

CohortVisualizer: An OHDSI Workbench for Cohort Exploration and Hypothesis Generation

Christelle Xiong¹, BS, Christopher A. Mecoli^{1,2}, MD, MHS

1. Division of Biomedical Informatics and Data Science, Johns Hopkins School of Medicine
2. Division of Rheumatology, Department of Medicine, Johns Hopkins School of Medicine

Background

OHDSI studies commonly start with cohort creation in ATLAS, followed by cohort diagnostics. However, there is limited information on temporal trends and aggregate patient trajectories, especially in a visual format. CohortVisualizer addresses this interpretive gap. It is a structured bridge between cohort definition, data quality awareness, targeted clinical review, and hypothesis generation.

Methods

CohortVisualizer ingests ATLAS JSON cohort definitions, through CirceR-compatible logic, or already-instantiated cohort tables. It then extracts cohort populations and structured Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) events using read-only queries. The core representation is a patient-by-domain-by-time-window feature matrix anchored at cohort entry, which allows the same data structure to support overview heatmaps, dimensionality reduction, clustering, anomaly scoring, review prioritization, and temporal rule checks. This workbench represents the aggregate cohort and data-density layer, and secondly shows the subgroup interpretation layer, where clusters are summarized by prevalent conditions, drugs, demographics, follow-up, and outcome signals. It also represents the final translation layer, where exploratory cluster comparisons become candidate hypotheses and exportable review material. This layered design keeps the workflow auditable: users can move from a population-level pattern to subgroup summaries, then to specific review prompts, without treating exploratory signals as confirmatory evidence.

Results

The current implementation provides an R package and Shiny application for OMOP CDM version 5.3/5.4 environments, with a synthetic demonstration mode that exercises the workflow without requiring database credentials. The figures show the intended user journey. First, the first layer frames the cohort as an empirical object, showing data density and concepts over time (Figure 1). Second, using unsupervised modeling, it generates clusters from the cohort(s), enabling side-by-side prevalence patterns and summary characteristics that help reviewers ask why subgroups differ (Figure 2). Third, the module converts those differences into ranked, explicitly exploratory hypothesis candidates and a clinician review packet (Figure 3). Finally, the

outcome exploration layer allows users to examine event timing, outcome distribution across the existing cohort UMAP, and outcome-specific clustering to identify outcome-enriched regions or subgroups for further review (Figure 4). The tool also supports review queues, signal explanation cards, temporal concern flags, patient trajectory inspection, and report export. Together, these functions create a reproducible path from cohort-level orientation to targeted review and research-question refinement. Unit tests cover feature construction, clustering support functions, review sets, temporal flags, ranking, signal explanations, export behavior, and data connectors.

Conclusions

CohortVisualizer fills a practical gap between aggregate OHDSI diagnostics and formal chart review by organizing structured OMOP evidence into a guided, hypothesis-generating workflow. Its value is the combination of visual cohort orientation, machine-learning-assisted subgroup discovery, transparent review prioritization, and cautious hypothesis reporting. The figures emphasize that the project is designed for study planning and validation support, not clinical decision support or causal inference. Used alongside ATLAS, CohortDiagnostics, DataQualityDashboard, and downstream HADES methods, CohortVisualizer may help OHDSI teams refine phenotypes, identify data-sparse or outlying records, plan targeted clinical validation, and generate better-specified analyses. Future work should evaluate usability across sites, calibrate prioritization signals, and assess whether this workflow improves the efficiency and quality of OHDSI study design.



Figure 1. Cohort-level orientation view summarizing synthetic/demo cohort size, follow-up, demographics, and domain-level data density over calendar time. This aggregate view helps investigators detect cohort composition and source-data coverage issues before reviewing individual records.

