LLM-Assisted Analytic Code Development on OMOP

Lowering the Bar Without Lowering the Standard

Adam Johnson, MD, MPH

Vascular Surgeon: Duke University School of Medicine & Durham VA Medical Center

OHDSI Community Presentation: June 9, 2026

adam.johnson@duke.edu

About me

Roles

Practicing Vascular Surgeon

Director, Duke Vascular Informatics Lab

Medical Director, Durham VA Limb Preservation Program

Chair, SVS VQI FIT Program

Epic Physician Builder

Research Focus

Improve clinical decision-making and systems of care for patients with peripheral artery disease

Strengthen transportability and reproducibility of prognostic models in vascular surgery

Train future surgeon-data scientists

June 2026
adam.johnson@duke.edu

Disclosures

Funding

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number K12TR005435. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Federal Employment

I am an employee of the US Veterans Health Administration. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US government or the VA Health Administration.

AI-Assisted Development

This work used Claude Code (Anthropic) for LLM-assisted code generation. All outputs were constrained to OHDSI standards via institutional guardrails — a central topic of this presentation.

adam.johnson@duke.edu

OHDSI has an amazing community, but accessing it requires a certain technical expertise

What exists:

- Open-source data model and robust analytic framework
- Rigor, documentation, and data provenance are fundamental to these tools
- Vibrant and open community
- Interactive and engaging EHDEN academy and ODHSI tutorials

What barriers I have seen

- Accessing the tools directly requires a technical and informatics expertise many clinicians do not have
- Many institutions do not allow LLM direct access to data or even data definitions
- Even at institutions with robust OMOP frameworks, there is limited awareness and access

LLMs create a genuine opportunity — but unconstrained LLMs actively undermine OHDSI standards. The question is whether we can make the guardrailed path the default path.

June 2026
adam.johnson@duke.edu

One Docker Pull. Full HADES Stack. Correct Structure From Day One.

- R 4.5, Java 17, full HADES stack, SQL Server JDBC — pre-configured
- GitHub Template Repository — fork once, study begins with correct structure
- Numbered workflow steps 01–09 encode OHDSI best practices as the default path
- study_params.yaml forces explicit study design declaration before any model runs
- Standardized environment → analysis code deploys cleanly to air-gapped clinical sites via portable bundle

For a clinician-researcher without local OMOP and HADES expertise, the environment itself is the first collaborator.

June 2026
adam.johnson@duke.edu

```
omop-dev-workspace/
├── docker-compose.yml          # Shared SQL Server container (Azure SQL Edge)
├── .devcontainer/              # VS Code dev container definition
│   ├── devcontainer.json
│   ├── docker-compose.yml      # Dev container service + volume mounts
│   ├── Dockerfile              # R 4.5.2 + Java 17 + system dependencies
│   └── setup-claude-headless.sh  # Claude Code headless auth bootstrap
├── renv.lock                   # Workspace-level R package lockfile (full HADES + tidyverse st
├── renv/activate.R             # renv bootstrap — sourced by .Rprofile on container start
├── .Rprofile                   # Activates workspace renv so all /workspace Rscript calls find
├── .env.example                # Template for local secrets (SQL Server password)
│
├── studies.yaml                # Registry of all study and ETL repos in this workspace
├── contributors.yaml           # Contributor identity and CRediT roles (source for CONTRIBUTOR
├── datasets.yaml               # Site topology: CDM schemas, connection env refs (gitignored)
│
├── docs/                       # Workspace-level setup guides and reference docs
├── archive/                    # Retired docs and reference material (not actively maintained)
├── scripts/                    # Workspace-level utility scripts (e.g. sync_contributors.R)
├── infrastructure/             # Workspace-level setup and vocabulary loader scripts
│   ├── setup/
│   └── scripts/
├── phenotype_library/          # Shared verified concept sets (catalog.yaml + lookup scripts)
├── federated/                  # Multi-site federated analysis workflow and site coordination
│
├── omop_vocab/                 # OMOP vocabulary CSVs (not committed; Athena license)
├── presentations/              # Slide decks (not committed; generated locally)
│
├── synthea-omop-template/      # Study template — copy this for each new observational study (
```


adam.johnson@duke.edu

```
<your-study>/
  config.R                    ← single source of truth for all settings
  workflow/                   ← numbered step scripts (01–09)
  R/                          ← reusable infrastructure functions
  setup/                      ← renv + package install helpers
  scripts/                    ← ETL, Synthea runner, QC utilities
  cohorts/                    ← SQL cohort definitions (edit these)
  covariates/                 ← covariate CSV spec files (edit these)
  synthea/modules/            ← Synthea disease module + diagram
  portable/                   ← self-contained bundle for external sites
  internal_repo/              ← prebuilt OHDSI package binaries
  drivers/                    ← JDBC driver archive
  .github/                    ← Claude Code / AI assistant instructions
  output/                     ← analysis outputs (gitignored)
```


