GILEAD

Agentic Cohort Builder

Gilead Clinical Data Science & EPAM Life Science

AGENDA

1. Executive Summary
2. Solution & Method (Agentic Cohort Builder)
3. What we built
4. Complex Cohort LLM & Human vs Human
5. Opportunities, Learnings and Next Steps



Our objective is not automated epidemiology. Our objective is **automated implementation of cohort specifications** so experts can focus on **phenotype validity, study design, and interpretation.** The goal is efficiency and compressing timelines.

Lee WC, Mankin B, Hood D. **A brave new world: the rise of agentic AI in rapid cycle RWE analytics**. *Expert Rev Pharmacoecon Outcomes Res*. 2026;26(6):739-746.

Executive Summary

Cohort building is the cornerstone of RWE evidence generation. It is also the most critical step to get right and traditionally the most time-consuming. The Agentic Cohort Builder automates and standardizes implementation while preserving expert oversight, enabling epidemiologists and analysts to focus on phenotype validity, study design, and interpretation.

In the therapeutic area pilots, the approach reduced development time from **55 to 6 hours**, with less than half a **percentage-point difference** from the human-built cohort. The path to scale is a governed human-agent model supported by rigorous QC, testing, validation, and auditability.

1 CONCEPT CREATION

Standardize code lists into evidence-linked OMOP concept sets

2 COHORT BUILDER

Translate eligibility and temporal logic into machine-readable cohorts

3 QC & VALIDATION

Compare protocol intent with executable definitions and flag gaps

4 OMOP COHORTS + DIAGNOSTICS

Produce execution-ready cohorts, attrition, characteristics, and metadata

Objective: **Not automated epidemiology. Automated implementation.** Experts focus on phenotype validity, study design, discrepancy resolution, and interpretation.

PILOT RESULTS

TIME TO DEVELOP

55 → 6 hours

~89% reduction in development time

ALIGNMENT TO HUMAN BENCHMARK

99.5% percentage-point

Concordance between final cohort retention

GOVERNED HUMAN–AGENT MODEL

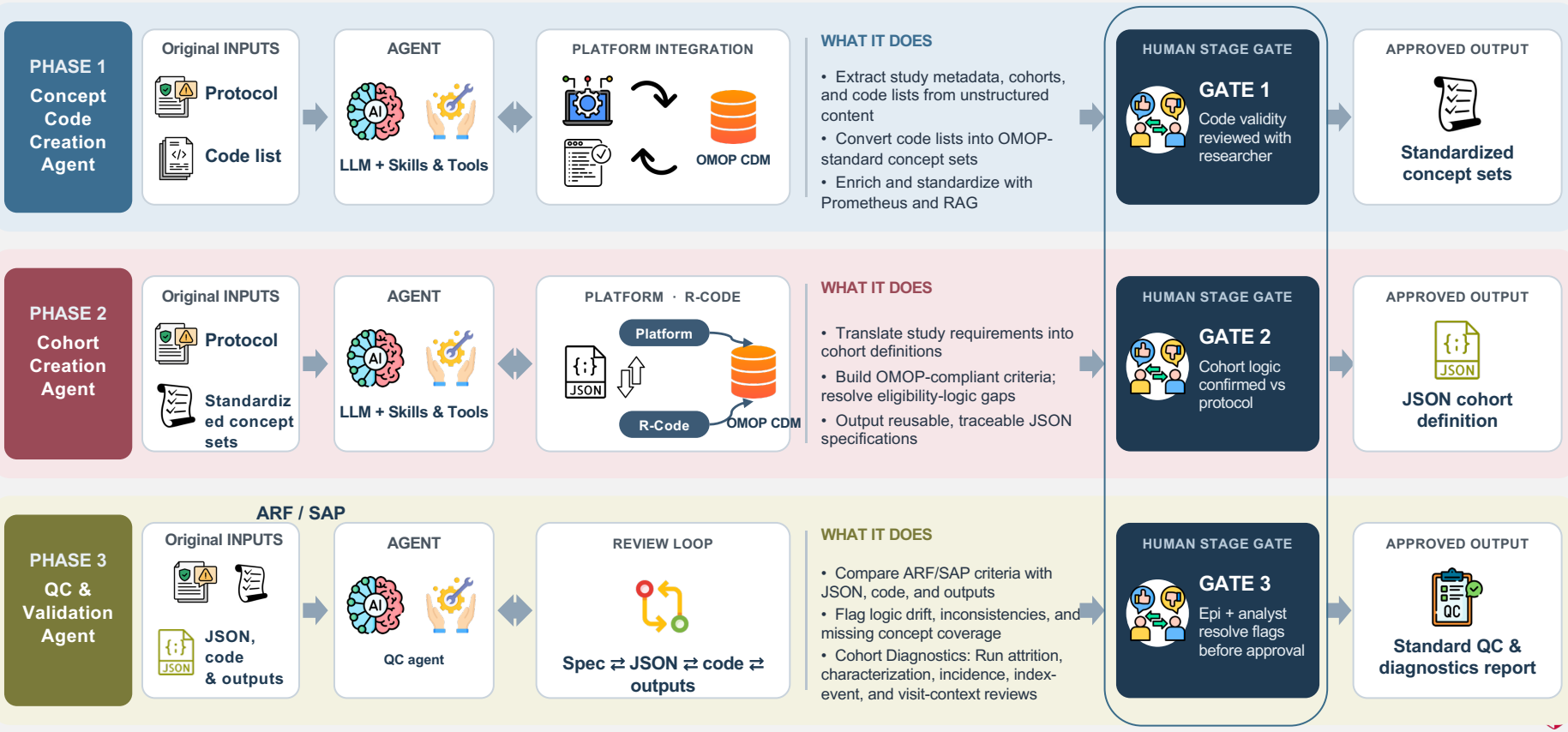
Strong first draft + Human stage gates + Rigorous QC