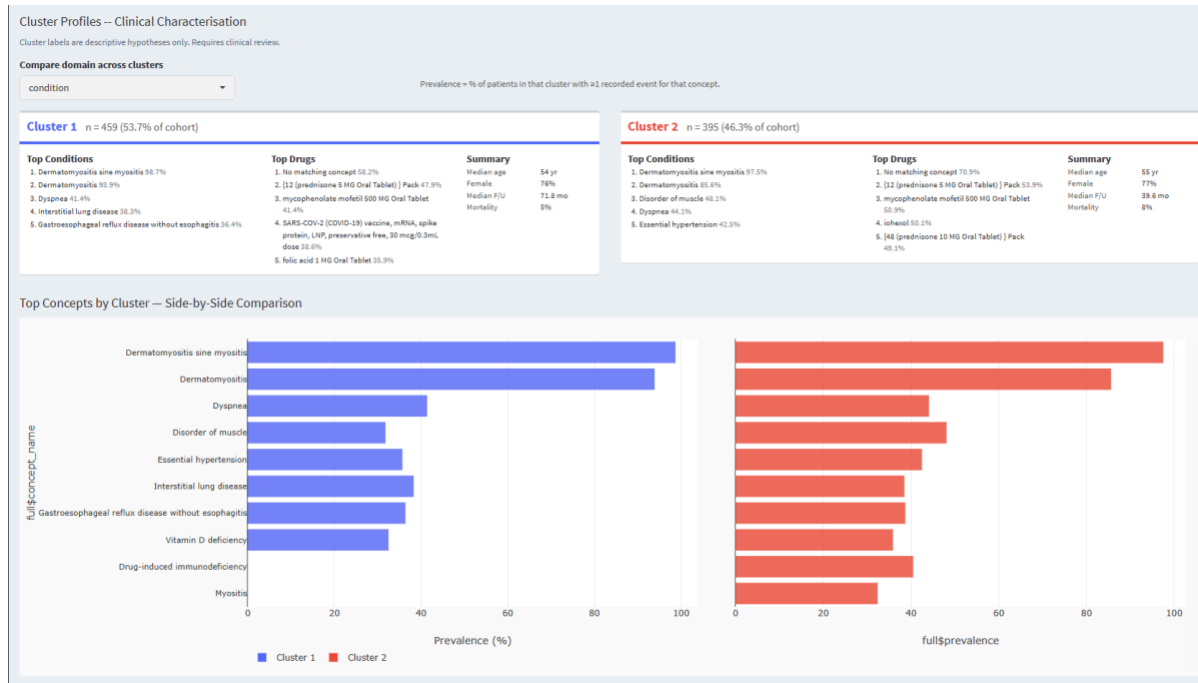


Figure 2. Cluster-profile comparison view showing cluster-level concept summaries and prevalence patterns. These descriptive profiles translate unsupervised groupings into clinically interpretable review prompts while avoiding confirmatory claims.

