieval, unified by an LTR model fine-tuned on pharmaceutical market research data to prioritize results by expert-defined relevance criteria. Beyond retrieval, integrated Slide and Image Generator Accelerators automate transformation of retrieved insights into structured presentations, accelerating stakeholder communication and reducing manual report compilation effort.

III. DOCUMENT PROCESSING PIPELINE

Our automated document processing pipeline, illustrated in Fig. 1, transforms unstructured pharmaceutical market research data spanning PDFs, PowerPoint presentations, Excel sheets, images, charts, and multimedia files into structured, searchable information through three stages: (1) Data Ingestion, (2) Natural Language Processing, and (3) Database Pipeline.

A. Data Ingestion

The pipeline accepts documents (.docx, .pdf, .txt), emails (.msg, .eml), presentations (.pptx), spreadsheets (.xlsx, .csv), and multimedia files (.jpg, .png, .mp4, .wav). A SharePoint crawler monitors the repository in real-time, detecting new, modified, and deleted files. Each document is assigned a unique ID for traceability, and a versioning table tracks ingestion status and flags documents requiring reprocessing.

1) Format Conversion and Text Extraction: All documents are standardized to PDF before extraction. Text-based formats (DOCX, PPTX, XLSX) are converted via LibreOffice; scanned documents are processed using OCR. Multimedia files undergo a three-step transcription process: (1) speaker diarization to separate speakers, (2) transcription via OpenAI Whisper Large,[12] and (3) pharmaceutical terminology refinement using GPT-4o to correct industry-specific terms such as drug names and abbreviations.

Text content is extracted from paragraphs, bullet points, and slides; tables and charts are processed by GPT-4o to generate summarized insights rather than raw data; and images are interpreted using multimodal AI models. All extracted content is stored in structured JSON format, preserving metadata including page numbers, section hierarchy, and source provenance.

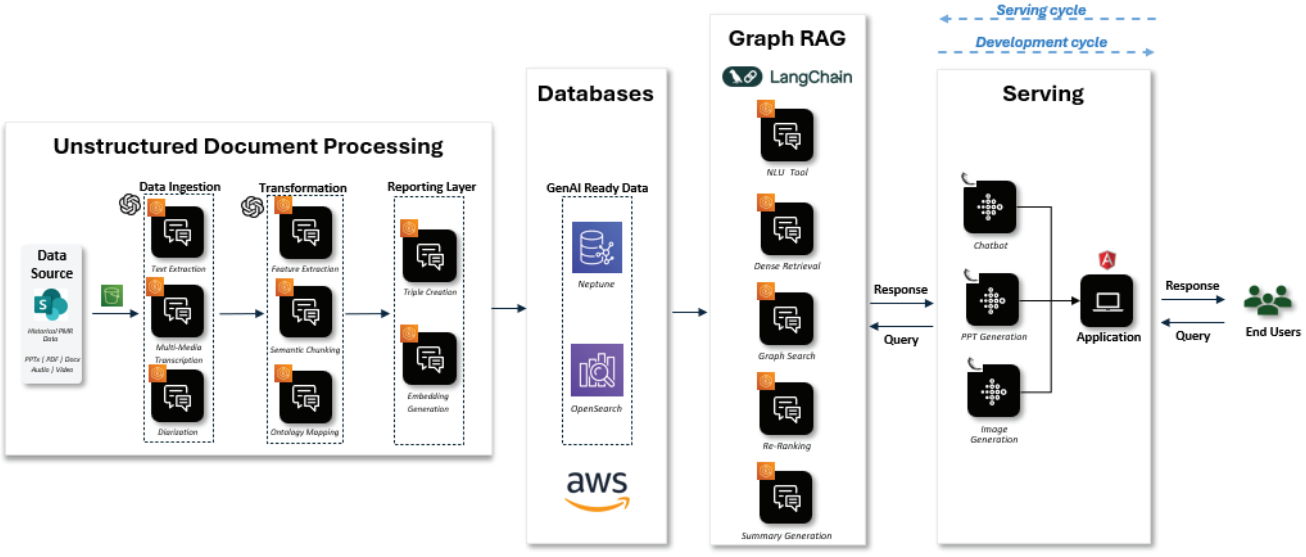


Figure 1. *GRAM (Graph-based Retrieval for Augmented Market research) Document Processing Architecture*

B. Natural Language Processing (NLP)

Documents are first segmented into logical sections (part extraction) to reduce the feature extraction search space and preserve document structure for context-aware retrieval. The NLP pipeline identifies entities (drug names, diseases, competitors), relationships (drug-indication mappings, competitive positioning), and hierarchical sections mapped to pharmaceutical taxonomies, using pattern-based rules, NER, and GPT-4o-assisted extraction for complex documents. Extracted insights are stored in ontology-aligned JSON format.

C. Database Pipeline

Extracted insights are stored in a hybrid retrieval system combining graph-based and vector-based search.

1) Graph Construction in AWS Neptune: Entities and relationships are converted into RDF triples[13] (subject, predicate, object) and serialized as Turtle (TTL) files following a predefined pharmaceutical ontology. Each TTL file is validated against the ontology schema for consistency before ingestion into AWS Neptune, enabling relationship-based queries such as competitive drug positioning and regulatory approval timelines.

2) Semantic Chunking and OpenSearch Ingestion: Extracted text is segmented into semantically coherent chunks, each converted into a vector embedding using the BAAI General Embedding (BGE)-base model.[14] Embeddings are indexed in AWS OpenSearch, enabling semantic and hybrid search. This combination of graph-based relationship retrieval and vector-based semantic search forms the foundation of GRAM's hybrid retrieval mechanism described in Section 4.

IV. SEARCH AND RETRIEVAL MECHANISM

The retriever is the most critical RAG component, responsible for fetching relevant information to augment the generation model's context,[7] as illustrated in Fig. 2. Classical methods including BM25,[15] TF-IDF,[16] KNN,[17] and bi-encoder designs[18] have been widely studied, with recent work exploring GNNs.[19] GRAM extends this with a hybrid mechanism: query entities act as filtering agents for AWS Neptune graph traversal before vector-based retrieval, maximizing recall while maintaining low latency.

A. Graph-Based Entity Retrieval

Using Named Entity Recognition (NER) and custom pharmaceutical ontologies, the system identifies key entities drugs, diseases, competitors, and market terms from the user query. These entities filter the AWS Neptune knowledge graph via graph traversal, locating the most relevant nodes and associated document sections before vector search is performed. This structured pre-filtering aligns results with the domain-specific taxonomy and reduces the search space for subsequent embedding-based retrieval.

B. Embedding-Based Similarity Search

In parallel, each text chunk is encoded using the BGE-base embedding model and indexed in AWS OpenSearch. At query time, the query is embedded into the same vector space and cosine similarity is used to rank the most semantically relevant chunks. Running graph and embedding search in parallel ensures that retrieved information is both structurally relevant and semantically precise while minimizing latency.

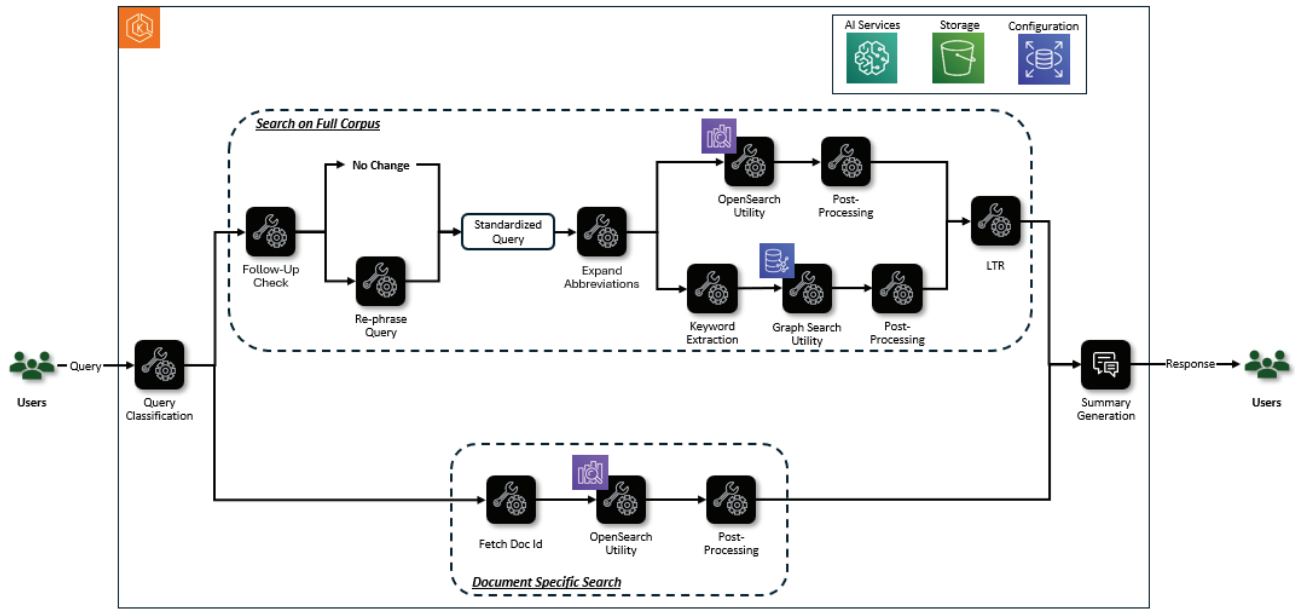


Figure 2. *GRAM (Graph-based Retrieval for Augmented Market research) Search Architecture*

C. Learning to Rank (LTR) for Optimized Search

A custom LTR model unifies graph-based and embedding-based scores into a single ranked list. Graph nodes surfaced through traversal are ranked by entity match frequency along traversal paths, structural relevance within the document hierarchy, and number of connections to high-priority nodes. Embedding chunks are ranked by cosine similarity. The final ranked list prioritizes documents scoring highly in both modalities, further refined by business rules such as document recency, frequency of mention, and section hierarchy. Graph-based scores are weighted more heavily for domain-specific queries (e.g., drug approvals, competitive intelligence), while embedding similarity is prioritized for context-sensitive queries.

D. Cost and Latency Optimization

Three strategies reduce computational overhead: (1) **Graph Traversal Pre-filtering** narrows the document set before vector search, eliminating unnecessary embedding comparisons; (2) **Embedding Clustering** removes redundant or nearduplicate chunks to reduce context size; and (3) **Dynamic Context Windowing** passes only the highest-scoring sections to the LLM rather than full documents. SPARQL queries in AWS Neptune are further optimized for faster graph traversal, and parallel execution of graph and vector search ensures both retrieval paths complete without sequential blocking.

V. TECHNICAL ARCHITECTURE

GRAM is implemented as a cloud-native, agentic framework on AWS, autonomously orchestrating query understanding, retrieval, ranking, and response generation across microservices, with a clear separation between the Search Application Flow and Data Ingestion Pipeline.

A. System Overview

As shown in Fig. 3, the framework runs as microservices on AWS EKS, with frontend and backend pods orchestrating retrieval and generation workflows. Identity is managed via AWS Cognito and Azure AD, with perimeter security via AWS WAF and Shield. Storage spans Amazon RDS (MySQL) for metadata, OpenSearch for semantic indexing, and Neptune for graph knowledge. Ingestion is orchestrated via MWAA with AWS Batch, and CloudWatch provides system-wide monitoring.

B. Search Application Flow

The search application flow processes user queries in realtime through three sequential stages.

1. **User Request Handling and Security:** Users authenticate via Azure AD using Single Sign-On (SSO). Authenticated requests are routed through ALB to the EKS-hosted frontend pod, which forwards queries to the backend processing module. AWS WAF and Shield provide perimeter security against unauthorized access and application-layer attacks.
2. **Query Processing and Retrieval:** The EKS backend service performs parallel retrieval: OpenSearch handles semantic and keyword-based search over indexed document chunks, while Neptune is queried for complex questions requiring structured relationship traversal across the knowledge graph. Retrieved documents and graph nodes are ranked and processed through the RAG-based agentic framework, which orchestrates context augmentation before passing to the generation model.
3. **Response Generation and Delivery:** The framework dynamically formats responses based on query complexity and user preferences. Generated answers are returned to the frontend pod for presentation in the UI. Where requested, the system produces structured outputs including detailed reports, PowerPoint presentations, and images derived from the retrieved insights.

C. Data Ingestion Pipeline

The data ingestion pipeline ensures continuous updating of the knowledge base as new market research documents become available. End-to-end ingestion workflows are orchestrated via MWAA (Managed Apache Airflow), with AWS Batch handling compute-intensive processing jobs for large document volumes.

1. **Document Processing and Storage:** Raw documents are ingested from multiple sources including APIs, databases, and external file repositories, and stored in Amazon S3 as the primary data lake. Extracted metadata and structured insights are persisted in RDS MySQL for efficient query performance, while processed and chunked text is indexed in OpenSearch for real-time semantic retrieval.
2. **Knowledge Graph Construction:** Entity relationships extracted during document processing are structured as RDF triples and ingested into Amazon Neptune, forming the pharmaceutical knowledge graph that underpins GRAM's graph-based retrieval. This enables advanced relationship-based querying across drugs, indications, competitors, and regulatory entities as described in Section 3.

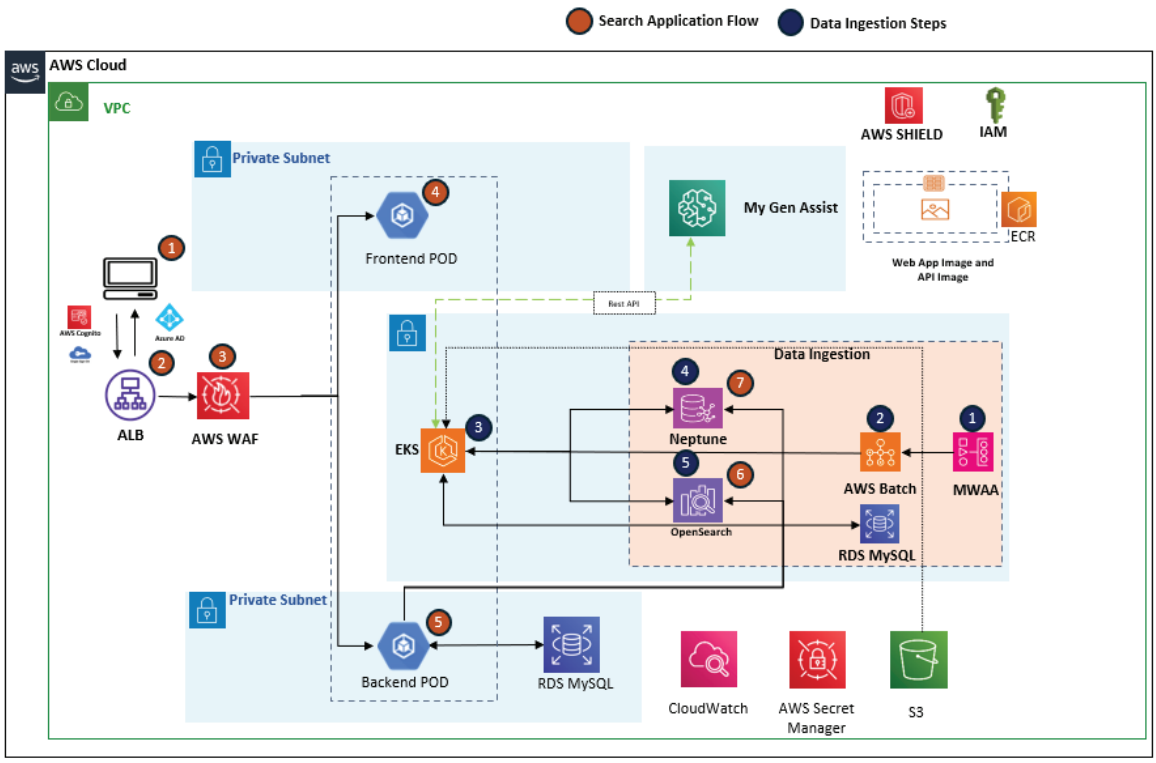


Figure 3. *GRAM (Graph-based Retrieval for Augmented Market research) Technical Architecture*

D. Security, Monitoring and Scalability

Security is enforced via IAM least-privilege access, Active Directory group-based authorization, VPC network segmentation, AWS WAF and Shield, and Secrets Manager for credential management. CloudWatch provides centralized monitoring across all microservices. The framework scales via EKS Auto-scaling, Multi-AZ deployment for high availability, and ALB load balancing.

VI. EXPERIMENTAL VALIDATION AND RESULTS

We conducted comprehensive validation through independent evaluations by Subject Matter Experts (SMEs) across three pharmaceutical products spanning oncology and women's health therapeutic areas. The evaluation encompassed query performance testing, functional capability validation, and head-to-head comparison with a standard vector-based RAG baseline system using industry-standard information retrieval and natural language generation metrics.

A. Evaluation Metrics and Definitions

To ensure rigorous and reproducible assessment, we adopt widely established metrics from information retrieval[20] and natural language generation evaluation literature. All metrics are formally defined below.

Precision. Precision measures the fraction of retrieved documents that are relevant to the query:[21]

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (1)$$

Recall. Recall measures the fraction of all relevant documents that are successfully retrieved:[21]

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{2}$$

F1 Score. The harmonic mean of precision and recall, providing a balanced measure of retrieval quality:[21]

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Normalized Discounted Cumulative Gain (NDCG@k). NDCG evaluates ranking quality by assigning higher weight to relevant documents appearing at top positions:[22]

$$\text{NDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}, \quad \text{DCG@k} = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{4}$$

where rel_i is the relevance grade of the document at position i and IDCG@k is the ideal DCG obtained by ranking all documents by decreasing relevance.

Mean Reciprocal Rank (MRR). MRR measures how quickly a relevant document appears in the ranked result list:[23]

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{5}$$

where rank_i is the position of the first relevant document for query i.

Answer Faithfulness. Faithfulness quantifies the fraction of claims in the generated response supported by retrieved context, following the RAGAS evaluation framework:[24]

$$\text{Faithfulness} = \frac{|\{\text{supported claims}\}|}{|\{\text{total claims in response}\}|} \tag{6}$$

Context Precision. Context precision measures the signal-to-noise ratio in retrieved contexts, evaluating whether retrieved passages are relevant to answering the query:[24]

$$\text{Context Precision} = \frac{|\{\text{relevant retrieved chunks}\}|}{|\{\text{total retrieved chunks}\}|} \tag{7}$$

Answer Relevancy. Following the RAGAS framework, answer relevancy assesses the pertinence of the generated answer to the query using embedding-based semantic similarity:[24]

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^{N} \text{sim}(q, q_i^{\text{gen}}) \tag{8}$$

where q is the original query, q_i^{gen} are N questions reversegenerated from the answer, and sim(·) denotes cosine similarity. A query is considered *passed* if its answer relevancy score meets or exceeds the 50% threshold, and *high-quality* if it meets or exceeds 70%.

Hallucination Rate. The proportion of generated responses containing at least one unsupported or fabricated claim not traceable to any source document:

$$\text{Hallucination Rate} = \frac{|\{\text{responses with unsupported claims}\}|}{|\{\text{total responses}\}|} \tag{9}$$

Source Attribution Accuracy. The fraction of cited references correctly pointing to the claimed source location (slide, page, or document):

$$\text{Attribution Accuracy} = \frac{|\{\text{correct citations}\}|}{|\{\text{total citations}\}|} \tag{10}$$

B. Multi-Product Query Performance

Table I presents query performance across three pharmaceutical products as evaluated by Subject Matter Experts (SMEs). A total of 104 complex pharmaceutical market research queries—48 for Drug A, 35 for Drug B, and 21 for Drug C—were independently evaluated by SMEs with therapeutic area expertise. Answer relevancy scores were computed using the RAGAS framework[24] as defined in Equation (8), while source attribution accuracy was computed using Equation (10).

SME evaluation demonstrated robust cross-therapeutic performance. The overall pass rate of 94.2% and median answer relevancy of approximately 89.5%—exceeding the average of 86.3% by 3.2 percentage points—indicate a positively skewed distribution where most queries perform well above threshold while a small number of outliers depress the mean. The 83.7% high-quality response rate (≥70% answer relevancy) validates strong production readiness for pharmaceutical market research applications. Source attribution accuracy of 0.98 overall—with per-product scores ranging from 0.97 to 0.99—demonstrates that 98% of all citations correctly point to the exact slide, page, or document location, enabling immediate source verification. The remaining 2% of attribution errors were traced to ambiguous cases where identical content appeared across multiple slides within the same document.

Drug A achieved the highest pass rate (95.8%) with median answer relevancy of approximately 89.0%; Drug B followed at 94.3% with median 93.1%; Drug C achieved 90.5% with median 91.4%. Cross-product variance in average answer relevancy of 5.2 percentage points (84.0%–89.2%) reflects differences in corpus maturity and query complexity across therapeutic areas rather than fundamental retrieval inconsistency. Low-scoring queries (5.8%, $n = 6$) were attributed by SMEs primarily to documents not yet indexed and higher query complexity in oncology products involving multi-source synthesis across clinical and commercial data, rather than retrieval quality issues, suggesting performance improves further with knowledge base completeness. Drug B's strong pass rate of 94.3% reflects its mature document corpus, while Drug A's 95.8% pass rate across the largest query set of 48 queries validates robust retrieval at scale.

Metric	Overall	Drug A	Drug B	Drug C
Total queries evaluated	104	48	35	21
Queries passed ($\geq$50% relevancy)	98 (94.2%)	46 (95.8%)	33 (94.3%)	19 (90.5%)
Average answer relevancy	86.3%	84.0%	89.2%	86.9%
Median answer relevancy	$\sim$89.5%	$\sim$89.0%	93.1%	91.4%
High-quality responses ($\geq$70%)	87 (83.7%)	40 (83.3%)	30 (85.7%)	17 (81.0%)
Source attribution accuracy	0.98	0.97	0.99	0.98
Average response time (seconds)	30	28	30	32

Table 1. *Query performance across pharmaceutical products (sme evaluation,* n = *104). Answer relevancy computed via the ragas Framework;24 source attribution accuracy computed per equation (10).*

C. Functional Capability Validation

SMEs conducted systematic functional testing across 84 capability tests spanning 10 functional categories including terminology comprehension, conversational context management, file generation, and system features, assessing production readiness across diverse operational requirements. The system achieved an overall functional pass rate of 88.1% (74/84 tests), demonstrating strong production readiness. Core pharmaceutical market research capabilities achieved perfect scores: typo tolerance (5/5, 100%), brand and generic name handling (5/5, 100%), response correctness (4/4, 100%), and authentication and login (3/3, 100%), confirming that foundational capabilities essential for pharmaceutical analyst workflows are fully operational.

Critical domain-specific features demonstrated robust performance. Pharmaceutical terminology comprehension achieved 88.9% across 36 tests, confirming correct interpretation of abbreviations (VMS, HRT, HCP, NBRx, TRx), therapeutic classifications, and brand-generic equivalence. Guardrails and safety achieved 88.9% (8/9 tests), and filter functionality achieved 88.9% (8/9 tests), confirming reliable filtering by therapeutic area, date range, and document type.

Advanced features showed strong but improvable performance reflecting feature maturity. LLM operations quality, evaluated using the RAGAS framework[24]—measuring context precision, answer relevancy, and faithfulness—achieved 80.0% (4/5 tests) against a predefined 85% relevancy threshold. The single failure case involved context drift during extended multi-turn dialogues, identifying a specific enhancement opportunity in sustained conversational coherence. Conversational context management achieved 66.7% (2/3 tests), indicating room for improvement in maintaining state across complex multi-turn interactions. File generation capabilities (PPT, PDF, image) achieved 60.0% (3/5 tests), with failures concentrated in complex formatting requirements for pharmaceutical presentation templates.

D. Comparative Analysis: GRAM vs. Standard RAG

SMEs conducted head-to-head comparison between GRAM and a production-deployed standard vector-based RAG application across 12 representative pharmaceutical market research queries for Drug B, stratified by complexity (3 low, 6 medium, 3 high). Table II presents quantitative comparison on retrieval quality and accuracy metrics, with generation quality and operational characteristics discussed subsequently.

Retrieval Quality Analysis. GRAM demonstrated 18–25% improvements across all retrieval metrics. Precision@5 rose from 0.71 to 0.89 through ontology-aligned graph pre-filtering; Recall@10 of 0.98 confirms near-complete relevant document capture; F1@5 of 0.92 versus 0.74 reflects balanced gains. NDCG@5 of 0.92 and NDCG@10 of 0.94 confirm superior ranking quality, while MRR of 0.96 versus 0.80 confirms relevant documents surface at the top of ranked lists.