Scale through testing, validation, versioning, and complete auditability.

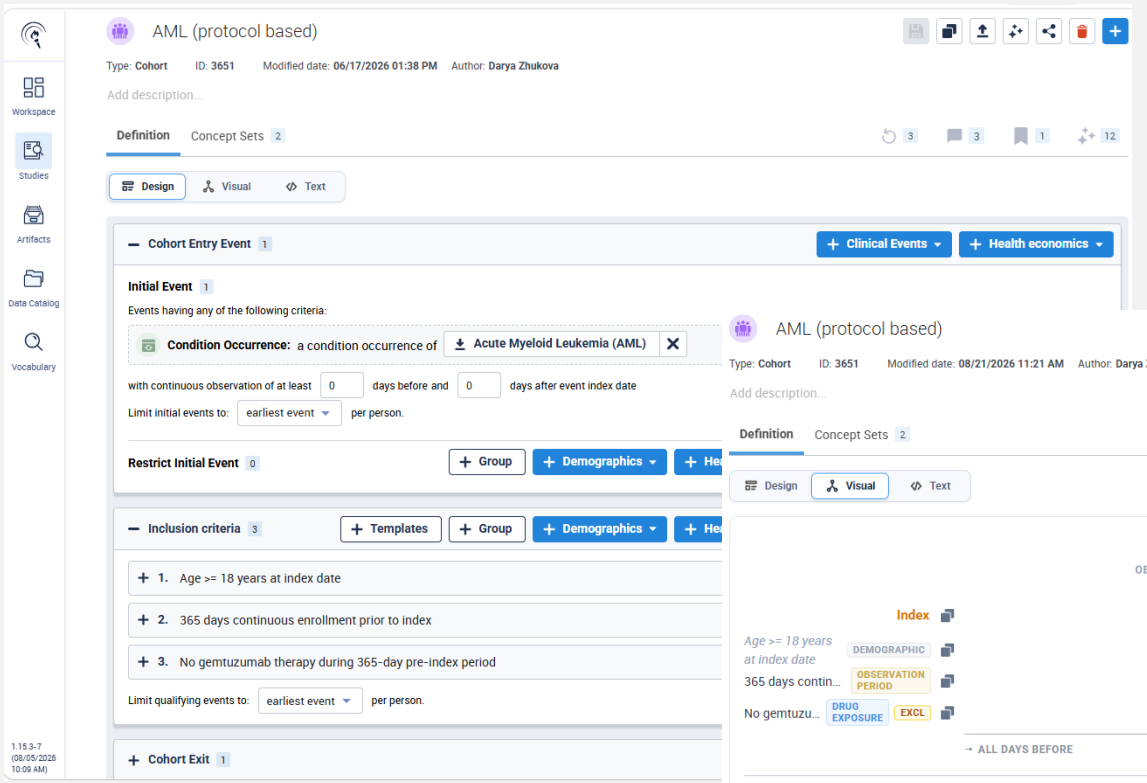
BOTTOM LINE: Accelerate cohort development without compromising the scientific rigor required for credible RWE.