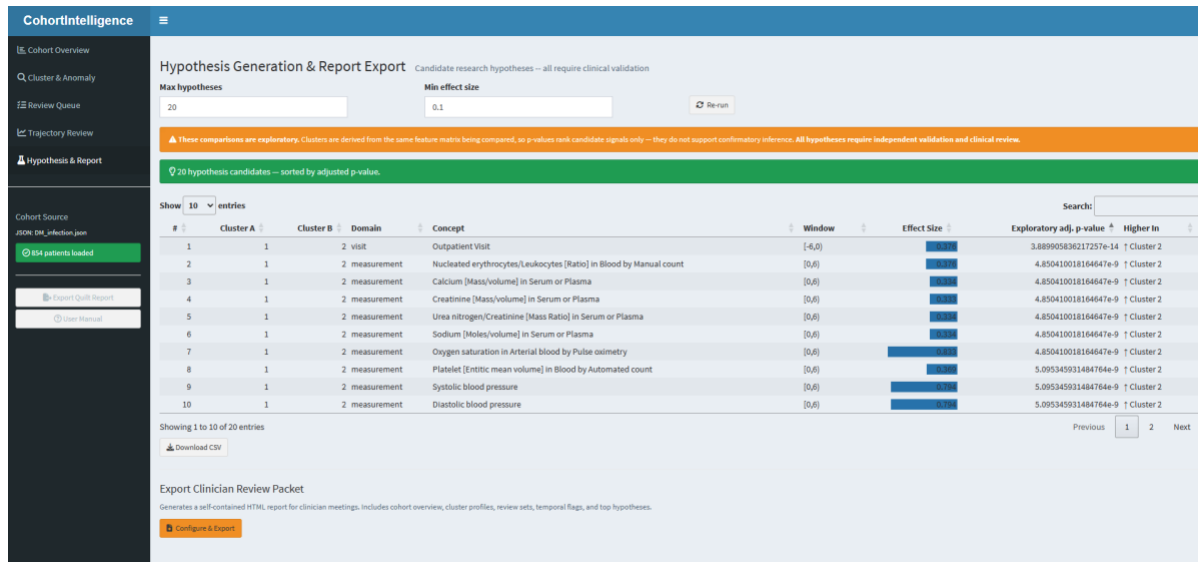


Figure 3. Hypothesis generation and report export view. Candidate signals are ranked for review, framed as exploratory findings, and packaged for clinician or methodologist discussion rather than treated as inferential results.

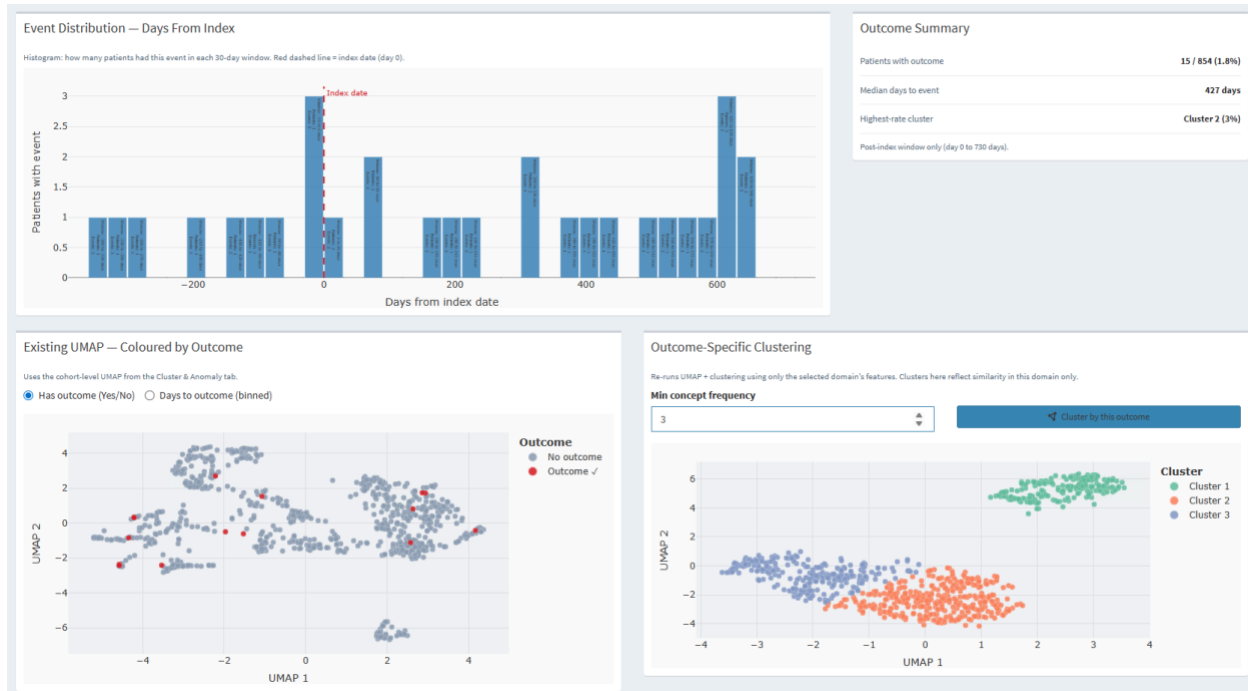


Figure 4. Outcome-centered exploration view linking event timing, cohort-level outcome distribution, UMAP localization, and outcome-specific clustering.