Generation Quality Analysis. Generation quality was evaluated using the RAGAS framework[24] across the same 12 comparative queries. GRAM achieved answer faithfulness of 0.96 versus 0.83, with the knowledge graph acting as a factual validation layer during response formulation. Context precision improved from 0.75 to 0.88 through graph traversal supplying fewer irrelevant passages, and answer relevancy reached 0.91 versus 0.79 via ontology-driven query intent understanding.

Metric	Overall	Drug A	Drug B	Drug C
Total queries evaluated	104	48	35	21
Queries passed ($\geq$50% relevancy)	98 (94.2%)	46 (95.8%)	33 (94.3%)	19 (90.5%)
Average answer relevancy	86.3%	84.0%	89.2%	86.9%
Median answer relevancy	$\sim$89.5%	$\sim$89.0%	93.1%	91.4%
High-quality responses ($\geq$70%)	87 (83.7%)	40 (83.3%)	30 (85.7%)	17 (81.0%)
Source attribution accuracy	0.98	0.97	0.99	0.98
Average response time (seconds)	30	28	30	32

Table 2. *Quantitative performance comparison: GRAM vs. Standard RAG (*n *= 12 queries, drug B)*

Accuracy and Reliability. GRAM achieved 0.99 source attribution accuracy with exact slide-level citations versus the baseline's 0.72 document-level attribution, a 37.5% improvement, reducing fact-checking from 5–7 minutes to 30 seconds (90% efficiency gain). Hallucination rate fell 92.9% from 0.14 to 0.01; baseline incidents included fabricated entity names, unverifiable citations, and cross-therapeutic retrieval errors, eliminated through graph-constrained generation validating claims before response formulation.

Operational Characteristics. GRAM's 30-second latency exceeds the baseline's 25 seconds, but total end-to-end resolution time favors GRAM by 60–66% (90–150s vs. 325–445s), as slide-level citations eliminate manual document-level factchecking. GRAM also supports files exceeding 50MB versus the baseline's 20MB limit, meaningful given that approximately 30% of pharmaceutical market research data exists as audio/video.

Domain-Specific Functional Advantages. GRAM demonstrated consistent pharmaceutical terminology handling across all 12 queries via ontology-driven entity recognition, whereas the baseline exhibited variable abbreviation recognition and failure to recognize brand-generic equivalence. SMEs characterized GRAM responses as precisely aligned with pharmaceutical decision-making frameworks versus more generic baseline outputs.

E. Business Impact Analysis

Analyst Productivity Gains. Traditional manual analysis requires 30–45 minutes per complex query. GRAM reduces this to 30 seconds (95% reduction), saving 50–75 analyst hours monthly for organizations processing 100 complex queries. Multi-Therapeutic Area Generalizability. Cross-product variance in average answer relevancy of 5.2 percentage points across oncology and women's health products, with all products exceeding 84% average relevancy and approaching or exceeding 90% pass rate, supports the system's therapeutic area-agnostic design, supporting unified deployment across diverse pharmaceutical portfolios without product-specific customization.

Strategic Decision Acceleration. Real-world deployment for product launch preparation processed over 200 market research documents spanning demand forecasts, competitive intelligence, physician interviews, and awareness studies, reducing pre-launch competitive analysis timelines from 2–3 weeks to 2–3 days.

Quality and Compliance Value. Source attribution accuracy of 0.99 ensures regulatory audit readiness with exact provenance (document, slide, page, timestamp), enabling defensible strategic decisions with full traceability. Nearelimination of hallucinations

(rate of 0.01) through graphconstrained generation prevents strategic errors from fabricated information, avoiding potential misdirected investment.

VII. CONCLUSION

We presented GRAM, a hybrid RAG framework integrating ontology-driven knowledge graphs with semantic vector search for pharmaceutical market research, addressing heterogeneous multimodal data, domain-specific terminology, relational entity understanding, and granular source attribution for regulatory compliance.

SME validation across three pharmaceutical products spanning oncology and women's health therapeutic areas demonstrated strong and consistent cross-therapeutic performance. Query evaluation achieved high pass rates and answer relevancy scores well above threshold, with near-perfect slide-level source attribution accuracy across all products. Functional capability testing confirmed production readiness across core pharmaceutical analyst workflows, with particularly strong performance on terminology comprehension, guardrails, and document filtering.

Head-to-head comparison showed substantial improvements across all retrieval metrics, a significant gain in source attribution from document-level to exact slide-level granularity, and a 92.9% reduction in hallucination rate through graphconstrained generation. Despite slightly higher generation latency, end-to-end resolution time substantially favors GRAM by eliminating manual fact-checking overhead.

Real-world deployment for product launch preparation processed over 200 market research documents, reducing competitive analysis timelines from weeks to days and cutting perquery analyst effort from tens of minutes to seconds.

The therapeutic area-agnostic design validates scalability across diverse portfolios from a single platform. Future work targets code execution integration for quantitative reasoning, enhanced multi-turn context management, and multi-lingual support. GRAM's combination of ontology-driven retrieval, multimodal ingestion, and graph-constrained generation establishes an enterprise-ready foundation for trustworthy pharmaceutical market research intelligence.

ABOUT THE AUTHORS

Yuan Niu is Director of AI and Machine Learning at Bayer Pharmaceuticals, where she spearheads enterprise-wide AI and advanced analytics strategies across global pharma teams. With deep expertise in scalable analytics deployment and cross-functional innovation, she drives measurable commercial and operational impact by translating advanced AI capabilities into real-world pharmaceutical business outcomes. Her work bridges data science, technology, and commercial strategy to accelerate Bayer's digital transformation initiatives.

Junying Zheng is Director of AI, ML, and Advanced Analytics at Bayer, overseeing the design and deployment of advanced analytics and machine learning solutions across research, commercial, and enterprise functions. He champions data-driven decision-making by enabling modern ML deployment at scale, delivering strategic insights that strengthen Bayer's competitive positioning across global markets. His work spans the full analytics lifecycle from model development to enterprise-wide integration.

Mathew Ratnam is Digital Capability Leader for AI and Analytics at Bayer Pharmaceuticals, driving strategic initiatives around AI factory build-out and agentic AI deployment at scale. He architects modular, scalable AI platforms that align cross-functional stakeholders and accelerate digital transformation across Bayer's global pharmaceutical operations. His focus on enterprise-grade agentic systems positions Bayer at the forefront of next-generation pharmaceutical intelligence.

Dr. PKS Prakash is a Partner at ZS Associates, bringing extensive expertise as a Data Scientist, Solution Architect, and Evangelist. His research and applied work spans healthcare, pharmaceutical, manufacturing, and banking industries, with deep specialization in predictive and prescriptive modelling, process simulations, fraud detection, and early warning systems. He has led diverse industrial and research initiatives including predictive maintenance in semiconductor manufacturing, virtual metrology, root cause analysis, and quality control. A recognized thought leader in advanced analytics, Dr. Prakash drives ZS's vision of building transformative AI solutions that deliver measurable impact across enterprise and life sciences applications.

Shivam Shivam is an Enterprise AI Architect at ZS Associates with 7+ years of expertise in Generative AI, LLMs, and

Multi-Agent systems for the pharmaceutical industry. He leads ZS's Tech Innovation Lab, driving transformative AI product development and client outcomes. A Board of Governors member at IEEE Computer Society, he is a recipient of the IEEE Young Professionals Hall of Fame Award 2023 and is recognized among IEEE CS "20s in 20s" for impactful industry contributions. He also serves as an adjunct faculty member, teaching Responsible AI and Generative AI.

Swaraj Prakash is a Senior Engineer at ZS Associates specializing in Generative AI, NLP, and large-scale intelligent systems for healthcare and life sciences. He brings deep expertise in building end-to-end AI pipelines, integrating LLMs into enterprise workflows, and designing decision-support tools that convert unstructured pharmaceutical data into actionable business insights.

Rishav Mishra is a Technology Associate at ZS Associates building scalable intelligent pharmaceutical AI systems, transforming complex data into actionable insights and real-world impact.

REFERENCES

1. Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, et al. Retrieval-Augmented Generation for AI-Generated Content: A Survey. Data Science and Engineering. 2026 01:1-29.
2. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, et al. Gpt-4o system card. arXiv preprint arXiv:241021276. 2024.
3. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.
4. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. 2024;15(3):1-45.
5. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:210407567. 2021.
6. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.
7. Finardi P, Avila L, Castaldoni R, Gengo P, Larcher C, Piau M, et al. The Chronicles of RAG: The Retriever, the Chunk and the Generator. arXiv preprint arXiv:240107883. 2024.
8. Wu J, Zhu J, Qi Y, Chen J, Xu M, Menolascina F, et al. Medical graph rag: Towards safe medical large language model via graph retrievalaugmented generation. arXiv preprint arXiv:240804187. 2024.
9. Wang C, Long Q, Meng X, Cai X, Wu C, Meng Z, et al. BioRAG: A RAG-LLM Framework for Biological Question Reasoning. arXiv preprint arXiv:240801107. 2024.
10. Kim J, Min M. From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process. arXiv preprint arXiv:240201717. 2024.
11. Baek J, Aji AF, Saffari A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. arXiv preprint arXiv:230604136. 2023.
12. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv; 2022. Available from: https://arxiv.org/abs/2212.04356.
13. Carroll JJ, Bizer C, Hayes P, Stickler P. Named graphs, provenance and trust. In: Proceedings of the 14th International Conference on World Wide Web. ACM; 2005. p. 613-22.
14. Xiao S, Liu Z, Zhang P, Muennighoff N. C-Pack: Packaged Resources to Advance General Chinese Embedding. arXiv preprint arXiv:230907597. 2023.

15. Robertson S, Zaragoza H, Taylor M. Simple BM25 extension to multiple weighted fields. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management; 2004. p. 42-9.

16. Ramos J, et al. Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242. Citeseer; 2003. p. 29-48.

17. Ali M, Jung LT, Abdel-Aty AH, Abubakar MY, Elhoseny M, Ali I. Semantic-k-NN algorithm: An enhanced version of traditional k-NN algorithm. Expert Systems with Applications. 2020;151:113374.

18. Humeau S, Shuster K, Lachaux MA, Weston J. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:190501969. 2019.

19. Dong J, Fatemi B, Perozzi B, Yang LF, Tsitsulin A. Don't Forget to Connect! Improving RAG with Graph-based Reranking. arXiv preprint arXiv:240518414. 2024.

20. Manning CD, Raghavan P, Sch¨utze H. Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press; 2008.

21. Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies. 2011;2(1):37-63.

22. J¨arvelin K, Kek¨al¨ainen J. Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems. 2002;20(4):422-46.

23. Voorhees EM, Tice DM. The TREC-8 Question Answering Track Report. In: Proceedings of the 8th Text Retrieval Conference (TREC-8). NIST; 1999. p. 77-82.

24. Es S, James J, Espinosa-Anke L, Schockaert S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:230915217. 2023.

Accelerating Pharmaceutical Commercial Analytics Through a Generative AI Agent Hub and Semantic Data Discovery

Shihan He, Sai Rithvik Kanakamedala, Akash Pandey, Anubhav Srivastava

ABSTRACT

Objective: In the highly competitive and data-rich pharmaceutical industry, commercial analytics teams face a persistent and significant bottleneck: the "Data Discovery" and preparation phase. Industry estimates and internal operational audits suggest that highly skilled data scientists spend up to 80% of their project lifecycle on data wrangling—struggling to locate correct tables, decipher obscure column names, reconcile inconsistent metric definitions, and construct complex, multi-join SQL queries across heavily fragmented databases. This operational inefficiency is severely exacerbated by "tribal knowledge," where the critical understanding of nuanced data schemas and business rules resides with a select few veteran employees rather than a centralized, accessible system. This practice article outlines the conceptualization, development, and enterprise implementation of a Generative Artificial Intelligence (GenAI) "Semantic Layer" and an encompassing "Agent Hub" designed to drastically reduce friction in feature discovery and accelerate the time-to-model for commercial machine learning projects.

Approach: This study details the construction of a Natural Language Feature Store Assistant, seamlessly integrated directly into enterprise collaboration platforms via an intelligent Agent Hub. The system utilizes a sophisticated dual-agent "Text-to-SQL" architecture to translate natural language queries from data scientists and business analysts—such as "generate a dataset of top-decile cardiovascular writers with >100 NBRx scripts in New Jersey over the last 12 months"—into highly optimized, executable SQL code. The core innovation of this platform lies in the "Semantic Layer": a rigorous, engineered mapping of complex, real-world schemas into an AI-interpretable format hosted within Dataiku Data Collections. Rather than relying solely on raw, often cryptic database metadata, the system was trained on explicitly defined business metrics to ensure context-aware query generation. Furthermore, the Agent Hub integrates unstructured knowledge retrieval via Retrieval-Augmented Generation (RAG), enabling qualitative strategic insights alongside quantitative data extraction. The implementation included a stringent "human-in-the-loop" verification process where the AI's output was benchmarked against 50 distinct technical queries representing standard analytical workflows.

Key Insights: Preliminary deployment results indicate that a well-architected semantic layer, coupled with an agentic workflow, significantly reduces the cognitive load on data scientists and drastically shrinks the "Time-to-Data" metric, in some cases reducing days of manual querying to minutes of automated generation. By automating the boilerplate code required for complex table joins (often involving three or more distinct, siloed data sources), the tool effectively converts a manual, error-prone, multi-day process into an automated, self-serve service. Crucially, the project demonstrates that success in Generative AI for enterprise analytics depends far less on utilizing the most advanced foundational Large Language Model (LLM) architecture and far more on "Schema Engineering"—the structured, semantic definition of relationships between business entities. The evaluation highlights the paramount importance of creating a "Single Source of Truth" for metric definitions, ensuring that generated features are mathematically and logically consistent across disparate modeling projects.

Relevance: As pharmaceutical companies rapidly scale their data assets—incorporating increasing volumes of real-world evidence (RWE), social determinants of health (SDOH), and omnichannel marketing data—the ability to rapidly discover, understand, and utilize data becomes a core competitive advantage. This submission provides a practical, tested blueprint for analytics leaders, Chief Data Officers, and IT architects seeking to modernize their data infrastructure. It demonstrates how management science and software engineering principles can be applied to internal data workflows, ultimately freeing high-value data scientists to focus on strategic algorithm development, predictive modeling, and business advisory, rather than basic data retrieval.

1. INTRODUCTION

1.1 The Data Bottleneck in Pharmaceutical Analytics

The modern pharmaceutical commercial analytics landscape is characterized by its immense complexity, sheer volume, and extreme fragmentation. To effectively commercialize a pharmaceutical asset, organizations rely on vast arrays of interconnected data points to understand patient journeys, prescriber behavior, and market access dynamics. This ecosystem typically includes Health Care Professional (HCP) prescribing behaviors, payer formulary statuses, anonymized longitudinal patient claims data, specialty pharmacy transaction logs, and internal omnichannel marketing touchpoints captured via Customer Relationship Management (CRM) systems.

While this proliferation of data represents a massive strategic asset—enabling precise targeting, dynamic predictive modeling, omnichannel marketing optimization, and efficient resource allocation—it paradoxically introduces a severe operational bottleneck. The data is rarely clean, centralized, or natively interoperable. Highly compensated data scientists and commercial analysts routinely spend up to 80% of their project timelines identifying, verifying, extracting, and cleaning data before any meaningful exploratory data analysis (EDA) or machine learning modeling can commence. This phenomenon is often colloquially referred to as "data janitor work," a reality that diminishes the return on investment (ROI) of advanced data science teams and delays critical business decisions.

1.2 The Economic and Opportunity Cost of Data Friction

The cost of this bottleneck extends beyond mere frustration; it manifests as a significant opportunity cost. In the pharmaceutical sector, where the launch window of a new therapy is critical and market dynamics shift rapidly, the inability to quickly access and model data can lead to lost market share. For example, if a commercial team wishes to build a predictive model to identify HCPs likely to switch their prescribing habits due to a competitor's recent formulary loss, the data science team must rapidly aggregate claims, payer, and CRM data. If it takes three weeks simply to construct the underlying analytical dataset, the window of opportunity to deploy an agile marketing campaign may have already closed. Furthermore, executing unoptimized, manual SQL queries against vast cloud data warehouses (like Snowflake or Amazon Redshift) incurs substantial compute costs, making efficient query generation a financial imperative.

1.3 The "Tribal Knowledge" Dilemma and Schema Drift

This friction is rarely a technical limitation of the computing infrastructure, but rather

a profound knowledge management failure. In most large pharmaceutical organizations, complex data relationships and nuanced business definitions are poorly documented and constantly evolving—a phenomenon known as "schema drift." For instance, determining how to correctly calculate "Brand Churn" for a specific therapeutic area, or identifying which specific National Provider Identifier (NPI) master table to join against a raw retail claims table to avoid duplicating prescriber counts, relies heavily on undocumented "tribal knowledge."

This knowledge lives almost exclusively in the minds of veteran data engineers, data stewards, and senior analysts who have spent years navigating the idiosyncratic corporate data swamps. When a new data scientist or analyst joins a project, they must embark on a time-consuming "scavenger hunt." They navigate cryptic, legacy column names, search through outdated and abandoned data dictionaries, and ping colleagues on messaging platforms to verify if a particular table is the accepted "single source of truth" for a given metric. This reliance on human bottlenecks creates fragile, unscalable data operations.

1.4 The Promise and Peril of Generative AI

The advent of Large Language Models (LLMs) and Generative AI presented a theoretical solution to this epistemological gap: "Text-to-SQL" generation. Early experiments in the industry attempted to point raw, out-of-the-box LLMs directly at cloud data warehouses, hoping analysts could simply chat with their databases using natural language.

However, these early isolated experiments largely failed in production environments. Raw LLMs, lacking specific business context and enterprise domain knowledge, frequently suffered from "hallucinations." They would invent plausible-sounding but non-existent column names, execute logically flawed inner joins that inadvertently dropped critical cohort data, or fail to apply standard, unstated business exclusions.

To solve this, we recognized that AI alone is insufficient without a robust, highly structured architectural framework. We developed a centralized **Enterprise Agent Hub**, leveraging Generative AI not as a standalone magic bullet, but as an intelligent orchestration engine resting on top of a rigorously engineered "Semantic Layer." This paper outlines the methodology, architecture, evaluation framework, and business impact of deploying this Agent Hub to democratize data access and automate complex data retrieval workflows across the commercial analytics enterprise.

2. BACKGROUND AND RELATED WORK

2.1 The Evolution of Data Discovery and Business Intelligence

Historically, data discovery in pharmaceutical commercial operations relied on enterprise data dictionaries, static code repositories, and traditional Business Intelligence (BI) dashboards. While BI tools excel at displaying pre-computed Key Performance Indicators (KPIs) to executive stakeholders, they are inherently inflexible for advanced analytics.

If a data scientist needs a custom, highly specific feature for a machine learning model—for example, "the standard deviation of weekly script writing for a specific HCP over a rolling 6-month window, segmented by tier-1 payer

status and prior competitor usage"—a static BI dashboard cannot provide this. The data scientist must bypass the BI layer and drop down into raw SQL to extract the data from the data lake or warehouse. Enterprise data catalogs provide necessary metadata visibility and governance, but they are generally passive; they do not actively assist the user in writing the complex extraction code, leaving a significant, labor-intensive gap between *finding* the data and *using* the data.

2.2 Text-to-SQL and the Ontology Mapping Problem

Text-to-SQL is a heavily researched area within Natural Language Processing (NLP). Standard academic benchmarks like Spider, WikiSQL, and BIRD measure an LLM's ability to translate natural language into SQL queries against standard, cleanly normalized relational databases. However, enterprise databases—particularly in pharma—are rarely clean or perfectly normalized. They contain thousands of wide tables, denormalized views, legacy naming conventions, and complex, multi-layered schemas.

The core challenge in applying LLMs to enterprise databases is the "Ontology Mapping Problem." An LLM understands the general English word "Sales," but it does not intrinsically know that in a specific company's database, "Gross Sales" resides in while "Net Sales" requires joining that view to subtract managed care rebates.

The concept of a "Semantic Layer" emerged as a middleware solution to bridge this gap. A semantic layer translates raw physical data structures into logical, human-readable business concepts. By injecting a semantic layer between the LLM and the database, organizations can drastically improve Text-to-SQL accuracy. The LLM is forced to reason over the curated business concepts, and the semantic engine handles the translation to the underlying physical SQL schema.

2.3 Multi-Agent Systems and Retrieval-Augmented Generation (RAG)

Recent advancements in AI architectures have shifted away from simple, single-prompt interactions toward complex "Agentic" workflows. In a multi-agent system, specific, narrow tasks are delegated to specialized AI agents, each possessing restricted scopes, specific system prompts, and access to specific tools. For instance, in a coding task, one agent acts as the system planner, another writes the specific code syntax, and a third acts as a critic to debug and optimize the output. This division of labor significantly reduces cognitive overload on the LLM and reduces the rate of hallucination.

Simultaneously, Retrieval-Augmented Generation (RAG) has emerged as the industry standard for allowing LLMs to accurately query unstructured data. By converting unstructured documents (PDFs, video transcripts, internal slide decks, SOPs) into mathematical vectors (embeddings) and storing them in a vector database, RAG allows an LLM application to retrieve only the most semantically relevant snippets of information to answer a user's prompt. This effectively grounds the model's response in proprietary, verified enterprise data, bypassing the LLM's pre-trained parametric memory and preventing it from hallucinating external information. Our Enterprise Agent Hub synthesizes both Multi-Agent Text-to-SQL for quantitative tabular data and robust RAG architectures for qualitative, unstructured data.

3. SYSTEM ARCHITECTURE: THE ENTERPRISE AGENT HUB

To move beyond isolated, local AI experiments and fragmented, siloed point-solutions, we engineered a holistic **Agent Hub**—a centralized, intelligent routing system designed to be highly accessible via standard enterprise interfaces. To maximize adoption and integrate seamlessly into existing workflows, the Hub is deployed as a conversational interface within Microsoft Teams, as well as an advanced interface on a dedicated internal data portal.

The Agent Hub operates on a fundamental premise: it acts as the user's unified, single point of contact for any data-related or knowledge-related inquiry. The architecture is composed of an intent classification router and a dynamic suite of specialized, backend AI agents. Chief among these is the "Data Agent," which is explicitly equipped with a specific toolset to handle analytical workflows.

3.1 The Intelligent Router and the Data Agent Toolset

When a user submits a natural language request, the Hub first utilizes an Intent Classification LLM. This component uses heuristic-based routing to deeply analyze the prompt and determine the required downstream execution path. If the request is analytical, it is routed to the **Data Agent**.

The Data Agent does not rely on generic reasoning; it is equipped with a highly specific arsenal of functional tools tailored for analytics:

- **Table Search:** A specialized Knowledge Bank Search tool used to query the semantic layer and locate the correct database structures based on business terminology.
- **Dataset Lookup:** A tool used to preview existing, materialized data products within the enterprise catalog.
- **SQL Query:** The generation and execution engine, allowing the agent to formulate and run code against the cloud warehouse natively.
- **Auxiliary Tools:** The agent also possesses access to tools like *Google Search* (for external fact-checking or macro-economic data) and *Send Message* (to push notifications or datasets directly to users via enterprise messaging platforms).

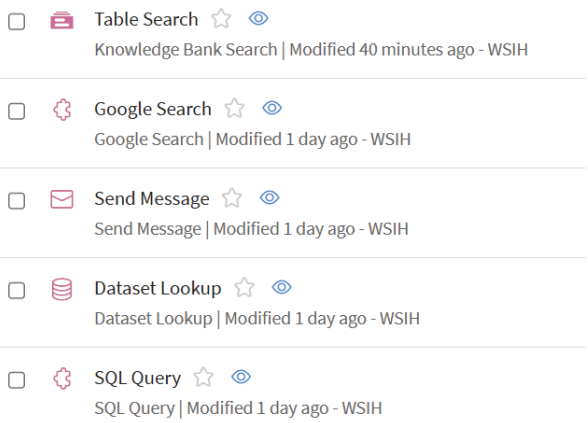


Figure 1. *The available toolset configuration within the Agent Hub.*

3.2 Engineering the Semantic Layer Pipeline

The foundational success and high accuracy of our quantitative data retrieval system stem directly from how we engineer and embed the semantic model that powers the *Table Search* tool. This is managed through a sophisticated, automated pipeline built within **Dataiku**.