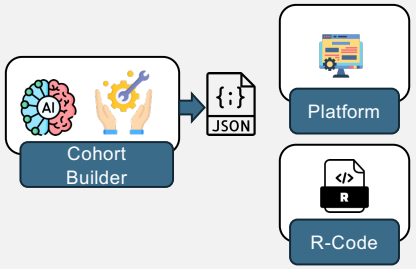
Solution: Agentic Cohort Builder

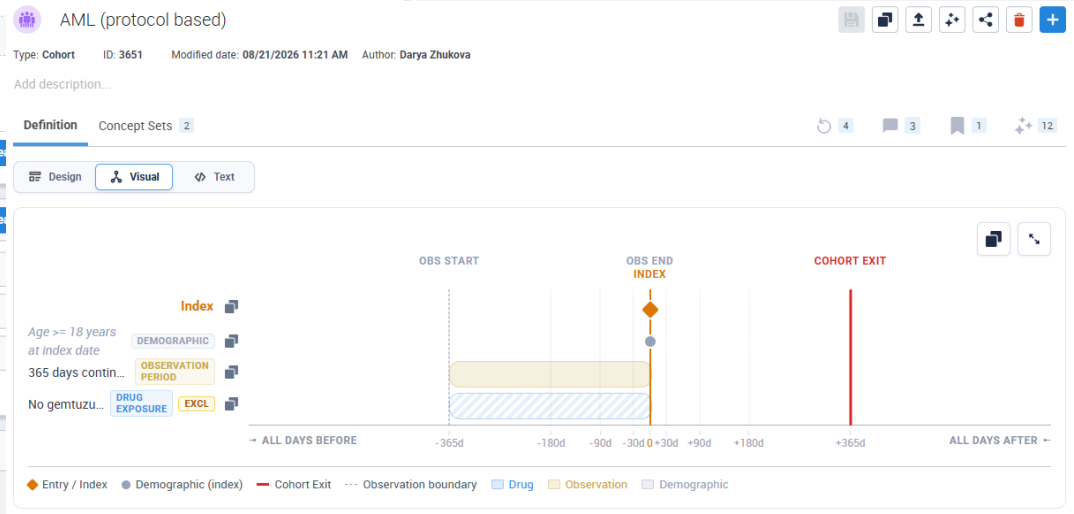
Three agent systems, one repeatable pattern — nothing becomes an output until a researcher approves it