adam.johnson@duke.edu

Step	Script	Purpose
1	`workflow/01_setup_synthea_etl_qc_env.R`	Install packages, verify DB connectivity, provision JDBC driver
2	`workflow/02_define_omop_cohort_outcome_covariates.R`	**Validate your study definition** — cohort SQL, covariate CSVs, concept IDs
3	`workflow/03_generate_synthea_module_artifacts.R`	Validate Synthea disease module and regenerate HTML diagram
4	`workflow/04_generate_synthea_csv.ps1` / `.sh`	Generate Synthea synthetic patients (skip for real CDM data)
5	`workflow/05_etl_csv_to_omop.R`	ETL Synthea CSV → OMOP CDM (skip for real CDM data)
6	`workflow/06_quality_check_defined_phenotypes.R`	Post-ETL data quality checks
7	`workflow/07_setup_analysis_env.R`	Verify analysis packages are installed
8	`workflow/08_run_analysis_and_manuscript_report.R`	**Your analysis and outputs**
9	`workflow/09_build_portable_analysis_bundle.ps1` / `.sh`	Package bundle for deployment to external sites

adam.johnson@duke.edu

An example of setting up rules of engagement

Tier	Source	Value for Prognostic Modeling
1	OHDSI Phenotype Library	Start from peer-reviewed phenotypes — don't reinvent
2	Local cross-study catalog	Verified features reused across models — compounding rigor
3	Live vocabulary query	Confirmed against actual vocabulary build

- Every concept ID tagged [vocab query] — confirmed against live vocabulary
- AI-generated IDs tagged [pretraining] with mandatory warning — never silently trusted
- Inline comments on every hardcoded ID — any collaborator can audit the feature set

Concept ID governance doesn't just protect model validity — it makes the clinician's work legible to the OHDSI expert community.

adam.johnson@duke.edu

Encoding What a Knowledgeable Collaborator Would Tell You on Day One

CLAUDE.md enforces:

- PatientLevelPrediction and HADES-first — prohibition on ad-hoc modeling
- Three-tier concept ID lookup before any ID enters code
- [vocab query] / [pretraining] tagging — pretraining IDs blocked without warning
- study_params.yaml drives analysis flags — study design declared first
- OHDSI-style commenting — every feature and SQL join documented

Known limitations:

- Instruction following is convention, not code
- No automated concept ID cross-validation against live vocabulary
- No ATLAS-compatible output — direct SqlRender SQL

A clinician using this environment produces outputs an OMOP expert can review and extend — without a knowledge transfer meeting.

June 2026
adam.johnson@duke.edu

Claude Code as Lab Manager — Key Tasks

- **Index studies and contributors**
- **Repo sync & git workflow** — commit/push across all repos, open/merge PRs
- **Concept lookup & phenotype catalog** — query live vocabulary, maintain `catalog.yaml`
- **Cross-study alignment** — check study repos against shared template, propagate updates
- **Setup validation** — audit placeholder values, zero concept IDs, disabled analysis flags
- **Issue triage** — fetch and prioritize open GitHub issues across all studies
- **Synthea module editing** — modify JSON state machines and config for synthetic data
- **PowerPoint drafting** — generate slides from study outputs via `/pptx`
- **Container-aware execution** — all Rscript/SQL routes through `docker exec`

Claude Code wrangles many of the tedious tasks related to maintaining an ever-expanding code base

June 2026
adam.johnson@duke.edu

We Have a Working Answer. We're Not Sure It's the Right One.

On making this scale:

- What is the minimum viable collaboration model for a clinician-researcher at a site with no local OMOP expertise?
- Is there a role for OHDSI in maintaining a shared CLAUDE.md-style instruction set as a community standard?

On guardrail robustness:

- Is instruction-following a sufficient enforcement mechanism for reproducible research?
- Is automated concept ID linting already a part of HADES tooling?

On community interoperability:

- Should LLM-assisted cohort development target ATLAS-compatible JSON rather than direct SqlRender SQL?
- Is there an established OHDSI workflow for batch vocabulary gap submissions from specialty registry ETLs?

On responsible use:

- Has the community developed guidance on use of proprietary registry data dictionaries with third-party AI services?

June 2026
adam.johnson@duke.edu

Next Steps and Where We Want Collaborators

Immediate next steps:

- Feedback on dev container and analysis templates to ensure they truly align with OHDSI standards
- Contributors for additional analysis templates and use cases for the dev container

- Could this dev container help onboard collaborators in federated studies with limited local data expertise or availability?

adam.johnson@duke.edu
Github: adam-mdmph