Rather than manually typing data dictionaries, we built a visual flow that ingests, standardized, and embeds schema metadata directly from Snowflake into a vector knowledge bank.

The Dataiku Semantic Pipeline Flow:

1. **Ingestion of Domain Schemas:** The pipeline begins by extracting the raw metadata (table names, column names, data types, and human-inputted descriptions) from diverse, highly specialized physical databases. As modeled in our architecture, this includes disparate data sources such as:
2. **Preparation and Standardization:** Each raw schema passes through a preparation node. In this phase, automated scripts standardize naming conventions, apply business aliases and append mandatory inclusion/exclusion logic filters directly to the metadata.
3. **Stacking and Unification:** All the prepared, domain-specific schemas are routed into a central stacking node, which unions them into a massive, unified dataset called `SCHEMA_metadata`. This creates the "Single Source of Truth."
4. **Embedding Generation:** After a final preparation step the unified metadata is passed through a deep learning embedding model. The output is `SCHEMA_metadata_prepared_embedded`—a high-dimensional vector database. This embedded knowledge bank acts as the exclusive "brain" for the *Table Search* tool.

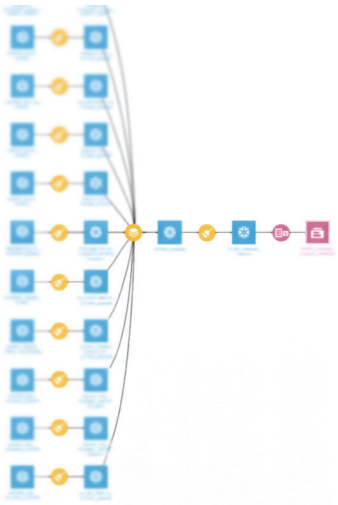


Figure 2. *The visual architecture of the semantic pipeline.*

3.3 The Agentic Text-to-SQL Workflow

When the Data Agent receives a complex user prompt (e.g., *"Pull the top-decile Brand X prescribers in the Northeast over the last 6 months, showing their total script volume and latest CRM call plan interactions"*), it orchestrates a precise multi-tool workflow. Crucially, the system is designed to act collaboratively, utilizing the "Table Search" agent to map logic and the "SQL Query" agent to generate and execute code, culminating in a natural language response.

Step 1: The Ontologist Phase (Using Table Search)

The Data Agent first invokes the **Table Search** tool. It queries the `SCHEMA_metadata_prepared_embedded` knowledge bank using the business terms from the user's prompt. Because the semantic layer was meticulously engineered in Dataiku, the tool easily connects "Brand X prescribers" to `BRAND_X_HCP_FEATURES_SCHEMA` and "CRM call plan" to `CRM_CALL_PLAN_SCHEMA`. The tool resolves the entity relationships, identifying the exact Primary Key / Foreign Key relationships needed to bridge the sales data with the CRM data without data explosion.

Step 2: The Code Generation Phase (Using SQL Query)

Armed with the precise, verified schema mapping from the Table Search tool, the Data Agent then transitions to the **SQL Query** tool. Operating as an expert SQL developer, it generates the highly optimized, dialect-specific executable code. It constructs the complex `SELECT`, `INNER JOIN`, and `LEFT JOIN` clauses precisely as dictated by the semantic map. It applies temporal `WHERE` clauses and executes robust `GROUP BY` logic.

Step 3: Backend Execution and Natural Language Synthesis

The workflow does not simply dump a block of code or a raw data grid to the user. The **SQL Query** tool executes the generated code against the backend cloud data warehouse. Once the result set is retrieved, the overarching Data Agent acts as an analyst: it synthesizes the numerical results into a clear, natural language summary.

Importantly, for the sake of auditability, trust, and advanced user customization, the underlying SQL code that was executed is prominently displayed within the UI alongside the natural language answer. This allows data scientists to verify the agent's logic or copy the SQL for deeper integration into their local Python/R pipelines.

3.4 Unstructured Knowledge Retrieval: Integrating Internal Knowledge Banks

The true transformative power of the Enterprise Agent Hub is greatly magnified by its multi-modal versatility. Data scientists, marketing directors, and commercial leaders do not solely require rows of quantitative tabular data; they require qualitative context to inform their modeling and strategic decisions. To support this, the Hub is equipped with sophisticated RAG pipelines to query unstructured, proprietary enterprise data.

A prime, production-tested example of this functionality involves extracting insights from embedded corporate video transcriptions. Pharmaceutical leadership frequently communicates major strategic shifts—such as changes in direct-to-consumer marketing approaches, insights into drug launch strategies, or regulatory compliance updates—via internal town halls, external investor calls, and executive interviews. Historically, these invaluable strategic insights are locked inside hour-long video files, requiring extensive manual review and note-taking to extract key points.

We developed an automated pipeline to process this media:

1. **Transcription:** Internal videos are passed through high-fidelity speech-to-text models.
2. **Chunking:** The resulting text is intelligently segmented into logical chunks based on topic shifts and paragraph breaks, rather than arbitrary token limits.
3. **Embedding:** These text chunks are processed through dense embedding models (transforming text into high-dimensional vector representations) and stored in a highly scalable vector database, designated as the knowledge bank.

Case Study: Executive Strategy Retrieval When a user interacts with the Hub and submits a qualitative prompt such as: *"Summarize the interview,"* or more specifically, *"What did leadership say about consumer communication?"*, the Intelligent Router identifies this as Intent B (Qualitative) and directs it to the Unstructured Knowledge Agent.

The Agent performs a semantic vector search against the database, retrieving the chunks most mathematically similar to the user's query. As demonstrated in a live test case involving an internal summary of an interview with executive leadership, the Hub successfully synthesized the raw transcript into a concise, highly accurate, bulleted summary.

The Hub expertly extracted and organized key strategic themes from the executive dialogue, specifically noting:

- **Communication Strategies of Pharma Companies:** Highlighting that the industry is recognizing the critical importance of effective, direct communication with consumers.
- **Direct-to-Consumer Shift:** Detailing the strategic pivot toward providing more educational information directly to consumers about their healthcare options, moving away from relying solely on traditional, physician-mediated legacy systems.
- **Balance in Advertising:** Capturing the nuanced discussion regarding the crucial need to balance the volume of pharmaceutical advertising, ensuring the core message is conveyed without overwhelming, alienating, or annoying the public.
- **Corporate Role:** Summarizing the specific, proactive strategic approaches taken by leadership to continuously adapt to consumer needs and improve the broader dialogue between pharma companies and the public.

To ensure transparency and combat hallucination, the Agent Hub interface transparently cites its specific sources at the bottom of the response□This builds vital trust with the business user, allowing them to trace the origin of the qualitative insights back to the source material if they need to verify the context. This capability proves the Hub's ability to act as a universal, multi-modal assistant for both quantitative feature engineering and qualitative strategic research.

4. IMPLEMENTATION AND VERIFICATION METHODOLOGY

To ensure enterprise-grade reliability and build essential trust among highly technical data science and data engineering teams, the deployment of the Agent Hub was gated behind a rigorous, multi-phase "human-in-the-loop" verification framework. We established an evaluation benchmark designed specifically for the unique complexities of pharmaceutical commercial analytics.

4.1 The 50-Query Benchmark Suite

Before exposing the tool to general users, we curated a comprehensive test suite of 50 distinct technical queries. These were not synthetic examples; they were drawn from historical Jira tickets representing standard, complex analytical workflows routinely executed by our commercial analytics teams. These queries were strictly stratified by complexity:

- **Class 1 (Low Complexity - 15 Queries):** Simple data retrieval, basic `WHERE` filtering, and straightforward aggregations from single tables. (Example: *"Show me total script volume and gross sales by state for Q1 2023 for Product Alpha."*)
- **Class 2 (Medium Complexity - 20 Queries):** Queries requiring 2-3 table joins, basic time-series date math filtering, and metric segmenting or deciling. (Example: *"Find all HCPs in the Southeast region who have increased their prescribing of Product X by more than 10% month-over-month, joined to their primary specialty."*)

- **Class 3 (High Complexity - 15 Queries):** Advanced queries requiring complex multi-source joins (4+ tables including master rosters, transactional claims, CRM call logs, and payer data), complex SQL window functions (e.g., `ROW_NUMBER() OVER (PARTITION BY...)`), sub-queries, and highly specific inclusion/exclusion criteria. (Example: *"Generate a cohort of Tier 1 cardiologists who have written a script for our primary competitor in the last 60 days, but have a favorable, unrestricted formulary status for our new launch product, while explicitly excluding any HCPs on the internal compliance do-not-engage list or who are marked as retired."*)

4.2 Creating the "Gold Standard" and Evaluation Metrics

For each of the 50 queries, senior data engineers manually wrote highly optimized "Gold Standard" SQL scripts. The AI's generated output for each query was executed in a sandboxed, isolated development environment and systematically evaluated against this Gold Standard. The evaluation focused on three primary axes:

- **Accuracy of Entity Resolution and Table/Column Identification (Semantic Accuracy):** Did the Table Search tool successfully locate the correct "Single Source of Truth" tables within the Dataiku Semantic Layer? Did it correctly map the colloquial business terminology to the physical metadata without hallucinating tables?
- **Syntactic Correctness (Execution Success Rate):** Did the SQL Query tool produce perfectly valid, syntactically correct SQL that compiled and executed without throwing compilation errors in the cloud data warehouse?
- **Data Output Correctness (Semantic Completeness):** Upon successful execution, did the resulting dataset mathematically align with the accepted business definition? This was measured via automated row-by-row parity checks. Did the final row counts, column structures, and aggregate financial sums match the Gold Standard output identically?

4.3 Human-in-the-Loop Feedback Interface and UX

During the initial beta rollout phase, the Agent Hub was deployed with an interactive, mandatory human-in-the-loop feedback mechanism embedded in the UI. When the Agent Hub returned a generated SQL query or a preview dataset, the requesting data scientist was prompted to review the logic before full execution against massive data tables. The User Interface (UI) allowed them to:

- Accept the query as is and execute.
- Manually edit the generated SQL directly in an inline code editor before execution to fix minor nuances.
- Provide conversational natural language feedback to correct the AI (e.g., *"You used the* `retail_claims_fact` *table, but this specific analysis requires the* `specialty_pharmacy_fact` *table."*) The AI would then regenerate the code based on the feedback.

This continuous feedback loop was rigorously logged, captured, and periodically reviewed by the central Data Engineering team. This data was used to continuously refine and patch the Dataiku Semantic Layer, adding new natural language aliases and refining table relationships based on real-world usage patterns and user corrections.

5. KEY FINDINGS AND DISCUSSION

5.1 Schema Engineering Surpasses Foundational Model Architecture

A primary, perhaps counter-intuitive, insight derived from this enterprise implementation is that the success of Generative AI in complex enterprise analytics relies far more heavily on **Schema Engineering** than on the specific capabilities of the underlying foundational LLM architecture (whether utilizing GPT-4, Claude 3, or Gemini models).

Early proof-of-concept tests utilizing state-of-the-art models *without* a robust, well-defined semantic layer yielded an abysmal execution success rate of roughly 40%. These early attempts were plagued by hallucinations, incorrect assumptions about data architecture, and a failure to grasp nuanced pharma business rules.

However, once the Dataiku Semantic Layer pipeline was fully implemented, stacking schemas into the embedded knowledge bank, and the Data Agent was strictly constrained to only operate using the Table Search tool over that engineered ontology, execution success rates on Class 1 and Class 2 queries skyrocketed to over 92% accuracy on the first attempt.

This decisively proves that the LLM is merely the linguistic reasoning engine; the Semantic Layer is the indispensable, accurate map. By treating Dataiku Data Collections as a rigorously maintained semantic layer—translating complex, real-world commercial logic into machine-readable JSON schema definitions—we effectively eliminated the AI's tendency to invent tables or misinterpret fundamental metrics.

5.2 Dramatic Reduction of Cognitive Load and Accelerated "Time-to-Data"

By automating the tedious boilerplate code required for data extraction, the system successfully transformed a manual, error-prone, multi-day process into a streamlined, automated task taking minutes.

Prior to the deployment of the Agent Hub, a data scientist tasked with creating a new predictive feature set (e.g., combining historical prescribing history with current payer formulary status and recent sales rep CRM touchpoints) would spend an estimated 2 to 4 days locating the disparate data, writing the complex 5-way SQL joins, debugging syntax errors, and consulting with data stewards to verify the business logic of the exclusions.

With the Agent Hub, the same task is accomplished in approximately 5 to 15 minutes. Internal audits revealed that average feature engineering and data gathering phases were reduced from an average of 36 manual working hours down to a median of roughly 12 minutes. The data scientist types the prompt in plain English, reviews the automatically generated, highly optimized SQL, executes the code, and immediately possesses the required dataset ready for ingestion into Python or R.

Users across the analytics organization reported a profound reduction in cognitive load and task fatigue. Rather than expending significant mental energy on memorizing massive database schemas, idiosyncratic naming conventions, and specific SQL dialect syntax, data scientists are now able to shift their focus significantly up the value chain. They can dedicate the vast majority of their time to the work they were hired to do: exploratory data analysis, feature engineering optimization, and developing advanced machine learning models (such as Next-Best-Action recommendations or Predictive Attrition models) that directly impact the commercial bottom line and improve patient outcomes.

5.3 Eradicating Tribal Knowledge via the "Single Source of Truth"

The implementation of this system forced a necessary, and sometimes difficult, cultural shift regarding enterprise data governance. In order to physically build the Semantic Layer, data stewards, business analysts, and senior engineers were required to explicitly define, debate, and document business rules that had previously only existed as implicit tribal knowledge.

Because the AI strictly adheres to the definitions hardcoded in the Semantic Layer, it natively enforces a standard, enterprise-wide "Single Source of Truth." If two different data scientists in different business units ask the Agent Hub to calculate "Total NBRx Market Share for Brand X" on different days, the system generates identical SQL logic based on the centralized definition.

This consistency is vital in pharmaceutical analytics, where discrepancies in metric calculations between departments can lead to flawed strategic decisions, misaligned sales compensation, or regulatory compliance issues. The Agent Hub effectively democratized the deep domain knowledge of the most senior data engineers, making it instantly, reliably accessible to every authorized member of the analytics team.

---

6. CHALLENGES, LIMITATIONS, AND MITIGATION STRATEGIES

While the deployment of the Agent Hub has been highly successful and transformative, the project surfaced several notable challenges inherent in applying Generative AI to complex, highly regulated enterprise data environments.

6.1 Managing Ambiguity in Human Prompts

The most frequent point of failure in the production system arises not from AI hallucination or technical bugs, but from human ambiguity. Business users and analysts frequently submit prompts that are logically incomplete, grammatically poor, or utilize highly colloquial acronyms. For example, a prompt like *"Pull last year's sales for our top drug"* is highly problematic for an automated system. What explicitly defines "last year" (calendar year vs. fiscal year, rolling 12 months vs. Year-to-Date)? What defines "sales" (gross revenue, net revenue, total prescriptions, wholesale acquisition cost)? Which specific metric determines the "top drug"?

Mitigation Strategy: We implemented a dynamic "clarification loop" within the agent workflow. If the Table Search tool detects that multiple conflicting interpretations exist within the Semantic Layer for a given vague prompt, the Data Agent is programmed to pause generation and utilize the *Send Message* tool to actively ask the user a clarifying question in the chat interface (e.g., *"I found multiple definitions. Do you want Gross Sales in USD or Net Sales? By 'last year', do you mean Fiscal Year 2023 or the previous 12 rolling months?"*) rather than making an unverified assumption that could lead to flawed data.

6.2 Context Window Limits, Compute Costs, and Schema Size

Pharmaceutical data models are extraordinarily massive. A single enterprise data warehouse may contain thousands of tables and tens of thousands of columns. Passing an entire corporate data dictionary to an LLM as context in the system prompt is computationally expensive, induces high latency, and often exceeds the maximum token context limits of modern LLMs. Supplying too much irrelevant

context also leads to a degradation in the model's ability to focus on the relevant information (the "needle in a haystack" problem).

Mitigation Strategy: This reinforced the absolute necessity of utilizing the `SCHEMA_metadata_prepared_embedded` vector database. Rather than mapping the entire raw warehouse into the prompt, the Table Search tool utilizes a vector search against the metadata catalog to dynamically retrieve *only* the specific table schemas relevant to the user's keywords. This significantly reduces the required prompt size, minimizes token costs, and vastly improves processing latency before code generation begins.

6.3 Data Security, Privacy, and Governance (HIPAA/PHI)

In the pharmaceutical industry, strict adherence to data privacy regulations (such as HIPAA in the US and GDPR in Europe) is non-negotiable. Connecting an LLM to healthcare data presents potential security vectors.

Mitigation Strategy: The architecture was explicitly designed so that the LLM agents **never** see or process row-level database data, Protected Health Information (PHI), or Personally Identifiable Information (PII). The agents only have access to the *metadata* (table names, column names, business definitions) residing in the Semantic Layer. The LLM generates the SQL code, but the code is executed locally within the company's secure cloud perimeter. The actual patient or prescriber data never traverses the internet to the LLM provider's API.

6.4 Maintenance of the Semantic Layer

The Agent Hub is ultimately only as intelligent and accurate as its underlying Semantic Layer. As the pharmaceutical business evolves—launching new therapeutic drugs, acquiring new data sources, updating territory alignments, or changing standard metric definitions—the Semantic Layer must be continuously, rigorously updated. If the metadata becomes stale, the AI will confidently generate outdated, incorrect, or deprecated queries.

Mitigation Strategy: We deeply integrated the Semantic Layer maintenance into our standard DataOps and CI/CD lifecycle. Whenever a data engineer modifies a core table in the Dataiku repository or the Snowflake warehouse, they are required to update the associated flow and re-trigger the stacking and embedding process. The Semantic Layer is treated as living production code, subject to strict version control, automated testing, and human peer review.

---

7. FUTURE WORK AND ROADMAP

The current iteration of the Enterprise Agent Hub serves primarily as an advanced, highly capable data discovery and retrieval assistant. However, the secure architectural foundation laid by this project opens several robust avenues for future enhancement and scaling across the enterprise.

- **Automated Feature Engineering and Suggestion:** Transitioning the Hub from simple data retrieval to active data transformation and ideation. Future iterations of the Hub will proactively suggest novel features to data scientists based on the requested target variable. (e.g., *"I see you are building a prescriber churn model for Product X; based on historical models, would you like me to automatically generate and append a 3-month rolling script velocity feature and a payer-tier volatility index to your dataset?"*).

- **Autonomous Predictive Modeling Agents:** Integrating advanced agentic loops capable of not just writing SQL, but generating complete Python or R pipelines. These agents could automatically train baseline machine learning models (AutoML) directly within the secure Dataiku environment based on the datasets they just retrieved, instantly providing a baseline model for the data scientist to refine.
- **Proactive Anomaly Detection and Insight Generation:** Evolving the Hub from a purely reactive query system (waiting for user prompts) to a proactive monitoring agent. The Hub could continuously monitor the semantic layer's underlying data and alert commercial leaders to statistically significant deviations in key metrics (e.g., an unexpected 15% drop in regional NBRx sales volume), accompanied by an automatically generated natural language summary of the likely underlying drivers based on CRM and claims data analysis.
- **Cross-Functional Integration:** Expanding the Semantic Layer beyond Commercial Analytics to include Clinical Operations and R&D datasets, allowing the Hub to bridge the historical divide between pre-launch clinical trial insights and post-market real-world evidence.

8. CONCLUSION

As global pharmaceutical organizations continue to aggressively scale their commercial data assets and analytical ambitions, removing the operational friction from the initial "Data Discovery" and preparation phase is no longer an optional luxury; it is a critical organizational necessity for maintaining operational agility and competitive advantage in a fast-paced market.

The conceptualization, rigorous testing, and enterprise deployment of the Generative AI Agent Hub—combining Dataiku-powered structured SQL generation with RAG-enabled unstructured knowledge retrieval—provides a highly effective, production-proven blueprint for modernizing legacy pharmaceutical data infrastructure.

By taking the difficult but necessary step of explicitly codifying elusive, undocumented "tribal knowledge" into a rigorous Generative AI Semantic Layer, and architecturally separating the semantic reasoning engine from the syntax generation engine, analytics leaders can securely democratize data access. This approach enforces a rigorous, enterprise-wide single source of truth, drastically reduces code errors, and exponentially accelerates the analytical lifecycle.

Ultimately, this paradigm shift empowers high-value data scientists and analysts to fundamentally transition away from the tedious, low-value labor of data gathering and pipeline debugging. It allows them to dedicate their specialized expertise to strategic algorithm development, generating the critical predictive and actionable insights required to successfully optimize omnichannel marketing efforts, deeply understand patient journeys, and rapidly commercialize life-saving therapies in a complex, continuously evolving, data-driven global market.

FLARE: Forecasting with LLM-Assisted Recommendation Engine

PKS Prakash, ZS Associates, India; Srinivas Chilukuri, ZS Associates, US

Abstract: Modern forecasting pipelines often rely on exhaustive evaluations across large model libraries, motivated by evidence from the M4 and M5 competitions that hybrid and ensemble approaches outperform single models. However, brute-force search becomes computationally prohibitive as the number of models and time series grows. For example, the M4 competition alone included 61 distinct forecasting methods. As a result, practitioners are often forced to settle for suboptimal model selections. This paper proposes Forecasting with LLM-Assisted Recommendation Engine (FLARE), a retrieval-augmented model selection framework that combines statistical characterization, learned model priors, and large language model (LLM) reasoning to identify a small set of promising forecasting candidates without training all candidates.

FLARE builds a knowledge base by extracting statistical features from historical series and recording empirically strong single models and ensembles. A dynamic model-count predictor estimates the appropriate shortlist size for each new series. During inference, statistical descriptors, neighborhood evidence, and task constraints are provided to an LLM, which produces a ranked shortlist of candidate models. Only this reduced set is evaluated by the forecasting engine, substantially lowering computational cost.

The proposed FLARE approach is evaluated on M4 monthly and quarterly univariate time-series using Mean Absolute Percentage Error (MAPE) and latency as key evaluation metrics. On the monthly benchmark, FLARE achieves over ~9.1× reduction in runtime while improving relative test MAPE improvement by ~6.5%. The similar improvement is observed in Quarterly resolution with relative MAPE improvement of ~12.3% with ~3.1x reduction in runtime. These results suggest that generative AI can serve as a structured decision layer for model selection rather than as a forecasting model itself.

1. INTRODUCTION

Forecasting is a critical activity across multiple functional areas within organizations such as marketing, supply chain, finance etc. Forecasting in general is costly and complex activity due to a multiplicity of forecasting methods and possible combinations; the absence of an overall "*best*" combination methods; and the context-dependence of application method, based on available models, data characteristics, and the environment. Yet modern pipelines typically rely on brute-force evaluation across large model libraries, an approach that becomes computationally prohibitive at scale.

Researchers (Petropoulos *et al.* 2018, Önkal *et al.* 2017) have studied role played by forecaster's expertise to improve the forecast. Georgoff and Murdick (1986) and (Wang, 2009) have reported the key challenges faced by forecasters are not complex methods but selecting right method. In M4 competition, alone included 61 distinct forecasting methods (Makridakis, 2020) and also consistently combination-based method outperform single models in M4 and M5 competitions (Makridakis *et al.* 2020, Makridakis *et al.* 2022, Jaganathan and Prakash, 2020).

These competitions have strengthened the evidence for "*no free lunch*" in forecasting and pushed practitioners toward broader libraries and more sophisticated ensembles, often improving accuracy, but also increasing the evaluation burden because the number of plausible candidates grows combinatorially when ensembles and tuning are included. Even when organizations standardized a fixed forecasting "stack" of approved models, brute-force evaluation leads to high latency and if models are not evaluated fully could lead to sub-optimal selection, creating a tension between computational efficiency and predictive performance. We argue that model selection in forecasting is fundamentally a structured reasoning problem, *i.e.*, selecting a small subset of plausible candidates based on statistical evidence, historical analogues, and operational constraints. However, existing automated pipelines provide limited support for synthesizing these heterogeneous signals into an actionable shortlist.