Continued - Example of Outputs from JSON

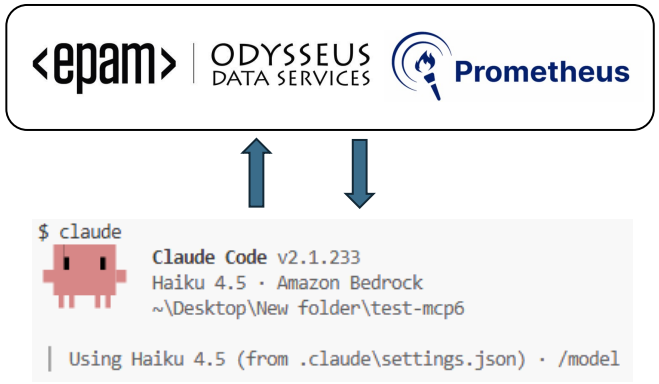






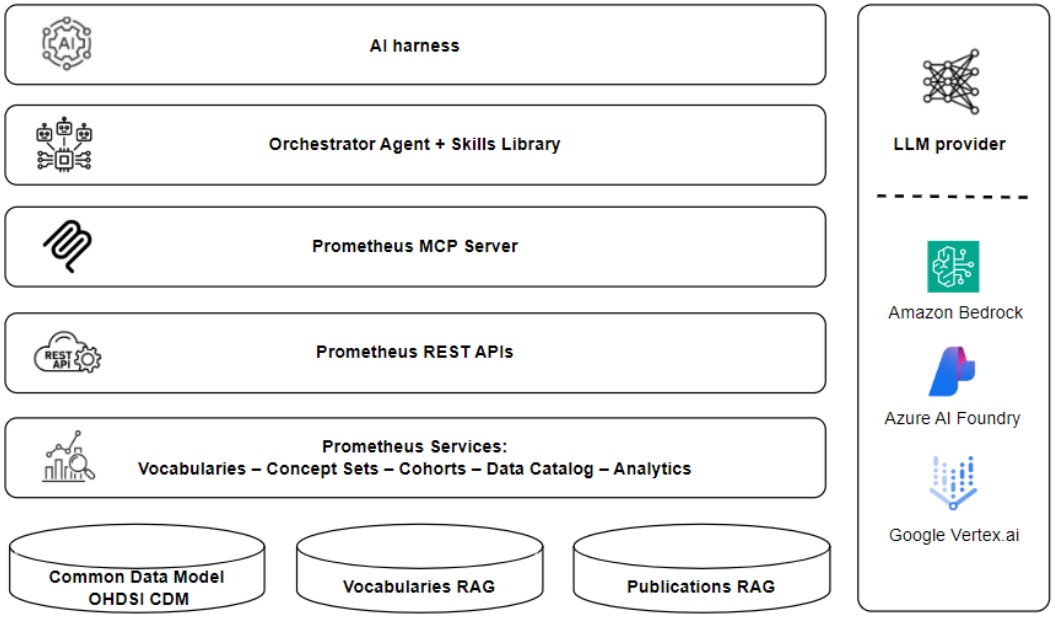
What We Built: Prometheus, Now Agent-Callable via MCP

- Prometheus MCP server exposing platform capabilities to AI agents
  (25+ callable tools/skills and growing)
- Anything available in the web UI becomes available via automatable agentic flows



cohort	generate_cohort	Draft an OMOP cohort expression from natural language using AI.
cohort	get_cohort_generation_result	Fetch the current status and result of a cohort generation job.
cohort	get_latest_cohort_executions	Get the latest execution per datasource for a persisted cohort.
cohort	list_cohort_executions	List all execution batches for a persisted cohort.
cohort	persist_cohort	Save a cohort expression to the Prometheus workspace.
cohort	search_cohorts	Search cohorts in the workspace or within a study.
conceptset	generate_concept_set	Generate a reusable OMOP concept set from a text description.
conceptset	get_concept_set_result	Re-fetch the status and result of a concept set generation job.
conceptset	search_conceptsets	Search concept sets in the workspace or within a study.
concepts	convert_concepts_to_standard	Convert non-standard OMOP concepts to their standard equivalents.
concepts	get_suggest_concepts_result	Fetch the current status and result of a concept suggestion job.
concepts	lookup_concept_record_counts	Look up record counts (RC/DRC/PC/DPC) for

What We built: Architecture



- **RAG-powered vocabulary and evidence grounding**
  Two dedicated Retrieval-Augmented Generation (RAG) vector stores – enable agents to ground concept lookups and suggestions in real OMOP vocabulary content and published medical literature, rather than relying on model memory alone – dramatically improving accuracy and traceability.

- **Vendor-agnostic AI harness**
  The AI harness acts as the CLI-level orchestration layer that coordinates agents and skills. It is fully vendor-agnostic. Today the demo runs on Claude Code, but the same harness pattern is supported by other AI harnesses such as Aider, OpenCode, CodeMie, OpenClaw, Codex, etc.