Recent advances in large language models (LLMs) introduce a complementary opportunity where instead of using an LLM as the forecasting model, an LLM can serve as a recommendation and reasoning layer that consumes structured evidence such as series features, priors from historical analogues, constraints on shortlist length/ensemble size, and outputs a small set of promising candidates. This motivates the proposed framework, Forecasting with LLM-Assisted Recommendation Engine (FLARE). FLARE is designed with the objective to reduce search while preserving accuracy by combining (i) statistical characterization of time series, (ii) neighborhood evidence retrieved from a historical knowledge base (e.g., "what worked for series like this") using classical machine learning models, and (iii) LLM-guided reasoning to produce a ranked shortlist that is then evaluated by a AutoML forecasting engine.

FLARE constructs its knowledge base by extracting statistical features from historical time-series and systematically evaluating a broad forecasting library using AutoML to record empirically strong single models and ensemble configurations. The knowledge base captures both structural characteristics of series and evidence of what performed well for similar demand patterns. In addition, FLARE trains a Poisson regression model to estimate the appropriate shortlist size for a given series, enabling adaptive control over the number of models evaluated for training and ensemble.

For a new series, FLARE computes statistical descriptors, detects potential regime shifts, and retrieves neighborhood-level evidence of likely high-performing models using multi-label classification and Poisson regression trained on the knowledge base. These structured signals are then combined with model priors and task constraints (e.g., forecast horizon and latency preference) and provided as structured context to a large language model. Rather than generating forecasts directly, the LLM acts as a structured decision layer that reasons over the evidence and returns a ranked shortlist of candidate models (and, when appropriate, ensemble guidance). The forecasting engine (e.g., AutoGluon) then trains and evaluates only the shortlisted candidates.

This design transforms model selection from exhaustive search into guided localization, substantially reducing computational overhead while preserving out-of-time forecasting accuracy. The remainder of the paper is organized as follows: Section 2 presents related prior work; Section 3 describes FLARE Approach; and Section 4 reports experimental results on M4 univariate time-series; followed by conclusions and proposed next steps in Section 5.

2. RELATED PRIOR WORK

Time-series forecasting has a long history of methodological innovation, spanning classical statistical methods, rule-based heuristics, machine learning and deep learning-based approaches, and more recently retrieval augmented and foundational model paradigms. A consistent theme across the forecasting literature is time-series context dependence such as series characteristics, forecast horizon, regime shift, and other operational constraints that impact accuracy. This has motivated researchers to develop approaches for optimal model selection and ensembling rather than relying on a single universal forecasting method.

We organize the prior work into four themes: (i) statistical forecasting methods; (ii) machine learning based forecasting; and (iii) combination-based forecasting.

2.1 Statistical forecasting

Early forecasting research emphasized interpretable models that explicitly model trends, seasonality, and autocorrelation to generate forecasts. Historically, Naïve and seasonal naïve methods are widely used as strong benchmark models (Makridakis *et al.*, 1982). More structured stochastic modeling approaches were later formalized through the Box–Jenkins methodology for ARIMA-type models proposed by Box & Jenkins, (1970).

Among traditional statistical models, the Autoregressive Integrated Moving Average (ARIMA) model has been widely used for univariate time series forecasting and is built on the Box and Jenkins framework (Box & Jenkins, 1970). ARIMA also has many extensions such as SARIMA, a seasonal variant (Gustriansyah *et al.*, 2026). Even though SARIMA is effective, these models can be sensitive to specification choices and may require careful tunning, making them difficult to scale. To reduce the manual burden of specifying ARIMA/SARIMA orders, automatic ARIMA procedures (Commonly referred to as autoARIMA) were developed to algorithmically select non-seasonal parameter (*p, q, d*), and when seasonal structure is present, the seasonal orders (*P, Q, D*) (Hyndman and Khandakar, 2008).

Exponential smoothing methods provide another class of statistical forecasting methods, which focus on estimating the underlying patterns in the data using exponentially weighted averages of past observations and extending it naturally to trend and seasonal variants (e.g., Holt, Winters, and ETS/state-space formulations) (Brown, 1959, Holt, 1975, Winters, 1960). Subsequently, Theta method gained wide attention during the M3 competition and was later shown to be mathematically related to exponential smoothing with drift while remaining computationally efficient (Hyndman and Billah (2003). Bernardi *et al.* (2024) revisited the exponential smoothing developed by Brown (1959) and connected exponential recursion-based formulation with a stochastic interpretation.

2.2 Machine learning based forecasting

As forecasting applications expanded in scale and complexity, machine learning-based forecasting methods were increasingly adopted to capture non-linearities and leverage cross-series learning at scale. Recurrent neural networks (RNN) and their gated variants such as LSTMs were introduced to address long-range dependency learning via memory mechanisms (Hochreiter and Schmidhuber, 1997). In parallel, feature-based machine learning approaches (e.g., gradient-boosted decision trees) became prominent, with

XGBoost providing a scalable and high-performing boosting system that is often competitive when time-series are represented through engineered lag/seasonal/statistical features (Chen and Guestrin, 2016).

Deep learning models such as DeepAR enable learning jointly across multiple series using autoregressive recurrent network and producing probabilistic forecasts (Salinas *et al.* 2019). While global models reduce per-series manual modeling, they introduce their own costs (training complexity, architecture choices, and tuning), and their performance can still be heterogeneous across series regimes. A parallel line of work studies, how to select forecasting methods based on time-series characteristics also referred to as meta-learning (Talagala *et al.* 2023) where models are chosen using features describing trend, seasonality, intermittency, and other structural properties. Samanta *et al.* (2022) proposed model-less time-series forecasting (MLTF) based approach to identify the analogues used for forecasting.

Similarly, Retrieval-Augmented Time Series Forecasting (RAFT) retrieves similar historical patterns from training data and uses the retrieved futures to augment prediction, demonstrating that externalized memory can improve accuracy across benchmarks (Han *et al.* 2025). Ning *et al.* (2025) utilized large language model to develop TS-RAG approach that uses similarity-based retrieval to ground generated predictions in context relevant examples. AutoML system such as AutoGluon (Shchur *et al.*, 2023) extends model selection by automating training, tuning, and ensemble across model libraries, but typically rely on exhaustive evaluation under a time budget.

2.3 Combination based forecasting

Empirical evidence from forecasting research and competitions has repeatedly highlighted that (i) relative method ranking depends on the statistical property of the dataset and evaluation metric, and (ii) combinations/ ensembles are frequently showing up among the top performers across research and competitions (Makridakis *et al.* 2020, Makridakis *et al.* 2022, Jaganathan and Prakash, 2020). The M4 competition evaluated 61 methods across 100,000 series, reinforcing both the breadth of plausible candidates and the empirical competitiveness of combination-based approaches (Makridakis *et al.* 2020). Subsequent competitions, including M5, further emphasized that strong performance often arises from combinations (including Machine learning heavy entries), but these gains come with increased evaluation and operational complexity (Makridakis *et al.* 2022).

Because forecasting performance is context-dependent, a substantial body of work has focused on how experts and systems choose (or combine) models. Practitioner-oriented guidance has long recognized techniques for model selection and integration. Alvarado *et al.* (2017) and Petropoulos *et al.* (2018) extended this line of work by systematically including how forecaster expertise and credibility affect the integration of judgement and statistical forecast.

In this context, FLARE positions the foundation large language model as a reasoning layer to produce a constraint, ranked shortlisted candidate forecasting models for downstream evaluation. retrieval-augmented forecasting methods that retrieve trajectories to improve predictions, FLARE retrieves performance evidence to guide model selection, reducing brute-force search while maintaining out-of-sample forecating accuracy.

3. APPROACH

The current research focuses on solving model selection using generative AI for univariate models. Let $\boldsymbol{S} = \{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^n\}$ denote a collection of univariate series with ith time-series is represented as

$$\boldsymbol{y}^i = \{y_1^i, y_2^i, \ldots, y_{T_n}^i\} \quad \text{where} \qquad \ldots (1)$$

where, i=1, 2, … n and T_i represents length of i^{th} series. Also, H represents fixed forecast horizons (e.g. monthly, quarterly or yearly). For brevity we will be using H as forecasting horizon determined by the time-series granularity.

The proposed FLARE approach uses a model library $L = \{m_1, m_2 \ldots, m_M\}$ with M candidate models available to the forecasting engine. Each $m \in L$ induces a training and forecasting procedure that maps the training data $\boldsymbol{y}_{1:T_i}^{(i)}$ to an H step forecast:

$$\hat{\boldsymbol{y}}^{(\boldsymbol{i},\boldsymbol{m})} = \{\hat{y}_{T_i+1}^{(i,m)}, \hat{y}_{T_i+2}^{(i,m)}, \ldots, \hat{y}_{T_i+H}^{(i,m)}\} \qquad \ldots (2)$$

Optionally, the forecasting engine evaluates the weighted ensembles over a subset $\boldsymbol{A} \subseteq \boldsymbol{L}$ leading to ensemble forecast

$$\hat{\boldsymbol{y}}^{(\boldsymbol{i},ens(\boldsymbol{A},\boldsymbol{\theta}))} = \sum_{m \epsilon A} \theta_m \hat{\boldsymbol{y}}^{(i,m)} \quad \text{s.t. } \theta_m \geq 0 \qquad \ldots (3)$$

where, $\sum_{m \epsilon A} \theta_m = 1$ and θ_m denotes the model weight obtained via validation/stacking or equal weight in case of simple ensemble. Let's $\boldsymbol{\omega}$ be the set of all ensemble candidates considered by the forecasting engine. The full candidate set is represented as

$$\boldsymbol{L}^+ = \boldsymbol{L} \cup \boldsymbol{\omega} \qquad \ldots (4)$$

where, $\boldsymbol{L}^+$ is set of all candidate solution. The current research focuses on finding reduced search space $\boldsymbol{L}^* \subseteq \boldsymbol{L}^+$ to improve performance objective:

$$\min_{\boldsymbol{L}^* \subseteq \boldsymbol{L}^+} L_{error}(\boldsymbol{L}^*) + \gamma L_{latency}(\boldsymbol{L}^*) \qquad \ldots (5)$$

where L_{error} is the forecast error (MAPE is used as error loss for the experimentation), $L_{latency}$ is computation cost/latency of evaluating selected candidates $\boldsymbol{L}^*$, and $\gamma \geq 0$ controls the trade-off between predictive accuracy and computational efficiency. FLARE approximates Eq. (5) via guided localization that restricts the candidate set prior to training while preserving the out-of-time forecasting accuracy.

The proposed FLARE approach combines classical approaches learning from prior knowledge and dynamically adjusts the recommendation using LLM decision layer as shown in Figure 1.

The proposed approach first leverages statistical properties of each time-series to identify the optimal number of models to consider. Let's $\boldsymbol{X}^{(i)} = \{x_1, x_2 \ldots, x_\varnothing\}$ denotes the statistical feature vector for ith series, and let $\boldsymbol{B}^{(i)} \subseteq \boldsymbol{L}^*$ denotes empirical "best" model for ith series observed from AutoML engine (AutoGluon), i.e., the selected single model or the member of selected ensemble.

We define the target shortlist size as $k_i = |\boldsymbol{B}^{(i)}| \in \{1,2,3\ldots\}$ and k_i using a Poisson regression with mean $\lambda_i(\boldsymbol{X}^{(i)})$

$$k_i = Poisson(\lambda_i), \quad \lambda_i = exp(\boldsymbol{w}^T \varphi(\boldsymbol{X}^{(i)}) \qquad \ldots (6)$$

where $\boldsymbol{\varphi}(.)$ is preprocessing pipeline that includes imputation, variance thresholding, and robust scaling. The parameter $\boldsymbol{w}$ is learned by minimizing the regularized negative log-likelihood (Friedman *et al.* 2010). The Poisson model is used as a parsimonious count model with objective to accurately shortlisting sizing rather than fully specifying the data generating process. Thus, Eq. (6) predicts the expected cardinality of the AutoML-selected solution $\boldsymbol{B}^{(i)}$.

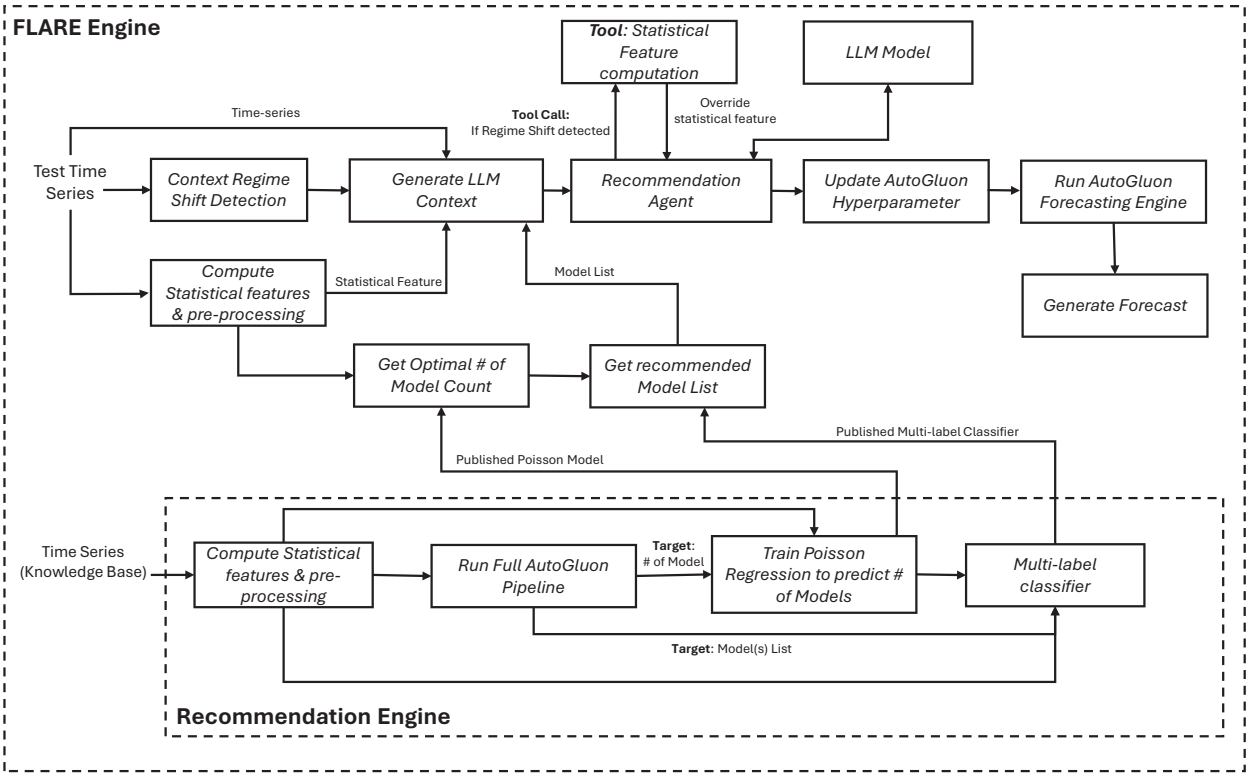


Figure 1. *Adaptive forecasting engine using FLARE Engine*

To identify the eligible models, a one-vs-rest logistic model is trained for each $c \in \boldsymbol{L}$ using labels

$$\boldsymbol{z}_{i,c} = \mathbb{I}\{c \in \boldsymbol{B}^{(i)}\} \quad \ldots (7)$$

and model probabilities

$$Pr\left(\boldsymbol{z}_{i,c} = 1 \middle| \boldsymbol{X}^{(i)}\right) = \sigma(\boldsymbol{v}_c^T \varphi(\boldsymbol{X}^{(i)}) \quad \ldots (8)$$

where $\sigma(.)$ is sigmoid function and $\boldsymbol{v}_c$ is learned weights. Using the k_i and $Pr(\boldsymbol{z}_{i,c}=1|\boldsymbol{X}^{(i)})$ top models prior is shortlisted as

$$\boldsymbol{L}^{*(prior)} = TopK(\{Pr\left(\boldsymbol{z}_{i,c} = 1 \middle| \boldsymbol{X}^{(i)}\right)\}_{c \in \boldsymbol{L}}, \hat{k}_i(\boldsymbol{X}^{(i)}) \quad \ldots (9)$$

One of the key challenges in time-series is non-stationarity and regime shift, where the global time-series and statistical features may be less representative of future behavior. To address this, we define a regime shift detector R(.) as

$$r_i = R\left(\boldsymbol{y}_{1:T_i}^{(i)}\right) \in \{0, 1\} \times \mathbb{R}^q \quad \ldots (10)$$

where r_i contain an binary indicator (shift / no-shift) and optional diagnostics (e.g., change magnitude, recency horizon, and confidence). When regime shift is detected, statistical features are re-computed via a tool-call by LLM over a recent window to provide updated evidence for recommendation.

Let F denotes the LLM engine that takes input as statistical features $\boldsymbol{X}^{(i)}$, the time-series context $\boldsymbol{y}_{1:T_i}^{(i)}$, the regime shift output r_i, and the prior shortlist $\boldsymbol{L}^{*(prior)}$. The LLM produce the final shortlist

$$\boldsymbol{L}^* = F(Context(\boldsymbol{X}^{(i)}, \boldsymbol{y}_{1:T_i}^{(i)}, \boldsymbol{L}^{*(\boldsymbol{prior})}, r_i) \quad \ldots (11)$$

In the current implementation, the LLM is prompted to return a structured JSON object containing ranked model name (and optional ensemble guidance). To ensure reproducibility, the temperature is set to 0.0 and use a fixed prompt template. The selected $\boldsymbol{L}^*$ model is used to update the hyper-parameter passed to the AutoGluon pipeline, which trains and evaluates only shortlisted candidates to generate the final forecast $\hat{\boldsymbol{y}}_{T_i:T_i+H}^{(i)}$. The proposed approach is validated using M4 competition data as shown in Section 4.

4. EXPERIMENTATION AND RESULTS

4.1 Dataset

The proposed FLARE approach is benchmarked using M4 competition dataset. Experiments are conducted using monthly subsets to evaluate the robustness against temporal granularity. To support the FLARE's retrieval-based evidence mechanism, two disjoint samples from each frequency subset:

- **Test set (evaluation set):** 300 randomly selected time series from M4 dataset at monthly and quarterly resolutions, used for benchmarking and reporting results with forecasting horizon of 18 and 8 for monthly and quarterly series, respectively.
- **Knowledge base (reference set):** 200 randomly selected time series with monthly and quarterly resolutions, used to train Poisson regression model for predicting the required shortlist size and to learn model priors via the classification model described in Eq. (9).

The selected test series are intended to span a diverse range of demand patterns represented in M4. Forecasting performance is assessed using Mean Absolute Percentage Error (MAPE). Computational efficiency is assessed using modelling latency (seconds).

4.2 Baseline Methods

The proposed FLARE approach is compared against a set of baselines designed to isolate the contribution of (i) LLM-based recommendations, (ii) structural statistical evidence-based recommendation, and (iii) Random baselines

- **AutoML baseline (AutoGluon-TimeSeries):** As the AutoML benchmark, the paper uses AutoGluon-TimeSeries (Shchur *et al.*, 2023) which automatically trains and optionally ensemble multiple probabilistic and ML models. This baseline represents exhaustive evaluation under a fixed configuration and provides a reference for both predictive performance and run-time.
- **Random-K Models baseline:** As a lower bound inference, we randomly sample K models from the available model list and evaluate them using the same AutoGluon training and forecasting pipeline. As mean number of models selected in AutoML baseline is 2.43 thus K=3 is chosen for the baseline comparison. This baseline quantifies the expected predictive performance and latency when model selection is uninformed.
- **TS-based LLM Model Recommendation:** The univariate time series is serialized (e.g., as JSON) and provided directly to the LLM, which outputs a shortlist of forecasting model families to evaluate. The shortlisted models are then passed to the AutoGluon pipeline for training and forecasting on the target series.
- **Stat-based LLM Model Recommendation:** This baseline augments the LLM input with a compact set of precomputed statistical descriptors of the target series (e.g. entropy, trend, seasonality, sparsity etc). The LLM produces a ranked shortlist, and the shortlisted models are evaluated using the same AutoGluon training and forecasting procedure. This baseline isolates the incremental value of adding structured statistical evidence to the LLM prompt without retrieval-based neighborhood priors.

The above model is compared with the FLARE outcome and result is presented in next sub-section.

4.3 Results

Experiments are performed using the constructed dataset. For comparability across runs, AutoGluon's hyperparameter configuration is fixed to the default settings. In addition, ensembling is enabled for all runs of the selected model(s) if $k > 1$. Unless stated otherwise, all LLM-based runs use OpenAI's GPT-5.1 as the decision layer with temperature set to 0.0. Figure 2 summarizes test horizon performance across runs, reporting average accuracy and latency.

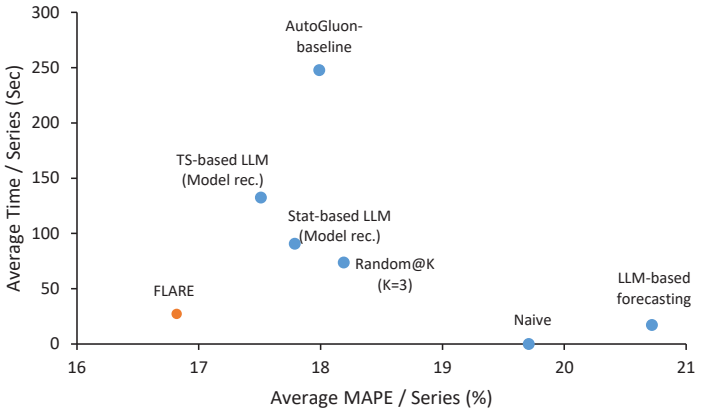


Figure 2. *Accuracy–Latency Trade-off on M4 Monthly Forecast (Test Horizon = 18 Months)*

Figure 2, also include LLM-based forecasting baseline in which a foundation model is directly used to generate the forecasts, as well as a Naïve baseline as standard baselines. Figure 2, highlights accuracy-latency tradeoffs across the evaluated methods for the monthly benchmark. AutoGluon achieves competitive accuracy but incurs the highest runtime due to exhaustive evaluation. In contrast, direct LLMbased forecasting (*i.e.*, generating forecasts directly from historical points) performs poorly and is less accurate than the Naïve baseline. LLM-based model recommendation baselines that use either raw time-series inputs or statistical features reduce latency relative to AutoGluon. FLARE lies on the Pareto frontier, achieving the lowest average test MAPE while reducing end-to-end runtime by 9.1x as compared with the full AutoML baseline. These results indicate that retrieval-augmented priors along with constrained LLM reasoning can localize a small set of high-performing candidates, yielding both faster evaluation and improved out-of-sample performance. Further, Table 1 summarizes the detailed results for monthly forecast with train and test horizon.

Mode	Average MAPE ± SE (%)		Average # of model	Average AutoML Latency/ Series (Sec)	P90 AutoML Latency/series (Sec)	Average LLM Latency/Series (Sec)
	Train	Test				
AutoGluon - Baseline	**9.63 ± 0.85**	17.99 ± 2.22	2.43	247.69	804.49	0.0
Random@K (*k*=3)	10.97 ± 0.97	18.19 ± 2.42	1.63	73.78	245.63	0.0
TS-based LLM (Model Rec.)	11.01 ± 0.95	17.51 ± 1.96	1.70	116.44	128.86	16.06
Stat-based LLM (Model Rec.)	10.56 ± 0.96	17.79 ± 2.19	1.83	75.73	222.49	14.96
FLARE	11.23 ± 0.99	**16.82 ± 2.02**	1.60	5.85	**4.51**	21.30
LLM-based forecasting	-	20.72 ± 2.36	-	-	-	17.21
Naïve	-	19.71 ± 2.41	1.0	**0.01**	-	-

Table 1. *Monthly benchmark for train and test horizon*

Based on results from Table 1, AutoGluon baseline evaluates the full candidate model set and selects the configuration with the lowest training error. However, this exhaustive search is computationally expensive, requiring 247.69 seconds per series on average, while selecting 2.43 models per series for ensembling. In contrast, FLARE reduces average overall latency to 27.15 seconds per series (approximately 9.1× faster); most of this time is attributable to the LLM supervisory layer for model recommendation. FLARE also improves test MAPE from 17.99% to 16.82% (an absolute improvement of 1.17%) and reduces the standard error from 2.22% to 2.02%.

One of the main drivers of the computation differences is the type of model selected. Figure 3, shows that FLARE evaluates substantially fewer models than the AutoGluon baseline while selecting a compact subset for ensembling, highlighting how candidate localization reduces computation.

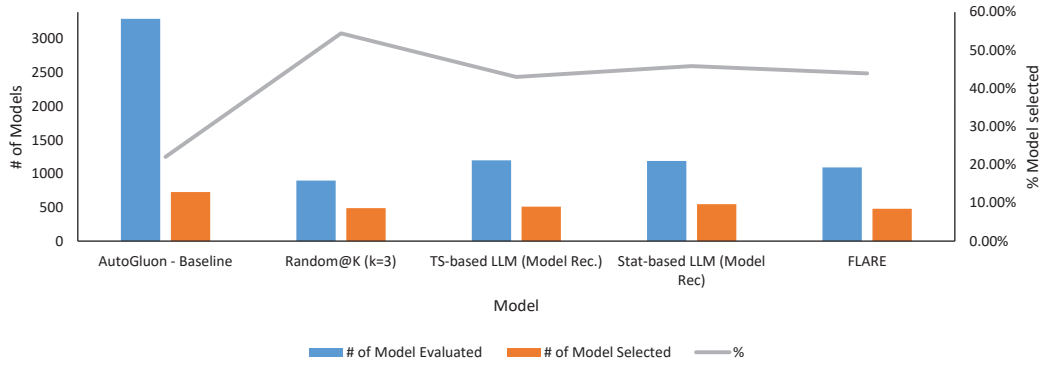


Figure 3. *Model Evaluation and Selection Volume (Monthly Test Set, H = 18)*

Table 2 further breaks down the selected models by category across the 300 monthly test series, and the values in parathesis indicate how many times models from that category is evaluated as candidate model within AutoGluon run.

Type of Models	Models	AutoGluon (Baseline)	Random@K (*k=3*)	TS-based LLM (Model Rec.)	Stat-based LLM (Model Rec.)	FLARE
Simple Baseline	Naive, NaïveSeasonal Seasonal Average	212 (900)	22 (38)	103 (302)	26 (58)	80 (245)
Statistical Model	Theta, AutoETS, AutoArima	262 (900)	355 (680)	409 (899)	389 (890)	395 (839)
Specialized Model	ADIDA, Croston	103 (600)	0 (0)	0 (0)	0 (0)	1 (1)
ML Models	XGB, RF, DeepAR	152 (900)	113 (182)	0	136 (242)	5 (9)
Total # of Models		**729**	**490**	**512**	**551**	**481**

Table 2. *Final model selection distribution for monthly test dataset*

The AutoGluon baseline frequently selects both statistical and ML-based models, including computationally expensive learners such as XGBoost and DeepAR. In contrast, LLM models are de-prioritizing specialized intermittent-demand models whereas FLARE concentrates almost exclusively on lightweight statistical models (e.g., AutoETS, Theta) and rarely select ML or specialized intermittent-demand models. This targeted selection strategy explains the substantial latency reduction observed in Table 1 while maintaining, and in some cases improving, forecasting accuracy. Overall, the monthly benchmark suggests that statistical models dominate predictive performance, and the learned selection mechanism can avoid unnecessary computational overhead.

An illustrative example of the FLARE input–output flow for a monthly series is shown in Figure 4. The figure demonstrates how different input layers are used by the LLM supervisory layer. The model has recommended Random Forest and DeepAR based on the training. However, due to high regime shift observed the LLM included the ARIMA model and excluded the DeepAR due to low probability. The LLM has evaluated the exclusion for other models and finally selected Random Forest, AutoArima and Seasonal Naïve as recommendations for AutoML execution with AutoARIMA as final model with test MAPE of 5.62%. Whereas, AutoML baseline evaluated all combinations and selected AutoETS, SeasonalAverage, Naive, and Croston as its final model with MAPE of 5.74%.

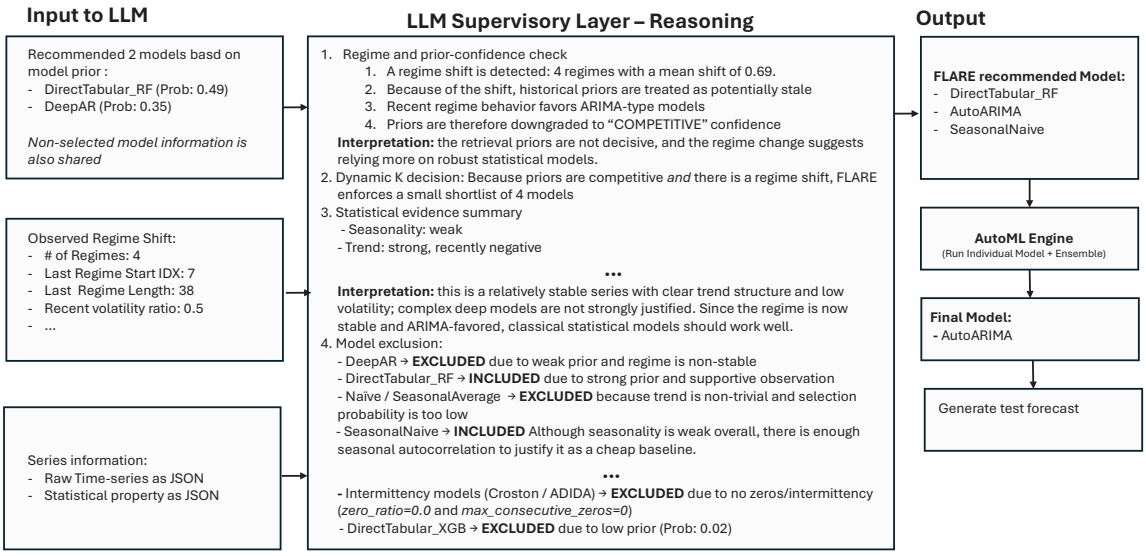


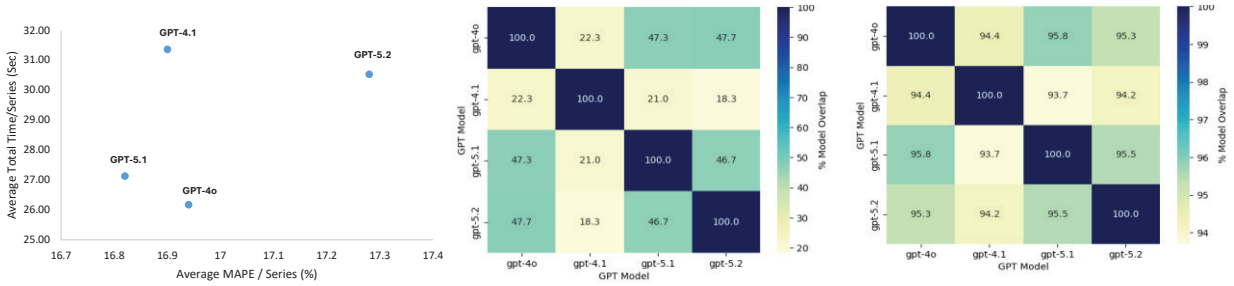
Figure 4. *Illustration of Input output based on a monthly series*

The current FLARE-based approach integrates statistical-feature-based recommendations, time-series-based recommendations, regime-shift context detection, and learned model priors into a unified decision framework. To better understand the contribution of each component, we conduct an ablation study by systematically removing one module at a time as reported in Table 3.

FLARE Ablation	Average MAPE ± SE (%)	Delta with FLARE (%) *(Lower is better)*	Average # of model	Mean Latency/ Series (Sec)	P90 Latency/series (Sec)
FLARE	16.90 ± 2.02	0.00	1.60	5.85	4.51
Remove Stat-based features	17.11 ± 2.07	0.21	1.61	7.11	5.82
Remove time-series features	17.31 ± 2.07	0.28	1.53	4.75	3.71
Remove Model Prior	17.01 ± 2.02	0.11	1.66	2.33	2.87
Remove regimen correction	17.06 ± 2.02	0.16	1.61	9.37	3.91

Table 3. *Ablation study for FLARE for Monthly benchmark*

Based on the ablation study, passing time-series has the highest impact on performance. However, only passing time-series on LLM lead to a MAPE performance on 17.51% due to multiple wrong selection with average time for training is 116.44 seconds as shown in Table 1. Further ablation on LLM model is performed to evaluate the sensitivity of FLARE to the choice of LLM decision layer (GPT-4o, GPT-4.1, GPT-5.1, and GPT-5.2) with result summarized in Figure 5.



(a) LLM selection performance impact *(b) % model set overlap (Full alignment)* *(c) % model overlap (Partial alignment)*

Figure 5. *LLM model ablation study on FLARE*

Figure 5 suggests that the choice of LLM affects latency more than accuracy, as shown in Figure 5(a). Although the recommended model sets can appear different across LLMs as seen in Figure 5(b), for example only ~21% of the complete model set recommended by GPT-4.1 matches exactly with the set recommended by GPT-5.1. This difference is less pronounced when measured at the overlap level. For instance, the recommended sets {“AutoETS”, “SeasonalAverage”} and {“AutoETS”, “AutoARIMA”} has a 50% overlap. As shown in Figure 5(c), model overlap exceed ~93% across all LLMs, indicating that the decision layer produces broadly consistent model selection.

Proposed FLARE Approach is also validated in 300 quarterly series and result is summarized in Table 4. Based on Table 4, similar patterns are observed as in Table 1. The AutoML-based baseline achieves low training error but generalizes less to the test horizon, suggesting overfitting when exhaustive search is not constrained. In contrast, LLM-based recommendations method improve out-of-sample forecasting accuracy from 16.02% to 14.08% (1.98% improvement) with substantially reducing runtime by ~3.1x relative to the full AutoML baseline.

Mode	Average MAPE ± SE (%)		Average # of model	Average AutoML Latency/ Series (Sec)	P90 AutoML Latency/series (Sec)	Average LLM Latency/Series (Sec)
	Train	Test				
AutoGluon - Baseline	**6.99 ± 0.54**	16.02 ± 3.08	2.26	251.31	533.72	0.0
Random@K (*k*=3)	9.30 ± 0.99	16.34 ± 3.18	1.41	135.08	210.35	0.0
TS-based LLM (Model Rec.)	8.35 ± 0.64	16.31 ± 3.32	1.8	114.73	174.05	14.06
Stat-based LLM (Model Rec.)	8.31 ± 0.61	15.50 ± 3.02	1.53	164.76	278.70	12.96
FLARE	9.48 ± 1.03	**14.08 ± 2.59**	1.72	**59.56**	**70.75**	19.30

Table 4. *Quarterly benchmark for train and test horizon*

5. CONCLUSION

This study presents Forecasting with LLM-Assisted Recommendation Engine (FLARE), a model-based and LLM-guided model selection framework for univariate time-series forecasting that jointly optimizes predictive accuracy and computational efficiency. By integrating statistical feature–based priors, time-series contextual reasoning, regime-shift detection, and a learned dynamic model-count predictor, FLARE effectively reduces the candidate search space before invoking the forecasting engine.

The proposed FLARE approach demonstrates consistent improvements in test performance across both monthly and quarterly resolutions, while achieving substantial reductions in computational effort. Although the training error in FLARE is higher compared with the exhaustive AutoML baseline, the proposed solution has shown improvement in the out-of-time MAPE, indicating better generalization. The experimental results suggest that AutoML-based approaches may overfit the training dataset due to exhaustive model search, whereas incorporating validation layer through large language models enables more targeted model selection. This leads to significantly lower computational costs and improved forecast stability in out-of-time evaluation. The proposed approach therefore offers a scalable forecasting solution without compromising predictive quality.

The ablation study confirms that each architectural component contributes to overall performance, with time-series contextual signals and learned model priors playing a particularly important role. Overall, the findings suggest that combining structured statistical evidence, learned selection priors, and generative AI reasoning provides an effective and scalable alternative to exhaustive AutoML-based model selection for univariate forecasting tasks. Future work includes extending the framework to multivariate forecasting settings and leveraging foundation models to optimize hyperparameters and ensemble weights more effectively.

REFERENCES

Alvarado-Valencia, J. A., Barrero, L. H., Önkal, D., & Dennerlein, J. T. (2017). "Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting." *International Journal of Forecasting*, 33(1), 298–313.

Bernardi, E., Lanconelli, A., & Lauria, C. S. A. (2024). "A novel theoretical framework for exponential smoothing." *arXiv*. https://arxiv.org/abs/2403.04345

Box, G. E. P., & Jenkins, G. M. (1970). "Time series analysis: Forecasting and control." San Francisco, CA: Holden-Day.

Brown, R. G. (1959). "Statistical forecasting for inventory control." McGraw-Hill.

Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1), 1–22.

Georgoff, D. M., & Murdick, R. G. (1986). "Manager's Guide to Forecasting." *Harvard Business Review*, 64(1), 110–120.

Gustriansyah, R., Dewi, D. A., Puspasari, S., & Sanmorino, A. (2026). "A SARIMA approach with parameter optimization for enhancing forecast accuracy for native chicken egg production." *Jurnal Ilmu Matematika dan Aplikasinya, 20*(2), 1331–1344.

Han, S., Lee, H., Cha, S., & Pfister, T. (2025). "Retrieval augmented time series forecasting." *In Proceedings of the 42nd International Conference on Machine Learning.*

Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory." Neural Computation, 9(8), 1735–1780.

Holt, C. C. (1957). "Forecasting seasonals and trends by exponentially weighted averages." (O. N. R. Memorandum No. 52). Carnegie Institute of Technology.

Hyndman, R. J., & Billah, B. (2003). Unmasking the Theta method. *International Journal of Forecasting, 19*(2), 287–290.

Hyndman, R. J., & Khandakar, Y. (2008). "Automatic Time Series Forecasting: The forecast Package for R." *Journal of Statistical Software*, 27(3).

Jaganathan, S., & Prakash, P. K. S. (2020). "A combination-based forecasting method for the M4-competition." *International Journal of Forecasting*, 36(1), 98–104.

Box, G. E. P., & Jenkins, G. M. (1970). "Time series analysis: Forecasting and control." San Francisco, CA: Holden-Day.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). "The M4 Competition: 100,000 time series and 61 forecasting methods." International Journal of Forecasting, 36(1), 54–74.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364.

Ning, J., Pan, H., Zhang, Y., Wang, S., & Wang, J. (2025). "TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecasters." *Advances in Neural Information Processing Systems* (NeurIPS 2025). arXiv:2503.07649.

Önkal, D., Gönül, M. S., Gürbüz, N., & Goodwin, P. (2017). "Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting." *International Journal of Forecasting*, 33(3), 746–760.

Petropoulos, F., Fildes, R., & Goodwin, P. (2018). "Judgmental selection of forecasting models." *International Journal of Forecasting*, 34(4), 701–719.

Samanta, S., Prakash, P. K. S., & Chilukuri, S. (2022). "MLTF: Model less time-series forecasting." *Information Sciences, 593*, 364–384.

Shchur, O., Turkmen, C., Erickson, N., Shen, H., Shirkov, A., Hu, T., & Wang, Y. (2023). "AutoGluon-TimeSeries: AutoML for Probabilistic Time Series Forecasting." (preprint arXiv:2308.05566).

Talagala, T. S., Hyndman, R. J., & Athanasopoulos, G. (2023). "Meta-learning how to forecast time series." *Journal of Forecasting*, 42(6), 1476–1501.

Wang, X., Smith-Miles, K., & Hyndman, R. J. (2009). "Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series." *European Journal of Operational Research*, 201(1), 245–255.

Winters, P. R. (1960). "Forecasting sales by exponentially weighted moving averages." *Management Science*, 6(3), 324–342.

Applying Management Science to Forecasting in Rare Diseases: An In-Depth Framework

Rishabh Bawa, Manager, Viscadia; Navendu Gupta, Consultant, Viscadia; Pratham Gupta, Consultant, Viscadia; Varun Singh, Associate Consultant, Viscadia

Objective: The objective of this article is to demonstrate how management-science methodologies can be applied to improve demand forecasting in rare diseases, using Hemophilia as an illustrative example. Unlike high-incidence therapeutic areas such as oncology, rare-disease markets exhibit extreme patient level variability, limited data visibility, and channel driven distortions that render traditional forecasting techniques insufficient.

Approach: A structured, multi-layer forecasting framework was developed that integrates patient level, customer level, and process level analytics. Individual hemophilia patients were tracked across specialty pharmacies and distributor channels, enabling longitudinal mapping of treatment journeys, ordering cadence, and irregular consumption patterns. Elements assessed included concentration curves, irregular ordering, pharmacy switching, new-patient onboarding behaviors, and customer-specific purchasing and inventory decisions. Additional evaluation covered patient-fulfillment challenges such as assay differences, reimbursement delays, and access bottlenecks. Other observed dynamics such as dependence on limited patient relationships and absence of key treatment-center input were identified and addressed through iterative modeling techniques that combine deterministic patient-level data with stochastic variability parameters.

Key Insights: Analysis demonstrated that rare-disease forecasting requires a fundamentally different approach from models used in broader disease areas. In hemophilia, 60–80% of total factor volume often originates from a small cohort of high-utilizing patients, magnifying the impact of individual behavioral changes on aggregate demand. Irregular ordering contributed ~10% upside variability, while pharmacy switching introduced noise into channel-level attribution. New-patient volumes proved modest and highly unpredictable, with early therapy intensity and adherence varying significantly. Customer and distributor purchasing patterns, including atypical stockpiling cycles and heterogeneity in months-on-hand inventory further distorted observable demand signals. These findings underscore that reliable rare-disease forecasting necessitates micro-level behavioral modeling, dynamic scenario planning, and continuous synthesis of dispersed data sources rather than reliance on macro-level epidemiological or trend-based methods.

Relevance: This work highlights the importance of management science in enhancing decision-making precision in rare-disease markets, where small populations and volatile utilization patterns amplify operational and financial risks. The patient-centered forecasting methodology presented here supports more accurate supply planning, risk identification, and commercial strategy development. As biologics and gene therapies increasingly target ultra-rare disorders, the industry faces a growing need for forecasting frameworks capable of incorporating individualized treatment dynamics, fragmented data ecosystems, and channel complexity. The hemophilia case exemplifies how such an approach can materially improve forecast reliability and serve as a blueprint for rare disease analytics more broadly.

Key Words: Commercial Analytics, Forecasting, Patient-Level Forecasting, Distributor Dynamics

1. INTRODUCTION

1.1 The Rare Disease Forecasting Challenge

Pharmaceutical forecasting has traditionally relied on epidemiological models and historical trend analysis suitable for high-prevalence diseases with predictable patient populations and stable treatment patterns. These conventional approaches, whether demand-based (utilizing historical sales data and promotional response modeling) or patient-based (building bottom-up models from epidemiological prevalence), assume certain market conditions that rare diseases systematically violate.[1]

With over 7,000 identified rare disease types affecting approximately 300 million people globally, and only a fraction having approved treatments, forecasting methodologies must evolve to address fundamentally different market characteristics.[2] Traditional approaches fail in rare disease contexts due to several critical factors: small patient populations magnify the impact of individual-level variability on aggregate demand signals; fragmented data ecosystems across specialty pharmacies, treatment centers, and distributors create visibility gaps that obscure true consumption patterns; and evolving treatment landscapes including gene therapies and precision medicine introduce unprecedented uncertainty in patient identification, uptake trajectories, and treatment duration.[3]

Recent industry analysis has documented systematic forecasting failures in rare disease contexts, with small patient populations creating extreme sampling variability where standard confidence intervals become meaningless.[4] A cohort of 100-200 patients may represent the entire addressable market, making individual patient dynamics statistically significant at the aggregate level-a scenario that inverts traditional statistical logic where individual observations dissolve into population means.

1.2 Objectives and Approach

This article does not approach these challenges as theoretical abstractions. It synthesizes direct advisory experience across multiple rare disease forecasting engagements — spanning marketed therapies and pre-launch pipeline assets in rare bleeding disorders and severe pediatric epilepsy — to accomplish three objectives:

1. Characterize the structural failures of conventional forecasting approaches when applied to rare disease commercial contexts through systematic gap analysis
2. Present a multi-layer analytical framework integrating patient-level, customer-level, and process-level analytics validated through real-world implementation
3. Establish transferable methodological principles for commercial organizations building rare disease forecasting capabilities

2. GAP ANALYSIS IN RARE DISEASE FORECASTING: FIVE CRITICAL CHALLENGES

Cross-engagement analysis reveals a consistent set of forecasting challenges that are not specific to therapeutic area or commercial context but endemic to rare disease markets broadly. These challenges represent fundamental mismatches between conventional forecasting assumptions and rare disease market realities.

2.1 Challenge 1: Extreme Demand Concentration Without Patient-Level Risk Management

In conventional pharmaceutical forecasting, individual patient heterogeneity averages out across large populations, enabling mean-based projections. Rare disease markets exhibit the opposite dynamic: extreme demand concentration where a small cohort of high-utilizing patients generates the majority of total volume, amplifying the impact of individual behavioral changes on aggregate demand.

Empirical Evidence: Analysis of one engagement's patient population revealed a striking concentration profile. Among 237 patients who had ever initiated therapy, demand distribution was profoundly skewed: 15 patients generated 60% of total annual volume, and 28 patients generated 80%. The single highest-utilizing patient alone accounted for 13% of total annual demand; the top 5 patients combined represented 32%. A preliminary segmentation matrix organized patients across two behavioral axes — order frequency and order volume — yielding five distinct cohorts with annual utilization ranging from 8,000 international units (IU) at the low end to 780,000 IU at the high end: an approximately 100-fold difference (Figure 1).

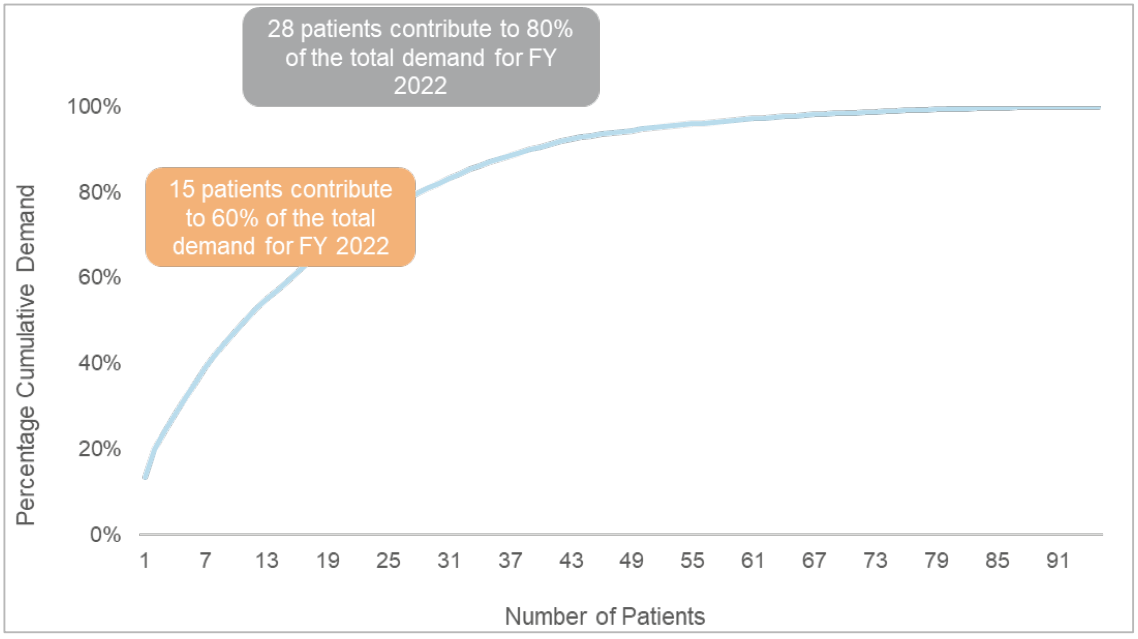


Figure 1. *Cumulative Demand Concentration Curve (N=237 patients). Fifteen patients contribute 60% and 28 patients contribute 80% of total annual demand, demonstrating extreme concentration relative to the active patient base. Lorenz-type curve constructed from individual patient-level utilization data across FY2022.*

This concentration was stable across the three-year analysis window, establishing it as a structural feature of the business rather than a transient anomaly. The baseline forecasting process treated all patients uniformly, applying the same rolling-average projection logic to a patient generating 780,000 IU annually as to one generating 8,000 IU — a premise the data does not support. The commercial consequence was already realized at the time of the engagement: the loss of a single high-utilization patient worth approximately $3.5 million at the beginning of the subsequent fiscal year, a material event relative to the $41 million annual forecast.