- **Agent + Skills Library layer**
  The Agent + Skills Library layer packages discrete workflow steps (extract, map, generate, suggest, merge, standardize, run analytics) as reusable, independently testable skills. New workflows can be composed from existing skills without re-engineering the platform layer beneath.

- **Platform-agnostic by design**
  The architecture decouples AI orchestration from the platform layer via the Prometheus MCP Server – a standardized interface over Prometheus REST APIs. Any external agent or orchestrator that speaks MCP can drive the workflow, regardless of the underlying technology stack.

- **Prometheus Services expose the full platform breadth**
  The REST API layer surfaces Vocabularies, Concept Sets, Cohorts, Data Catalog, and Analytics as callable services – meaning the same architecture that drives the demo workflow today can progressively automate a growing range of Prometheus capabilities with no architectural changes.

- **Pluggable LLM provider**
  No vendor lock by the LLM Providers. Amazon Bedrock, Azure AI Foundry, and Google Vertex AI – switch or upgrade models without changing the workflow.

What We Built: MCP server capabilities



- Authentication/Navigation
  - Authentication/SSO
  - Tenants
  - Studies/Workspaces
  - Data Sources



- Concept sets/Vocabularies
  - Concept lookup/search
  - AI Concept Set generation
  - Enrichment
  - Standardization
  - Metadata
  - Validation



- Cohorts
  - AI Generation
  - Attrition
  - Metadata
  - Validation



- Analytics
  - Counts with Demographics
  - Characterization
  - Incidence Rates
  - Pathways
  - Feasibility
  - Cost, Time-to-Event, etc.

Cohort attrition: human benchmark vs agentic builder

Patient retention at each sequential eligibility step

Both funnels start at N = 245,779



Final retention is closely aligned: 9.4% human benchmark vs 9.1% agentic builder (670 fewer patients; −0.3 percentage points).

Accuracy

<0.5% difference between 'Agentic Builder' vs. Human Benchmark

Significant Time Savings

89% reduction in time taken to create a validated cohort

QC Agent

When doing manual QC, the agentic system does a through job and creates a standard report

Bi-Directional Auditability

When doing manual QC, the agentic system does a through job and creates a standard report

ARF/SAP: Analytics Request Form / Statistical Analysis Plan
QC: Quality Control (check)

Opportunities, Learnings and Next Steps

1 TRUSTED FOUNDATION

WHAT THE FOUNDATION IS BUILT OF

Skills
Detailed, human-backed instructions with examples, guardrails, acceptance criteria, and repeatable tests for each agent skill.

Tools
Approved analytical tools integrated through explicit, version-controlled paths that define when and how agents can use each capability.

HOW WE KEEP IT TRUSTWORTHY

LLMOps / AgentOps
Versioned prompts, models, agents, tools, and knowledge sources, with complete decision traces and performance metrics.

Testing
Unit, regression, edge-case, performance, and cross-database tests for every agent, tool, and end-to-end cohort workflow.

Validation
Gold-standard cohorts, human-versus-agent benchmarks, automated diagnostics, and independent QA-agent review.

ENABLES SCALE

2 EXPANDED VALUE

BUILD & VALIDATE

Intelligent cohort development
Translate study requirements into standardized concept sets, executable OMOP cohort logic, and transparent inclusion and exclusion criteria.

Automated quality & validation
Compare narrative specifications with executable logic, identify discrepancies, assess attrition, and produce traceable outputs for expert review.

REUSE & SCALE

Reusable knowledge
A version-controlled library of validated phenotypes, standing cohorts, concept sets, benchmarks, and decisions.

Scalable execution
Consistent cohort definitions across therapeutic areas, data sources, and federated networks, with data-fitness assessment.

EXPAND THE ECOSYSTEM

Marketplace of tools and analysis
A centralized environment/platform to discover, reuse, and contribute cohort definitions, diagnostics, methods, and study-ready assets.

Expansion beyond cohorts
Feasibility assessment, population characterization, sensitivity analyses, automated documentation, and downstream evidence workflows.

HUMAN ACCOUNTABILITY: Epidemiologists and analysts remain responsible for phenotype validity, study design, interpretation, and approval at defined stage gates.