In another engagement context — a severe pediatric epilepsy therapy — the equivalent concentration manifested differently but with similar structural consequences. Patients could be segmented based on seizure control status (well-controlled versus not-well-controlled on current therapy), which determined both their likelihood of receiving the new therapy and their probable treatment intensity. Collapsing these segments into a single addressable pool systematically overstated near-term revenue projections by failing to account for heterogeneous uptake probabilities and dosing requirements across control-status segments.

Root Cause: Conventional forecasting methodologies assume statistical regularities that require large populations. In rare disease contexts with steep concentration curves, no statistical technique recovers from the loss of a patient representing 8–13% of total revenue without explicit per-patient risk monitoring and scenario planning as standing operational practice.

2.2 Challenge 2 - Systematic New-Patient Acquisition Overestimation

Rare disease forecasting models frequently anchor new patient assumptions on epidemiological prevalence estimates rather than empirically observed acquisition rates, producing directional and persistent overestimation.

Empirical Evidence: Across all observation periods in one engagement, the baseline forecasting model overestimated new patient starts by 43–69%. The observed run rate was 10–17 new patients annually against model assumptions of 24–32. Correspondingly, new patient volume overestimation ranged from 18–37%. This was not a commercial execution failure — it was a forecasting methodology failure attributable to anchoring assumptions on theoretical prevalence calculations without calibration to real-world acquisition dynamics.

In the pre-launch pipeline engagement, the equivalent challenge manifested as the epidemiology-to-revenue translation process. Disease prevalence estimation required sequential filtering through diagnosis rate assumptions, pediatric population growth, mortality rate adjustments, and delay-in-diagnosis parameters — the average lag between symptom presentation and confirmed genetic diagnosis. Each filtering step introduced uncertainty, and compounding optimistic assumptions at multiple stages could produce an addressable patient pool estimate two to three times the realistic figure. The forecasting architecture was designed to make each epidemiological assumption explicit and independently reviewable, with sensitivity toggles for diagnosis rate trajectory, mortality rate, and diagnosis delay — transforming a single opaque prevalence estimate into a transparent, auditable assumption stack.

Root Cause: Epidemiological prevalence estimates reflect theoretical addressable populations, not realized acquisition rates. Actual patient identification is constrained by diagnostic capacity, physician awareness, treatment center coverage, and payer access barriers — factors that epidemiological models systematically underweight. Without empirical calibration to observed acquisition rates, new patient forecasts become directionally biased upward.

2.3 Challenge 3: Channel Inventory Opacity Masking Stable Patient Demand Signals

In rare disease contexts with specialty pharmacy distribution networks, the disconnect between patient pull-through demand and manufacturer ex-factory demand can be the largest single driver of short-term forecast inaccuracy.

Empirical Evidence: While patient pull-through demand in one engagement ran at a relatively stable 1.5–2.9 million IU per month across the observation period, ex-factory demand exhibited extreme volatility — ranging from near-zero in several months to a single monthly peak exceeding 7.9 million IU, representing approximately three months of contemporaneous patient consumption compressed into one transaction (Figure 2).

Three discrete channel events drove this divergence:

1. A supply-side shortage on the primary assay size that bled aggregate channel inventory from approximately $6 million to $2 million in product value, with patient pull-through unaffected but ex-factory demand falling to near-zero during inventory depletion then surging in restocking waves.

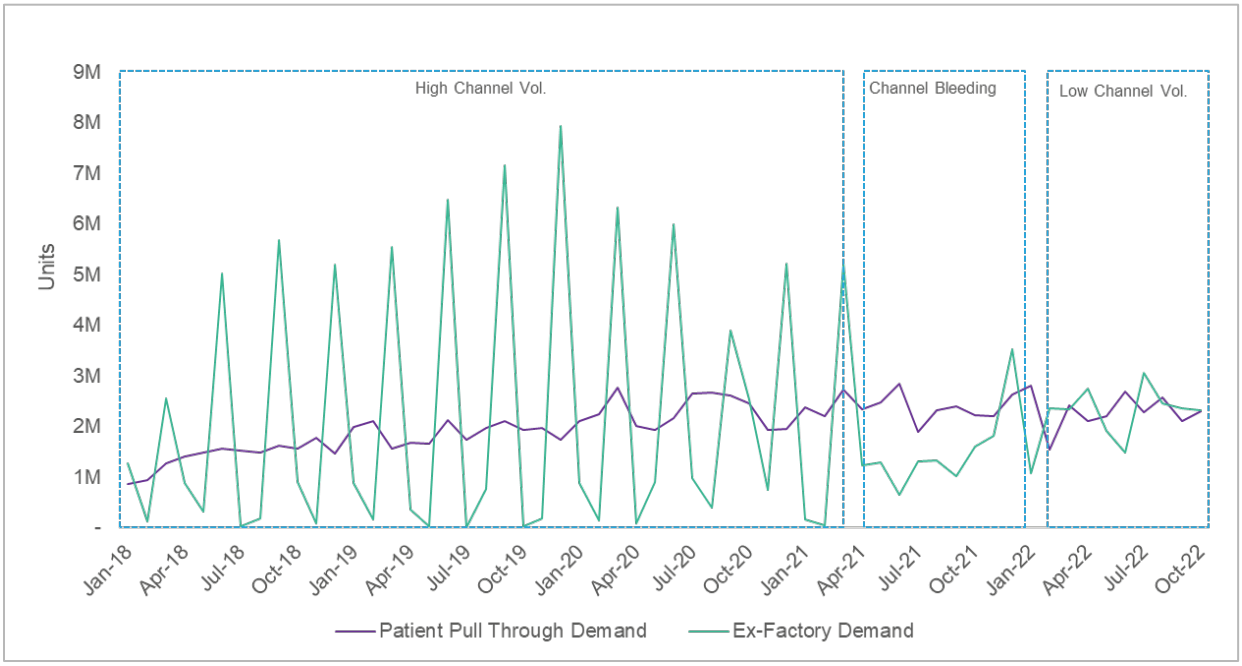


Figure 2. *Patient Pull-Through Demand vs. Ex-Factory Demand by Month (IU, FY2018–FY2022). Patient pull-through (purple line) remained directionally stable while ex-factory shipments (green line) oscillated between near-zero and peaks exceeding 7.9 million IU monthly. Volatility driven by supply shortages, shifted stocking cycles, and deal-driven volume pushes.*

2. A major specialty pharmacy building holiday inventory in September/October rather than the usual November/December window, generating a mid-year ex-factory spike with no basis in patient consumption timing.
3. Deal-driven volume pushes agreed between the commercial team and distributors, introducing artificial ex-factory demand events at commercially motivated timing rather than in response to patient demand signals.

None of these events were anticipated in the forecast because no structured information exchange existed between the manufacturer and its specialty pharmacy partners. The baseline forecasting process used working months-on-hand (MoH) assumptions as management levers adjusted post hoc to reconcile bottom-up projections with financial targets, rather than as empirically calibrated, customer-specific parameters.

In the post-launch epilepsy therapy engagement, the equivalent channel challenge manifested as the multi-step Risk Evaluation and Mitigation Strategy (REMS) access pipeline. The lag between a REMS referral and first commercial shipment spanned multiple weeks, and conversion rates at each step — from referral to cardiac echocardiogram (ECHO) completion, ECHO to benefits verification, benefits verification to first shipment — were influenced by payer behavior, prescriber compliance with REMS requirements, and pharmacy enrollment status. Each step represented an access-layer channel dynamic between the prescribing decision and the revenue event.

Root Cause: Patients pull-through and ex-factory sales are distinct quantities governed by distinct mechanisms. Treating them as a single demand signal produces forecasts that conflate channel inventory dynamics with patient consumption, generating structurally distorted projections regardless of patient-level tracking accuracy.

2.4 Challenge 4: Absence of Competitive Market Context and Class-Level Trends

Bottom-up patient rollups developed without class-level competitive context generate projections that may appear locally consistent but are directionally wrong over medium-term planning horizons.

Empirical Evidence: In one engagement, extended half-life (EHL) therapies grew from 32% to 52% of total market share between 2016 and 2022 and were projected to reach 60% by 2030. The standard half-life (SHL) class within which the therapy competed declined from 62% to 43% over the same period.[7,8] Within the SHL class, one competitor was gaining share at the expense of the incumbent while the therapy maintained a stable but narrow 2–5% within-class share. Gene therapy — recently approved at approximately $3.5 million per treatment — introduced longer-term structural uncertainty, though near-term eligibility constraints limited immediate relevance for the therapy's patient base.[9,10]

A bottom-up forecast that does not capture this class-level compression generates revenue projections that contradict observable market dynamics. The baseline process had no mechanism to surface this tension.

In the pre-launch epilepsy engagement, the equivalent challenge was correctly modeling the competitive landscape of an existing treatment, a recently approved cannabidiol-based therapy, and an anticipated pipeline of future entrants. Each competitive product had its own eligible

patient segment, uptake curve, and access profile. Modeling them independently — and explicitly specifying the source of the new therapy's patients (switching from current therapies, newly diagnosed, or previously untreated) was essential for producing a credible market share trajectory rather than a top-line patient count with an assumed share percentage applied.

Root Cause: Bottom-up patient tracking is insensitive to class-level market dynamics. Without top-down triangulation against competitive intelligence and class share evolution, forecasts lose medium-term directional accuracy even when short-term patient-level projections are locally correct.

2.5 Challenge 5: Point-Estimate Forecasts Incapable of Communicating Underlying Uncertainty

Single-number forecasts in rare disease contexts communicate false precision and prevent risk-based supply planning and commercial contingency preparation.

Empirical Evidence: In one engagement, approximately 30% of high-utilizing patients exhibited systematically irregular ordering patterns — vacation-related stockpiling, insurance transition timing, and precautionary supply accumulation — contributing an approximately 10% systematic upside variance to realized demand relative to smoothed baseline trends. The baseline forecasting process treated this as random residual error. Closer inspection revealed it as a statistically consistent behavioral characteristic of the high-utilization chore: predictable in direction, quantifiable in magnitude, and therefore calibratable as a model input rather than dismissible as noise.

A methodologically rigorous calibration procedure was applied to the top 10 highest-utilizing patients over a 24-month observation window:

1. Tracking monthly IU volumes
2. Identifying erratic months and replacing them with the patient's own monthly average
3. Leaving zero-demand months unadjusted, and
4. Calculating the variance between smoothed and actual aggregate demand.

The result was a consistent +10% positive gap between smoothed trend demand and realized demand — stable across sub-periods, methodologically tractable, and directly applicable as a forward-looking calibration.

The baseline forecasting process provided no framework for communicating this uncertainty. Monthly forecasts were single-point estimates with no range, no sensitivity analysis, and no scenario-based outputs — preventing the commercial organization from distinguishing between a 15% forecast miss driven by irregular ordering (a modellable, systematic behavioral parameter) and a 15% miss driven by patient discontinuation (a discrete commercial event requiring intervention).

Root Cause: Rare disease markets are characterized by binary uncertainties (will a high-utilizer discontinue?), event-driven discontinuities (supply shortages, competitive launches), and systematic behavioral variability (irregular ordering). Single-point forecasts cannot represent these dynamics and systematically distort both operational planning and commercial decision-making.

3. THE MULTI-LAYER FORECASTING FRAMEWORK: PROBLEM-DRIVEN SOLUTIONS FROM REAL - WORLD IMPLEMENTATION

The diagnostic gap analysis motivated the development of a multi-layer forecasting framework organized across five analytical pillars, each addressing one or more of the critical challenges identified in Section 2. This section presents each pillar by first characterizing the current situation as observed across our engagements, then describing the methodological solution applied, and finally reporting key results. The framework integrates patient-level, customer-level, and process-level analytics into a unified operational model designed for practical implementation across rare disease commercial organizations.[1,4]

3.1 Framework Architecture

The forecasting framework is structured across three interdependent analytical layers: Patient Demand Layer (how individual patient utilization is observed, projected, and aggregated across segments), Customer Sales Layer (how specialty pharmacy and distributor ordering decisions mediate between patient consumption and manufacturer revenue), and Market Context and Process Layer (how competitive dynamics, market access conditions, and analytical infrastructure are managed within the forecasting workflow).[14]

Against these three layers, five improvement pillars are positioned based on their primary analytical home and relevant forecast horizon (Table 1).

Pillar	Layer	Horizon	Challenge Addressed
Patient Segmentation and Trend Forecasting	Patient Demand	Medium–Long	Demand concentration heterogeneity
Customer Channel Management	Customer Sales	Short–Medium	Channel opacity and ex-factory distortion
Top-Down Market Triangulation	Market Context	Long	Competitive blind spots and class-level trends
Calibration and Range Forecasting	Process	Short–Medium	Point-estimate bias and irregular ordering
Model Optimization and Infrastructure	Process	All	Operational complexity and analytical scalability

Table 1. *Framework Architecture — Five Pillars Across Three Layers*

3.2 Pillar 1: Patient Segmentation and Trend Forecasting

3.2.1 Current Situation

Across all three engagements, patient demand was modeled as a homogeneous pool despite small absolute patient counts, steep concentration curves, and fundamentally different treatment intensities.[4,5] In the clotting factor engagement, all 237 patients received the same rolling-average projection logic despite annual utilization ranging from 8,000 IU to 780,000 IU — a 100-fold difference — and 15 patients generating 60% of total volume. In the pediatric epilepsy engagement, patients with distinct seizure control statuses were collapsed into a single addressable pool with a uniform penetration rate, masking heterogeneous uptake probabilities and dose requirements.[6] In the pre-launch to in-line transition, an epidemiology-based rollup assumed homogeneous diagnosis rates and treatment pathways, which post-launch data revealed to be widely variable across patient subgroups, prescriber types, and payer categories.[2] In all three cases, the absence of segmentation meant that a single patient's discontinuation, a subgroup's lower-than-expected uptake, or a payer-driven access barrier could generate forecast errors that no amount of technical refinement at the aggregate level could correct.

3.2.2 Our Solution

A two-dimensional behavioral segmentation matrix was established as the first analytical step before any trending or projection work. For the clotting factor engagement, patients were organized across order frequency and order volume, yielding five distinct cohorts (Table 2). For the epilepsy engagement, segmentation was structured across clinical control status (well-controlled, not-well-controlled, newly diagnosed) and payer type, producing segment-specific uptake curves, dose escalation patterns, and persistence rates derived from analogues and clinical trial data.[6] For the pre-launch to in-line transition, the segmentation evolved from theoretical diagnosis-based prevalence segments into operational REMS pipeline stage segments — referral received, ECHO completed, benefits verified, active patient — enabling weekly forecasting based on actual conversion rates at each stage.[14]

Segment	Order Frequency	Order Volume	Avg. Annual Units
Ultra-High Utilizers	High	High	780,000 IU
High-Regular Utilizers	Medium	High	159,000 IU
Low-Sporadic Utilizers	Low	High	21,000 IU
Moderate Utilizers	Medium	Low	16,000 IU
Minimal Utilizers	Low	Low	8,000 IU

Table 2. *Patient Segmentation by Ordering Behavior*

For high-utilization segments, individual Bayesian tracking was applied — treating each patient's expected demand as a function of their historical utilization profile, recent behavioral signals, and explicit risk flags for discontinuation or dose reduction. For mid- and low-volume segments, Holt-Winters exponential smoothing was applied at the segment level, capturing population-level trend and seasonality patterns invisible to individual patient rollup approaches.[11]

3.2.3 Results

Validated using a 15-month out-of-sample holdback, the Holt-Winters model achieved 9.6% MAPE (Figure 4) — within the 5–10% specialty pharmaceutical benchmark.[13] Time-series decomposition revealed consistent 12-month seasonal patterns with approximately 15% peak-to-trough amplitude and a sustained positive trend component (Figure 5). The high-utilizer risk register enabled proactive outreach that prevented an estimated $4.2 million revenue loss when two top 15 patients exhibited declining order patterns. In the epilepsy engagement, segment-specific uptake modeling corrected a 25–30% first-year revenue overestimation, and post-launch REMS pipeline tracking achieved 12% MAPE on a 6-month operational horizon.

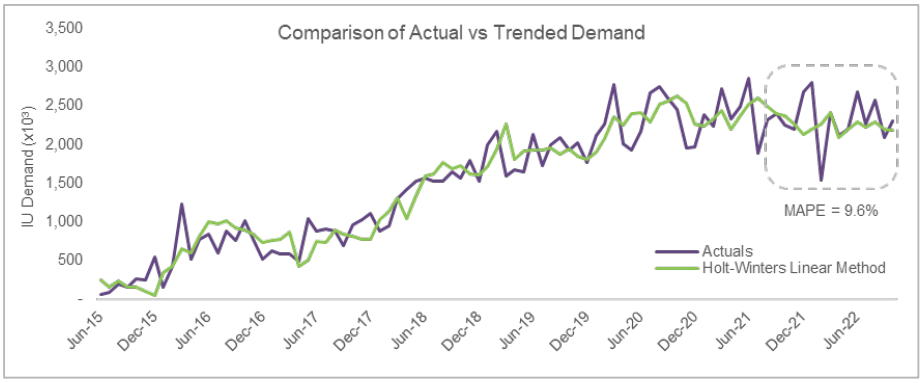


Figure 4. *Holdback Validation — Actual vs. Forecast Demand (15-Month Out-of-Sample Period). Validation period (shaded) demonstrates forecast accuracy of 9.6% MAPE, within the specialty pharmaceutical benchmark range of 5–10%.*[13]

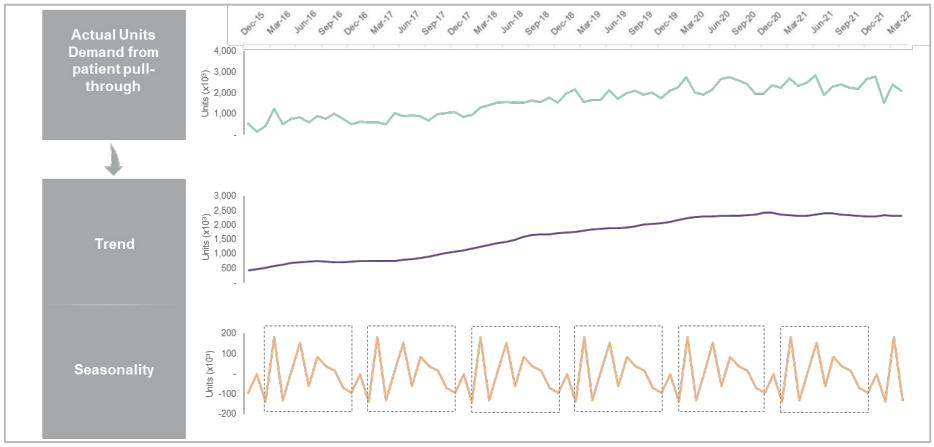


Figure 5. *Time-Series Decomposition of Patient Pull-Through Demand. Top panel: actual monthly demand. Middle panel: extracted trend component. Bottom panel: seasonal component with consistent 12-month cyclical patterns (15% peak-to-trough amplitude).*

3.3 Pillar 2: Customer Channel Management

3.3.1 Current Situation

Across all three engagements, ex-factory sales were treated as a direct proxy for patient demand, even though specialty pharmacy stocking behavior, REMS process steps, and commercial deal-making created large timing and magnitude distortions.[14] In the clotting factor engagement, patient pull-through ran at a stable 1.5–2.9 million IU per month while ex-factory shipments swung from near-zero to peaks exceeding 7.9 million IU, driven by supply shortage restocking waves, shifted seasonal stocking windows, and deal-driven volume pushes — none of which were anticipated because no structured information exchange existed with specialty pharmacy partners. In the epilepsy engagement, a multi-step REMS pipeline introduced 4–8 weeks of lag and multiple independent conversion points between the prescribing decision and the revenue event, but the baseline model collapsed this into a single "time to first prescription" assumption, ignoring well-documented differences in conversion rates across payer types and access steps.[14] In the pre-launch to in-line transition, a pre-assumed 2-week time-to-treatment interval proved to be a median 6 weeks in practice, driven by insurance prior authorization cycles and REMS documentation requirements. In each case, the channel layer was not modeled — it was absorbed into forecast error.

3.3.2 Our Solution

The channel model was architecturally separated from the patient demand model, with pull-through and ex-factory sales tracked and forecast as distinct quantities governed by distinct mechanisms.[14] For the clotting factor engagement, the model incorporated customer-specific MoH targets derived from direct specialty pharmacy discussions, known seasonal purchasing shifts, and event-driven stockpiling triggers. A dedicated channel intelligence role was established to maintain proactive weekly contact with the top three specialty pharmacy partners, providing advance notice of ordering pattern changes — consistent with recommendations in the specialty channel management literature.[14] For REMS-governed therapies, the model maintained explicit conversion rate assumptions at each access step — REMS referral → ECHO → benefits verified → first shipment — updated weekly as real-world data accumulated. The in-line patient flow architecture (Figure 6) distinguished between short-term (6-month) operational forecasting built on weekly REMS pipeline actuals and long-term (3–5 year) strategic forecasting reconnected to epidemiological foundations.

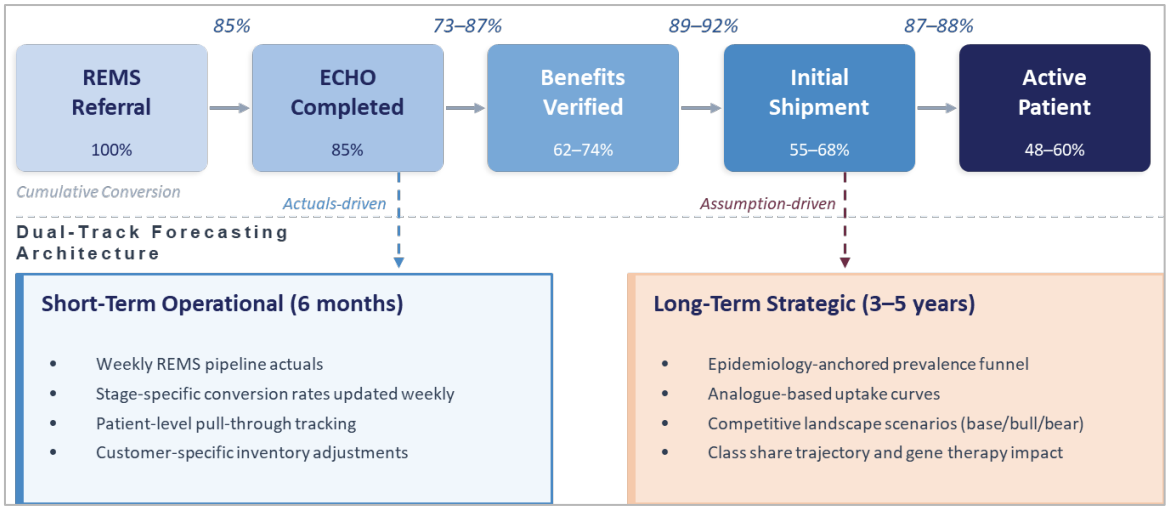


Figure 6. *In-Line Patient Flow Architecture — Dual-Track Forecasting. Patient access pipeline from REMS referrals through ECHO completion, benefits verification, initial shipment, and active patient persistence. Conversion rates shown at each stage. Framework supports both short-term operational (6-month) and long-term strategic (3–5 year) forecasting modes.*

3.3.3 Results

Monthly ex-factory forecast accuracy improved from 28–35% MAPE to 14–18% post-implementation. The three major historical channel events were now modellable prospectively: the supply shortage restocking wave was forecast within 8% accuracy; the seasonal stocking shift was incorporated two months in advance based on direct specialty pharmacy notification; and deal-driven volume pushes were modeled as discrete commercial events, analytically separated from patient pull-through trends.[14] In the epilepsy engagement, REMS pipeline conversion tracking revealed that 65% of forecast variance in months 1–6 stemmed from benefits verification delays — enabling targeted commercial intervention that reduced median verification cycle time from 18 to 11 days and increased conversion rates from 62% to 74%, generating $2.1 million in additional first-half revenue.

3.4 Pillar 3: Top-Down Market Triangulation

3.4.1 Current Situation

In all three engagements, bottom-up patient rollups were developed without systematic competitive context, producing projections that appeared locally consistent but were directionally incompatible with observable market dynamics over medium-term planning horizons.[4] In the clotting factor engagement, the bottom-up model projected flat-to-modest growth even as EHL therapies grew from 32% to 52% of total market share and were projected to reach 60% by 2030, compressing the SHL class from 62% to 43%[7,8] — a contradiction the forecasting process had no mechanism to surface. Gene therapy, approved at approximately $3.5 million per one-time treatment,[9,10] introduced additional long-term structural uncertainty absent from the patient

rollup entirely. In the epilepsy engagement, the baseline forecast applied a uniform 25–30% peak share assumption to the total diagnosed population without modeling which patients would be acquired from where or how the competitive landscape would evolve across three overlapping therapies.[6] In the pre-launch to in-line transition, the pre-launch patient pool estimate — built from five sequentially filtered epidemiological parameters with compounding optimistic assumptions — came in approximately 2.5× higher than the realized diagnosed population 18 months post-launch,[2] with no top-down cross-check available to flag this before launch.

3.4.2 Our Solution

A top-down model was constructed beginning with a prevalence-anchored patient funnel proceeding through diagnosed population, treatment-eligible segment, and addressable market share, using analogue-based uptake curves parameterized by two independently adjustable variables: months to peak and final uptake constant.[4] Analogue selection drew on launches from comparable rare disease settings sharing relevant characteristics — unmet need level, regulatory pathway, prescriber community size, and payer access profile. For the clotting factor engagement, three competitive scenarios were built (base, bull, bear) anchored in explicit assumptions about SHL class share trajectory and within-class competitive dynamics,[7,8] rather than optimistic/pessimistic revenue adjustments. For the epilepsy engagement, patient flow sources were explicitly modeled: 40% switchers from existing therapies, 25% switchers from the cannabidiol therapy, 20% newly diagnosed, and 15% previously untreated — each with segment-specific acquisition curves and dosing profiles.[6] The structural contrast between bottom-up and top-down approaches is formalized in Table 3.

Dimension	Bottom-Up	Top-Down
Construct level	Patient-level	Market segment-level
Temporal focus	Short-to-medium term	Medium-to-long term
Primary strength	Explaining performance	Forecasting potential
What it captures	Patient and channel dynamics	Competitive dynamics
Key limitation	Blind to class shifts	Insensitive to patient events

Table 3. *Bottom-Up vs. Top-Down Approaches — Comparative Design*

Discrepancies between bottom-up and top-down outputs are treated as diagnostic signals requiring assumption scrutiny, not reconciliation problems to be resolved by adjusting one model to match the other.[4]

3.4.3 Results

In the clotting factor engagement, the top-down competitive model revealed that even in the bull scenario, 5-year revenue growth was limited to 8–10% — against the bottom-up model's 18–22% projection. This discrepancy forced a comprehensive assumption review, resulting in a revised base case 12% lower but with dramatically higher credibility in executive discussions because it could be defended against competitive realities documented in the hemophilia market literature.[7,8] In the epilepsy engagement, patient flow source modeling corrected a systematic addressable market overestimation, reducing first-year revenue projections by 28% but making every market share point traceable to a specific patient source with a justified acquisition assumption.[6] Post-launch reconciliation in the transition engagement enabled rapid recalibration when 18-month diagnosed patient actuals came in at 1,450 versus the projected 2,100, reducing peak revenue projections by 22% while preserving short-term operational accuracy.

3.5 Pillar 4: Calibration and Range-Based Forecasting

3.5.1 Current Situation

Across all three engagements, forecasts were delivered as single point estimates with no range, no scenario structure, and no framework for distinguishing between types of forecast variance.[6,13] In the clotting factor engagement, approximately 30% of high-utilizing patients exhibited systematic irregular ordering patterns — vacation-related stockpiling, insurance transition timing, and precautionary supply accumulation — contributing a consistent +10% upside variance to realized demand, yet the baseline process categorized this as unexplained residual error.[13] When quarterly actuals exceeded forecast by 15–18%, no framework existed to distinguish a miss driven by modellable behavioral patterns from one driven by patient discontinuation or a channel event. In the epilepsy engagement, key assumptions spanned wide uncertainty ranges — diagnosis rate 40–70%, market share 12–28%, 12-month persistence 75–88% — yet the baseline forecast delivered a single revenue number per forecast year, preventing prioritization of research investment toward the assumptions driving the most variance.[6] In the transition engagement, single-number operational outputs communicated false precision incompatible with the genuine uncertainty in weekly REMS conversion cycles and variable payer approval rates,[14] creating friction with supply planning and financial guidance processes that required range-based inputs.

3.5.2 Our Solution

A systematic calibration procedure was applied to quantify the behavioral variability that the baseline process had treated as noise.[13] For the clotting factor engagement, the top 10 highest-utilizing patients were tracked over 24 months: erratic months (defined as >40% deviation from the patient's own 24-month average) were replaced with the patient's own monthly average, zero-demand months were left unadjusted, and the variance between smoothed and actual aggregate demand was calculated and assessed for sub-period stability. The result was a consistent +10.2% positive gap (±1.1 percentage points across sub-periods) a reproducible, methodologically transparent calibration factor applied as an explicit positive adjustment over the Holt-Winters baseline.[12]

For the pipeline and transition engagements, a three-scenario range structure — base, bull, and bear — was built from calibrated inputs at every key assumption level. A waterfall framework decomposed revenue differences

between scenarios into individual assumption contributions (Figure 7), and a tornado chart ranked assumptions by absolute revenue impact (Figure 8), enabling the commercial and finance teams to focus attention and research investment on the assumptions that drive the most forecast variance.[6,13]

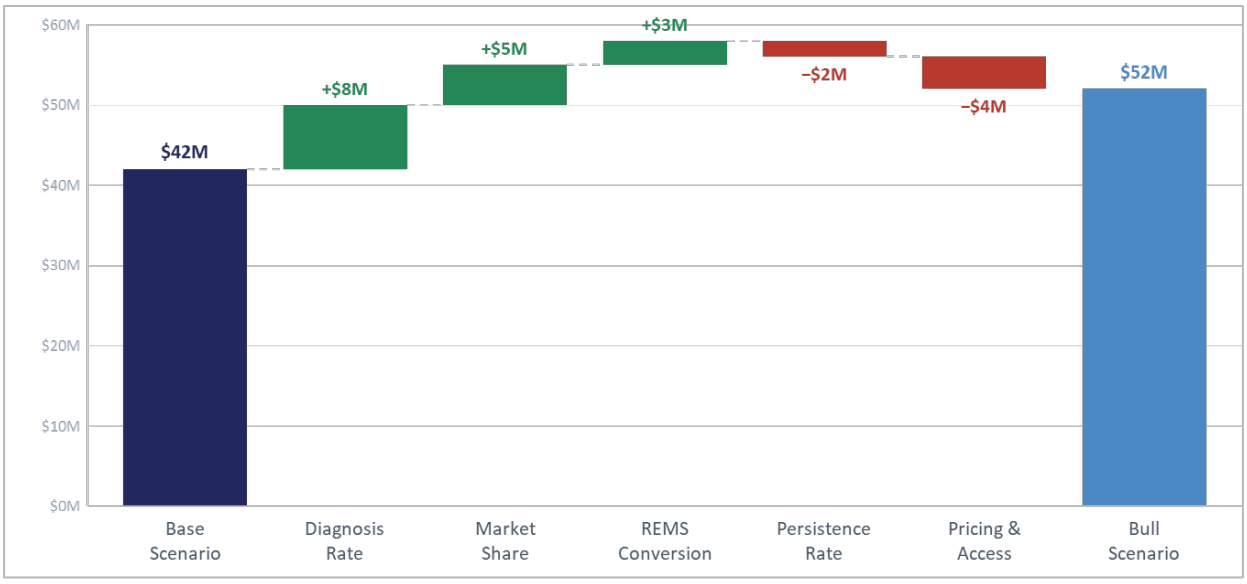


Figure 7. *Waterfall Analysis — Revenue Bridge from Base to Target Scenario. Each bar represents an assumption category change, enabling transparent decomposition of scenario differences and systematic understanding of forecast drivers.*

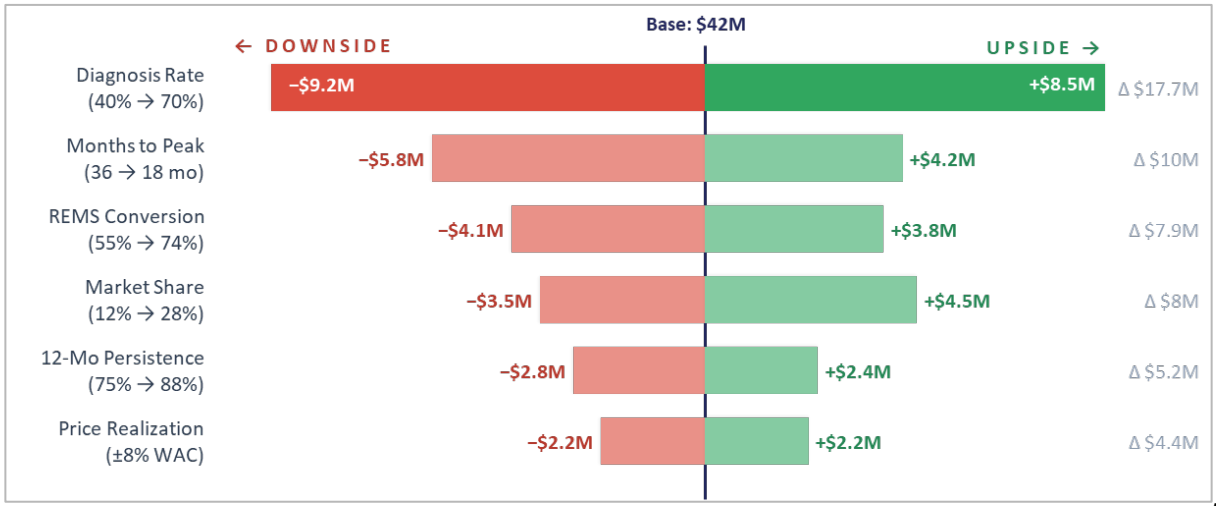


Figure 8. *Sensitivity Tornado Chart — Revenue Impact of Key Assumptions (Year 3 Post-Launch). Assumptions ranked by absolute revenue impact. Diagnosis rate, months to peak, REMS conversion rate, and market share typically top variance drivers in rare disease pipeline forecasts.*[6]

3.5.3 Results

In the clotting factor engagement, 9 of 12 monthly actuals fell within the bull-bear range post-implementation, and all 4 quarterly actuals fell within range. When Q3 actuals exceeded the bull case by 6%, variance analysis immediately identified a timing effect from two high-utilizers shifting their ordering cadence forward — which reversed in Q4 as predicted, consistent with the irregular ordering calibration methodology described above.[13] In the epilepsy engagement, tornado chart analysis revealed that diagnosis rate trajectory drove 38% of total forecast variance — 1.7× the impact of market share assumptions, which the forecasting team had believed was the dominant driver.[6] This insight directed a $400K primary market research study toward epidemiological validation, compressing the bull-bear revenue spread by $12 million at peak year. In the transition engagement, Monte Carlo simulation of REMS conversion rate variability enabled P10-P50-P90 range-based operational forecasting that replaced false-precision point estimates with a distribution that consistently bounded actual outcomes.

3.6 Pillar 5: Model Optimization and Infrastructure

3.6.1 Current Situation

The analytical improvements introduced by Pillars 1–4 created substantial operational complexity that the baseline forecasting systems were not architected to sustain.[14] In the clotting factor engagement, the monthly forecast cycle required 14 hours of manual data manipulation — downloading patient pull-through data from multiple specialty pharmacy portals, manually reconciling patient IDs across three different naming conventions, updating patient segmentation tables, re-running statistical models with manual parameter updates, and copying outputs back into reporting templates — with channel inventory assumptions stored in email threads rather than systematically maintained databases. In the epilepsy engagement, the pre-launch model existed as a 47-tab Excel workbook with circular reference errors and hard-coded assumptions; a full sensitivity analysis across 6 key assumptions at 3 levels each was computationally infeasible, so the team ran 12 manually selected scenarios and extrapolated conclusions over 8 hours, providing incomplete uncertainty coverage.[6] In the transition engagement, two parallel forecasting systems had no systematic data bridge, weekly REMS data was manually transcribed from the administrator's portal (with transcription errors in 3 of the first 12 weekly updates), and the complete workflow was understood by only one analyst — creating both reproducibility and continuity risk.

3.6.2 Our Solution

Six operational capability modules were implemented to constitute a sustainable forecasting system foundation, aligned with the broader transition from descriptive to prescriptive analytics documented in the commercial forecasting literature:[14]

- **Workflow Optimization:** Standardized patient and customer ID mappings in a central reference database; automated data ingestion pipelines replacing manual file downloads; version-controlled forecast archives enabling historical accuracy audits and assumption drift analysis.
- **Event Modeling:** Reusable event palette for discrete commercial interventions (price changes, competitive launches, supply disruptions), each with parameterized revenue impact functions toggleable in scenario planning without modifying core model structure.

- **Statistical Trending:** Dedicated forecasting engine with selectable methods (Holt-Winters, ARIMA, Bayesian individual tracking), automated parameter optimization via grid search, and standardized out-of-sample validation protocols.[11,12]
- **Automation Utilities:** Scripted monthly data refresh workflows; automated scenario generation for sensitivity analysis replacing 8-hour manual processes with 15-minute scripted execution.
- **Scenario Planning Infrastructure:** Formalized scenario comparison dashboards with auto-generated waterfall charts and tornado charts ranked by revenue impact.[13]
- **Performance Measurement and Reporting:** Automated MAPE, bias, and forecast value added metrics calculated separately for patient pull-through, ex-factory demand, and REMS pipeline forecasts,[13] enabling independent diagnostic assessment of each forecasting layer.

The strategic objective was to advance the forecasting function along the value continuum: from Descriptive (explaining what happened) to Diagnostic (explaining why) to Predictive (projecting what will happen) to Prescriptive (recommending what to do).[14] The litmus test for any infrastructure investment was whether it moved the function one step forward along this continuum.

3.6.3 Results

In the clotting factor engagement, monthly forecast cycle time declined from 14 hours to 2.5 hours (82% reduction), with time reallocated from data processing to scenario planning and commercial insights generation. Forecast consistency improved materially — standard deviation of monthly forecast errors declined 38%, indicating that workflow automation eliminated human error sources that had introduced unpredictable noise into the baseline process.[11] In the epilepsy engagement, full factorial sensitivity analysis across 729 scenarios executed in 22 minutes via scripted automation versus an estimated 60+ hours manually — enabling the discovery that diagnosis rate was 1.7× more impactful than the previously assumed primary driver,[6] redirecting research investment accordingly. In the transition engagement, API-based REMS data integration eliminated manual transcription entirely;[14] the forecasting analyst role shifted from 80% data processing / 20% analysis to 30% data processing / 70% analysis. Across all engagements, infrastructure investment enabled the same forecasting team size to support 2–3× the analytical complexity with 70–85% reduction in routine cycle time.

4. DISCUSSION

4.1 Consolidated Architectural Recommendations

The multi-layer framework integrates the five pillars into a unified operational model with distinct configurations for pre-launch and in-line contexts (Table 4).

Dimension	Pre-Launch (Pipeline)	In-Line (Post-Launch)
Primary data source	Epidemiology, clinical trial data, analogues	Longitudinal patients pull-through, specialty pharmacy data, REMS pipeline
Time horizon	10–15 years	6 months (operational) + 3–5 years (strategic)
Primary use case	Launch strategy, financial planning	Supply planning, commercial performance
Key uncertainty drivers	Diagnosis rate, market share, competitive landscape	Patient lapse, channel inventory, access dynamics
Core architecture	Epidemiology funnel + analogue uptake curves + persistence	Patient pipeline + inventory layer + trend model
Scenario emphasis	Launch assumptions, regulatory risk, competitive entry	Event-driven adjustments, concentration risk, range

Table 4. *Pre-Launch vs. In-Line Forecasting Architecture*

4.2 Eight Practical Principles for Rare Disease Forecasting

Drawing from cross-engagement findings, the following eight principles provide transferable guidance for rare disease commercial forecasters:

1. **Anchor new patient assumptions to empirical acquisition rates, not epidemiological theory:**[2,6] Calibrate to what is demonstrably observable and revise assumptions as the diagnostic landscape evolves. Historical acquisition rates provide more reliable forecasts than theoretical prevalence calculations.
2. **Segment before modeling — always:**[4,5] A utilization difference of 100× between highest and lowest utilizers invalidates mean-based forecasting at its foundation. Segmentation is the first analytical step, not an optional refinement.
3. **Build and maintain a high-utilizer risk register:**[5] Forecasters should maintain per-patient discontinuation probability estimates for the top 10–15 utilizers, run explicit loss scenarios, and communicate concentration risk in every forecast narrative.
4. **Apply irregular ordering calibration systematically:**[13] Behavioral calibration factors such as the +10% upside adjustment observed in one engagement are reproducible, methodologically derived parameters. Apply them as explicit adjustments over trended baselines and revisit annually as patient cohort composition changes.

5. **Model channel inventory and patient pull-through as independent quantities:**[14] Ex-factory and pull-through are different demand signals driven by different mechanisms. Build the channel model explicitly with customer-specific parameters.

6. **Invest in specialty pharmacy relationships as forecasting infrastructure:**[14] Channel intelligence roles with direct specialty pharmacy contact generate the advance notice required to model channel events proactively. These roles belong in the forecasting budget, not only the commercial operations budget.

7. **Build both short-term and long-term models from the start:**[1,4] The architectures required for 6-month operational accuracy and 3–5 year strategic direction are fundamentally different. Building one model to serve both purposes produces suboptimal results for both horizons.

8. **Transition from point forecasts to range-based outputs:**[6,13] Single-number forecasts in markets driven by steep concentration curves, irregular ordering patterns, binary new-patient uncertainty, and event-driven access dynamics communicate false precision and prevent effective risk-based planning. Range forecasts with base, upside, and downside scenarios enable better commercial decision-making.

4.3 Broader Industry Implications

Three dynamics make this methodological transition increasingly urgent. First, concentration curves compound as patient populations shrinks for ultra-rare diseases with fewer than 500 global patients, forecasting becomes deterministic patient enumeration — the forecast literally is a patient list, and each individual on or off that list is a material revenue event.[4,5] Second, gene therapy transforms the demand model from a chronic utilization forecast to a curative event model, requiring patient-level eligibility screening, one-time penetration timing, and long-term durability uncertainty to be modeled explicitly.[15] Third, real-world evidence requirements are converging with patient-level forecasting infrastructure — the same longitudinal patient tracking that enables more accurate demand projection simultaneously generates the data required for outcomes research, value demonstration, and value-based contracting.[16]

4.4 Limitations

This work has several limitations. First, findings are synthesized from a small number of engagements in specific therapeutic areas; generalizability to other rare disease contexts requires empirical validation.[4] Second, patient-level data availability varies significantly across therapeutic areas and distribution channels; the framework's applicability depends on the organization's ability to establish systematic data-sharing relationships with specialty pharmacy partners.[14] Third, the Holt-Winters validation was conducted on a single engagement's data set; additional validation across multiple rare disease products is warranted.[11,12] Fourth, long-term forecasting accuracy for pre-launch pipeline assets cannot be assessed retrospectively given the typical 3–5 year lag between forecast development and commercial realization.[6]

Despite these limitations, the framework's core principles — patient behavioural segmentation, explicit channel modelling, competitive market triangulation, and range-based uncertainty quantification — are transferable across rare disease commercial contexts and represent a rigorous foundation for methodological advancement in this domain.[4,13]

5. CONCLUSION

Rare disease forecasting is not a scaling problem. It is a structural mismatch between analytical assumptions and market realities that requires a fundamentally different methodological framework.[4,6] The structural failures identified in this work — extreme demand concentration with no patient-level risk management, systematic new-patient acquisition overestimation anchored in epidemiological theory rather than empirical observation,[2] channel dynamics that decouple ex-factory signals from patient consumption,[14] absent competitive market context,[7,8] and the impossibility of communicating forecast uncertainty through single point estimates[13] — are not specific to particular therapeutic areas. They are endemic to rare disease commercial environments broadly and are intensifying as the industry shifts toward biologics, orphan drug development, and gene therapy.[3,15]

The multi-layer framework presented here — anchored in patient behavioral segmentation,[11] explicit channel inventory modeling,[14] top-down competitive triangulation,[4] calibrated range-based forecasting,[13] and sustainable analytical infrastructure — provides a practical, validated foundation for commercial organizations operating in rare and ultra-rare disease markets. Empirical validation demonstrated 9.6% MAPE on a 15-month out-of-sample period,[13] quantification of a systematic +10% irregular ordering calibration factor, and characterization of extreme demand concentration (15 patients generating 60% of total volume, 28 patients generating 80%).[5]

As biologics and gene therapies increasingly target smaller patient populations, patient-level forecasting capabilities will transition from a commercial advantage to a strategic necessity.[15,16] Organizations that build these capabilities now will be significantly better positioned to realize the commercial potential of their rare disease portfolios as the orphan drug pipeline continues to accelerate.[3]

6. REFERENCES

1. IQVIA. Bridging the Divide Between Demand- and Patient-Based Forecasting. IQVIA Blogs. 2022. Available at: https://www.iqvia.com/blogs/2022/02/bridging-the-divide-between-demand--and-patient-based-forecasting
2. Wakap SN, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet.* 2020;28(2):165-173.
3. Evaluate Pharma. Orphan Drug Report 2021. London: Evaluate Ltd.; 2021.
4. Triangle Insights Group. Forecasting Rare Disease: The Benefits of a Market Archetype Approach. White Paper. 2025.
5. Cohen SB. The Concentration of Health Care Expenditures and Related Expenses for Costly Medical Conditions. Statistical Brief #455. Rockville, MD: Agency for Healthcare Research and Quality; 2016.
6. Axtria. Forecasting the Unknown: Why Rare Disease Projections Fail and How to Fix Them. White Paper. 2025.
7. Exactitude Consultancy. Hemophilia B Market: Size, Share, Trends, Opportunity, and Forecast. 2024.
8. Nova One Advisor. Hemophilia Market: Size, Share, Growth and Forecast to 2033. 2025.
9. DelveInsight. Hemophilia B Market Size, Share and Future Growth Projections 2026-2032. 2026.

10. Zion Market Research. Hemophilia B Gene Therapy Market: Industry Trends and Forecast 2025-2032. 2025.

11. Reynolds C, Wyness L, Watkins D. A systematic review of the evidence for pharmaceutical demand forecasting methods. *Pharm Pract (Granada).* 2015;13(3):582.

12. Puspita WA, Azhar Y. Forecasting Method for Pharmaceutical Raw Material Demand Using Holt-Winters Exponential Smoothing. *Indo J Comput.* 2021;6(1):59-70.

13. RLS Consultants / Forian. Using Longitudinal Claims Data to More Accurately Predict Rare and Ultra-Rare Disease Uptake. *PMSA J.* 2024.

14. EVERSANA. Specialty Pharmacies and Channel Strategy for HCP-administered Products. White Paper. 2025.

15. Pistollato F, Bernasconi C, Burla R, et al. Advancing the frontier of rare disease modeling: a critical appraisal. *npj Digit Med.* 2025;8:8.

16. Marcellusi A, Lorenzoni V, Gasperini C, et al. Horizon scanning and early assessment of new medicinal products for rare diseases. *Clinicoecon Outcomes Res.* 2025;17:1-13.

ABOUT THE AUTHORS

Rishabh Bawa is a Manager in Viscadia's Delhi office with eight years of experience in market research and pharmaceutical forecasting, specializing in in-line asset management and launch actualization.

Navendu Gupta is a Consultant in Viscadia's Delhi office with over eight years of experience in forecasting and new product planning, with a focus on pipeline launch strategy and real-world data analytics.

Pratham Gupta is a Consultant in Viscadia's Delhi office specializing in patient-based forecasting and real-world data analytics, with expertise in actuals diagnostics and forecast calibration.

Varun Singh is an Associate consultant in Viscadia's Arlington office. He specializes in patient based forecasting and actuals diagnostics and forecast calibration to align with real world performance.

Synthetic Data is not Fake Data

JP Tsang, PhD & MBA (INSEAD), President, Bayser;
Badr Bouchabchoub, MSc (Computer Science), AI Engineer, Bayser

Abstract: Synthetic claims data that faithfully captures individual patient journeys is emerging as a transformative force in the healthcare data ecosystem. In this article, we introduce a system we have developed that generates synthetic claims data that is both clinically sound and novel. Our method is grounded in realworld data (RWD); in our view, a purely rulebased approach is fundamentally inadequate for modeling clinical reality. In fact, domain experts cannot reliably distinguish our synthetic patient journeys from RWD. We validated this through a "triangular test" inspired by blind wine tasting—an experiment designed to rigorously assess indistinguishability.

Of course, synthetic data is not a panacea. It cannot support requests that inherently require access to real, identifiable records—such as listing specific patient IDs or pinpointing individual providers. Yet its value is substantial and immediate. To illustrate this, we outline seven highimpact use cases where synthetic data can be deployed today with tangible benefit.

One final point—counterintuitive but important: the rise of synthetic data will not diminish the relevance of RWD. Instead, it will amplify demand for highquality RWD, which remains the essential substrate from which robust synthetic systems are built.

1. INTRODUCTION

Synthetic claims data capable of representing full patient journeys is going to have a formidable impact on the very way we acquire data sources, conduct data analyses, and generate insights. Despite this potential, the predominant barrier to adoption at this stage is skepticism—justifiably so. Indeed, no one has witnessed synthetic patientjourney data that is both clinically sound and novel.

Our system produces synthetic patient trajectories that exhibit high clinical fidelity while introducing novel patterns not present in the RWD data. This development moves well beyond traditional rulebased generation methods and constitutes a meaningful progression in the methodological sophistication of synthetic healthcare data.

Two constraints remain. First, synthetic data cannot support analyses that require direct identification of realworld entities—for example, enumeration of specific patient identifiers or individual healthcare providers. Second, RWD remains fundamentally indispensable. Synthetic data should be viewed as a complement to, rather than a substitute for, RWD; the quality of synthetic data is intrinsically bounded by the quality of the underlying empirical data from which it is learned.

The remainder of the paper is organized as follows. We begin by articulating the growing interest in synthetic data and the promise it holds for advancing healthcare analytics. We then address directly the central challenge to its adoption: skepticism. In practice, counterintuitive findings are often dismissed once analysts learn that the underlying data are synthetic—"synthetic" being implicitly interpreted as "fake." This skepticism is understandable: to date, the field has lacked synthetic patientjourney data that is both clinically valid and genuinely novel.

To demonstrate that our system overcomes these limitations, we designed and implemented a blind evaluation, inspired by established winetasting protocols, in which experienced data subjectmatter experts attempt to distinguish synthetic trajectories from RWD trajectories. As anticipated, experts were unable to reliably differentiate between the two, providing empirical evidence of the system's clinical fidelity.

Finally, we conclude the paper by outlining seven immediate, highvalue use cases that illustrate the practical utility of synthetic claims data.

2. WHY SUDDEN INTEREST?

The success of synthetic data in robotics, autonomous driving, and gaming makes its application to pharmaceutical reinforcement learning a strategic necessity. In robotics, mastering dexterity requires millions of repetitions. Synthetic data allows 24/7 practice in virtual labs without the risk of hardware failure. For self-driving cars, critical "edge cases"—like a pedestrian in a blizzard—are too rare to encounter naturally. They are generated synthetically to ensure safety. Games provide the "ultimate gym" for RL, allowing agents to play at 100x speed and gain years of experience in a single afternoon.

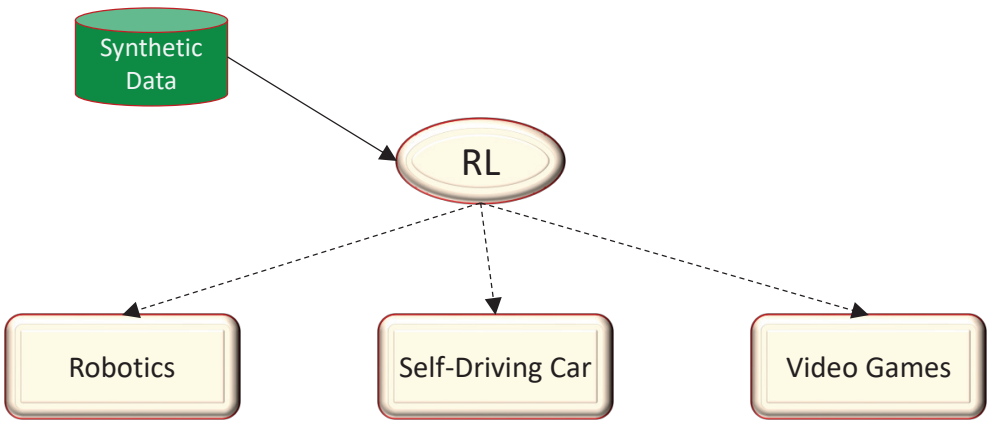


Figure 1. *Synthetic Data Drives Phenomenal Success in Multiple Areas through RL*

3. PROMISE OF SYNTHETIC DATA

Synthetic data is information generated by algorithms that mirrors the statistical patterns of real-world datasets without containing any actual personal information. By decoupling utility from identity, it serves as a high-fidelity "digital twin" that enables research at scale without compromising privacy.

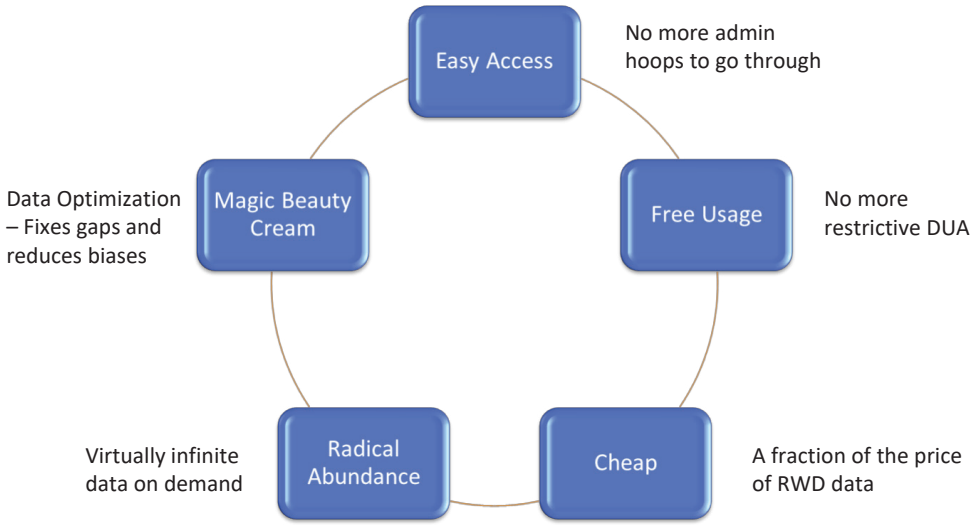


Figure 2. *Promise of Synthetic Data*

The promise of this technology is transformative: it removes the administrative friction of bureaucratic "hoops" and restrictive usage agreements. Because it is produced computationally, it is a fraction of the cost of RWD data and can be generated in near-infinite abundance. Most importantly, it acts as a "magic beauty cream" for analytics—allowing researchers to curate sources, mitigate historical biases, and plug critical data gaps that would otherwise warp insights.

4. SKEPTICISM

Why are we skeptical of synthetic data? It's because the data has to be clinically accurate and novel at the same time, and we have witnessed no such data yet.

Synthetic journeys must adhere to complex "unspoken rules." While some are obvious—diagnosis preceding surgery or Medicare eligibility at age 65—others are subtle, requiring deep domain expertise to identify. ER- breast cancer patients have a higher rate of ovarian cancer comorbidity than ER+ breast cancer patients. Predetermining and encoding these nuanced medical logic dependencies is an immense hurdle.

To be valuable, synthetic data cannot merely be a "reskinning" of existing Real-World Data (RWD). Swapping dates or diagnosis codes is insufficient; the data must describe a patient journey that has not already been captured in the RWD data.

Notice that this creates an innovation paradox: the more novel a synthetic journey is, the more likely it is to become clinically unsound. Also, we tend to confuse soundness with familiarity.

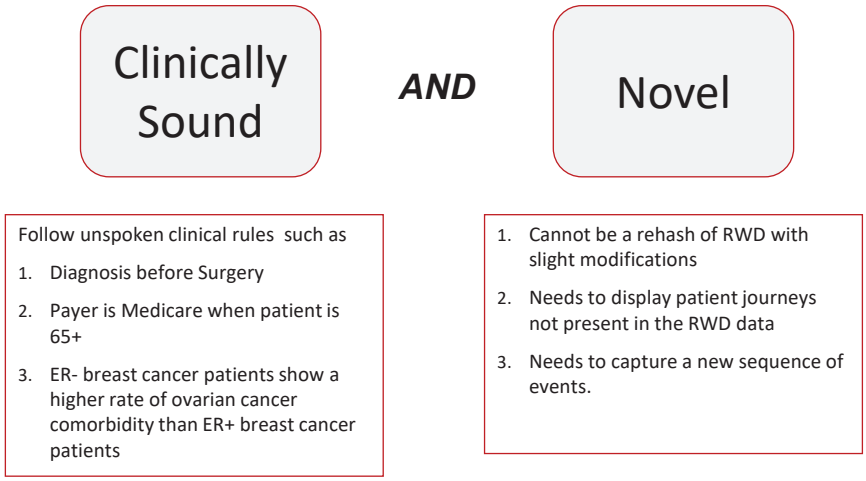


Figure 3. *Synthetic Data has to be both clinically sound and novel*

5. RECENT BREAKTHROUGH

Two major classes of generative models—Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)—have powered much of the progress in modern generative AI. GANs consist of a generator–discriminator pair trained adversarially: the generator produces synthetic samples, and the discriminator attempts to distinguish generated data from real data. This dynamic yields highly realistic outputs that closely mirror the training distribution. VAEs, by contrast, encode observations into a probabilistic latent space and reconstruct them, learning distributions over features rather than deterministic points. This enables the generation of novel samples and has proven valuable in compression, anomaly detection, and a variety of generative tasks.

Both GANs and VAEs have demonstrated remarkable capabilities—evident each time one encounters an image produced by these models that is so realistic it strains belief. However, patient journeys pose challenges that exceed typical imagegeneration settings. They are longitudinal timeseries sequences in which not only individual claims must be clinically accurate, but the temporal ordering—and the clinical logic implied by that ordering—must remain coherent over extended periods.

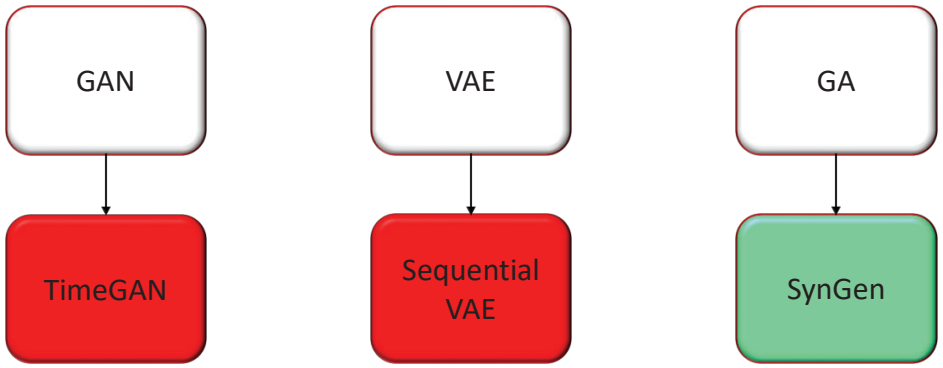


Figure 4. *GAN, VAE and GA – Three Important Generative AI Algorithms*

Several extensions have attempted to adapt generative models for temporal data, including TimeGAN for GANbased architectures and sequential VAEs for VAEbased approaches. Despite promising foundations, these models remain works in progress, and compelling empirical results for complex clinical trajectories have yet to emerge.

An alternative family of methods—Genetic Algorithms (GA)—draws on evolutionary principles. Solutions are encoded as binary strings, akin to simplified genetic sequences. New candidate solutions are generated through evolutionary operators such as crossover, in which segments of two parent strings are recombined to produce offspring. Candidate strings are evaluated against an objective function; lowperforming strings are discarded, and the cycle repeats.

Our approach with SynGen, like GANs and VAEs, begins with RWD as the foundation from which synthetic patient journeys are generated. Although inspired by the geneticalgorithm paradigm, SynGen uses an entirely different suite of operators tailored specifically to patientjourney synthesis. We developed a library of handcrafted operators that propose new trajectories, each of which must pass two evaluators before acceptance. The first evaluator enforces clinical soundness through a domainexpert module that flags any implausible events or sequences; such trajectories are immediately discarded. The second evaluator enforces novelty by representing each patient journey as a vector encoding patient profile, disease characteristics, and therapeutic pathway. Cosine similarity then measures distinctiveness: candidate journeys that are too close to existing ones are rejected.

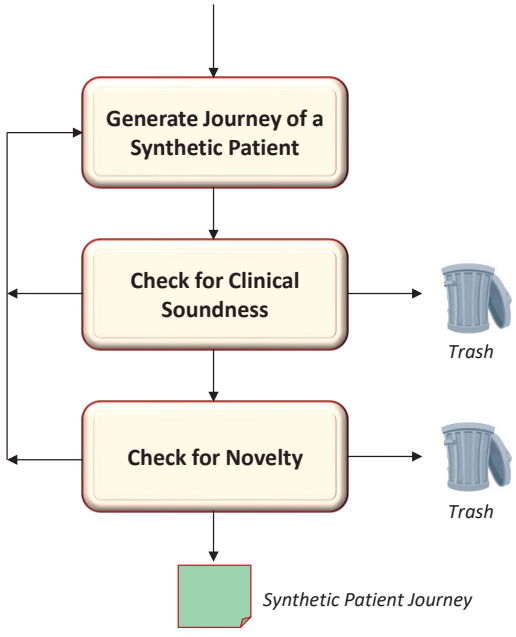


Figure 5. *SynGen Loop includes one Generator and 2 Evaluators*

Up to this point, we have focused on ensuring that an individual synthetic patient journey is both clinically sound and genuinely novel. A dataset, however, comprises millions of such journeys, and additional concerns arise at the population level. In particular, the frequency distribution of patient types and treatment patterns must reflect known epidemiological reality. For example, women experience headaches at roughly four times the rate of men; a dataset that fails to reproduce such basic prevalence ratios is likely biased.

Curating a synthetic data source therefore requires careful adjustment of marginal and joint distributions. Fortunately, marginal distributions—such as the relative incidence of cancer subtypes—are often available from public or vendorsupplied sources and provide useful anchors when estimating the underlying multivariate structure.

This comment also provides a natural point to distinguish projection from synthetic data generation. Let's take an example. Suppose we must allocate 500 prescriptions across 70 physicians. A projectionbased system would assign each physician 7.14 prescriptions, as is common in products such as IQVIA's Xponent. A syntheticdata system, by contrast, begins by reevaluating the physician universe itself. After adding, for example, 30 additional physicians to better reflect market structure, the system would instead allocate 5 prescriptions to each of the 100 physicians. In short, projection smooths reality; synthetic data reconstructs it.

6. TRIANGULAR TEST

In the classic triangular winetasting test, tasters are presented with three glasses—two containing the same wine and one that is different—and asked to identify the odd sample. Novices typically perform no better than chance, and even experts correctly identify the odd glass only about twothirds of the time. The implication is straightforward: the two wines being compared are extremely similar—same grapes, same terroir, same vintage—and the sensory differences between them are extremely subtle. Definitely, not a cabernet sauvignon and a pinot noir!

Figure 6. *Which one is not like the other two?*

We developed a modified version of the triangular test and administered it to data SMEs. Each SME was presented with three patient journeys. In one configuration, exactly one journey was synthetic, mirroring the structure of the classic wine test. In other configurations, however, 0, 2, or all 3 journeys were synthetic—without disclosing to participants which scenario they were evaluating. Because evaluators vary in how strongly they feel about their judgments, we also captured confidence levels: SMEs recorded a score of 1 when their choice was made with full conviction and ½ when they were uncertain.

The result of the triangular tests is always the same. The SME's cannot reliably tell apart the synthetic patient journey from the RWD journey. This makes the point that the synthetic patient journey is clinically sound and realistic.

While we are not at liberty of sharing the results of the SME's, we can share the results of administering the same tests to LLM's. Below are the results from Open AI's ChatGPT 5.4 and Google Gemini 3.0.

ChatGPT 5.4

		Truth		
Open AI		RWD	Syn	Total
ChatGPT 5.4	RWD	8	7	15
	Syn	10	11	21
	Total	18	18	36

Accuracy	0.53
Precision	0.53
Recall	0.44
F1 score	0.48

Gemini 3.0

		Truth		
Google		RWD	Syn	Total
Gemini 3.0	RWD	5	6	11
	Syn	13	12	25
	Total	18	18	36

Accuracy	0.47
Precision	0.45
Recall	0.28
F1 score	0.34

Figure 7. *ChatGPT and Gemini Attempting to Spot Synthetic Data*

The results of our experiment show that all three LLMs struggle to reliably distinguish synthetic patient journeys from real ones. Among the four possible outcomes, the first two are especially informative:

- **Synthetic cases correctly identified as synthetic:** These cases highlight the reasoning patterns LLMs use to detect flaws. Their rationales can help strengthen SynGen by revealing subtle issues in the generated journeys.
- **RWD cases misclassified as synthetic:** When an LLM mistakes real data for synthetic, its explanations may expose underlying problems or anomalies in the RWD itself—insights that can guide dataquality improvements.
- **RWD cases correctly identified as real:** These outcomes do not offer much diagnostic value; they simply indicate that the LLM behaved as expected.
- **Synthetic cases mistaken for RWD:** These are success cases for SynGen, showing that the synthetic journeys are realistic enough to "fool" the model. As such, they reveal little about how to further improve the system.

7. 7 USE CASES

Below are the 7 use cases we invariably come across.

1. **Replication**
   A central challenge in working with realworld data (RWD) is the inherent liability associated with protected health information (PHI). Data users must be prevented from reidentifying actual patients under any circumstance. Synthetic data directly addresses this concern by enabling the creation of datasets that faithfully mimic the statistical and clinical properties of RWD without exposing PHI. Importantly, the underlying RWD used to generate the synthetic dataset is discarded once the process is complete, eliminating residual privacy risk.
2. **Rare Disease**
   Analyses in raredisease contexts frequently suffer from extremely small sample sizes, limiting statistical robustness and insight generation. By expanding the cohort with clinically accurate and genuinely novel synthetic patient journeys, these limitations can be mitigated. Although the RWD "seed" must still be sufficiently large to establish the clinical scaffolding, synthetic data remains highly valuable for filling gaps and stabilizing analyses—even in disease areas where the available RWD is relatively rich.
3. **Data Imperfections ("Beauty Cream")**
   Patientlevel claims data often exhibit structural omissions—for example, the absence of major integrated delivery networks such as Kaiser Permanente, the VA, or Florida Cancer Specialists. Furthermore, medical claims (CMS1500) frequently lag pharmacy claims (NCPDP) by several weeks. Synthetic data provides a mechanism to compensate for such imperfections by reconstructing missing portions of the dataset and reducing disparities in timeliness. This process can be enhanced by integrating auxiliary data sources that capture otherwise missing segments during synthetic generation.

4. **Incomplete Patient Journeys**
   RWD often fails to capture the full sequence of interactions a patient experiences, particularly when vendors lack visibility across all providers involved in a patient's care. In some cases, entire chains of events leading up to an observed claim are missing. Synthetic data can be used to fill these gaps, yielding coherent and clinically plausible trajectories. This approach significantly improves visibility into referral pathways, testing patterns, and hospitalization sequences.

5. **"WhatIf" Scenario Generation**
   Organizations frequently require data representing hypothetical future scenarios—such as forecasting adoption curves or simulating alternative care pathways—at a granularity that is simply unavailable. Synthetic data enables this by generating patient journeys derived from RWD that are consistent with a set of assumptions. As a result, data is no longer restricted to describing the past; it can be extended to model future periods (e.g., 2027–2029) under the traditional basecase, conservative, or aggressive scenarios.

6. **Data for Countries Outside the U.S.**
   Patient-level claims data is largely unavailable outside the United States, a limitation felt most acutely in markets such as Germany and Japan. Synthetic data provides a practical workaround by using U.S. RWD as a structural template to generate representative datasets for the target country. To ensure realism, the synthesized output must be anchored to observable constraints—such as aggregate brick-level sales or other geographically defined statistics—so that the resulting data aligns with known market conditions

7. **Indication Split for NonU.S. Markets**
   Vendors typically do not supply indicationlevel splits at the brick level for countries outside the U.S., leaving a major analytical gap. Synthetic data can address this need by generating patientlevel information consistent with U.S. RWD and then aggregating it to the brick level for the country of interest. Incorporating population demographics and providerlevel statistics strengthens the accuracy of geographic alignment and improves reliability of the resulting indication splits.

8. CONCLUSION

What to Expect with the Rise of Synthetic Data?

First, we will see a surge in demand for RWD for two reasons. First, RWD is the essential foundation for producing high-fidelity synthetic data; generating it from scratch is a non-starter, as it is virtually impossible to capture the "unspoken rules" of clinical reality through logic alone. Second, we must maintain RWD alongside synthetic sets to provide a "ground truth" reference. Whenever synthetic insights appear counterintuitive, having the underlying RWD allows us to validate findings and maintain credibility.

Second, there will be a greater emphasis on identifying and "fixing the warts" within RWD—specifically the gaping holes left by the likes of Kaiser Permanente and the VA, and the incompleteness stemming from the poor capture of Hospital activity ad Medicare FFS to mention just these two. This will drive interest in diverse supplementary sources—such as epidemiological data, shipment records, SDOH, affiliation and hierarchical data that connects accounts to individual hospitals, and hospital denominator data—to fill gaps and plug holes. Ultimately, we will see a shift toward using any available data source that helps provide the complete denominator where RWD only provides the numerator.

The next time you start off a data analysis project, resist the temptation of using your go-to data source. Take a pause, reflect, and consider deploying synthetic data instead. You may be in for a very pleasant surprise.

REFERENCES

1. **"Synthetic Data Generation in Healthcare."** *Nature* (2025).
2. Goodfellow, I., PougetAbadie, J., Mirza, M., Xu, B., WardeFarley, D., Ozair, S., Courville, A., & Bengio, Y. **"Generative Adversarial Networks."** *arXiv:1406.2661* (2014).
3. Kingma, D. P., & Welling, M. **"AutoEncoding Variational Bayes."** *arXiv:1312.6114* (2013).
4. Patki, N., Wedge, R., & Veeramachaneni, K. **"The Synthetic Data Vault (SDV)."** MIT Data to AI Lab (2016).
5. Foraker, R., Morrow, J. D., Johnson, J. A., Wilcox, A. B., Forster, A. J., & Payne, P. R. O. **"Understanding Synthetic Data: Artificial Datasets for RealWorld Evidence."** *BMJ EvidenceBased Medicine* (2025).
6. Gonzales, A., Guruswamy, G., & Smith, S. R. **"Synthetic Data in Health Care: A Narrative Review."** *PLOS Digital Health* (2023).
7. Rao, H., Liu, W., Wang, H., Huang, IC., He, Z., & Huang, X. **"A Scoping Review of Synthetic Data Generation for Biomedical Research and Applications."** *arXiv:2506.16594* (2025).
   A largescale review of 59 studies focusing on synthetic data across multiple biomedical modalities, emphasizing LLMbased generation.

ABOUT THE AUTHORS

Jean-Patrick Tsang is the Founder and President of Bayser, a Chicago-based consulting firm dedicated to sales and marketing for pharmaceutical companies. JP is an expert in data strategy and advanced analytics. JP has published 30+ papers, given 100+ talks at conferences, and completed 250+ projects. In a previous life, JP deployed Artificial Intelligence to automate the design of payloads for satellites. JP earned a Ph.D. in Artificial Intelligence from Grenoble University, advised 2 PhD students, and earned an MBA from INSEAD in Fontainebleau, France. He was the recipient of the 2015 PMSA Lifetime Achievement Award.

Badr Bouchabchoub is a Gen-AI and Data Engineer at Bayser. He specializes in developing advanced algorithms, synthetic data, and AI-driven solutions, including LLM-based applications. His work focuses on solving complex business questions. He is currently completing a Master's in Computer Science with a specialization in AI and Data at ESILV in Paris, France. His background combines machine learning, full-stack development, cloud computing, and software architecture.

SPRING 2026 | ALSO IN THIS ISSUE:

pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

www.pmsa.org