

PhenoEx: An interactive phenotype builder

Authors: Kevin Crouse^{a,c}; Emily Clark, MPH^{a,b}

Institutions:

- a. National Center for Advancing Translational Science (NCATS), National Institutes of Health (NIH)
- b. Axle Informatics
- c. Nextonic Solutions

Background

The Observational Health Data Sciences and Informatics (OHDSI) research community¹ supports reproducible, collaborative health research, in part by the management of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)². OHDSI has overseen the development of open-source applications, such as ATLAS, which visualize concept relationships. However, there are few researcher-facing tools that allow users to input complex phenotype definitions in a graphical user interface and apply those definitions to their clinical data in a consistent, repeatable way.

Our team at NCATS has been developing *PhenoEx* (Phenotype Explorer) for the researchers involved in the National Clinical Cohort Collaborative (N3C)^{3,4}, for whom NCATS is the steward. N3C is a national consortium that provides secure access to de-identified, real-world data with the goal of increasing the speed of medical findings through collaborative engagement⁵. N3C has adopted the OMOP-CDM as its data model, and the NCATS team has developed several tools to support them. These tools include Set Architect (a Concept Set management tool) and ORBIT (an authentication, governance, and user provisioning tool). PhenoEx is a tool to define complex phenotypes and provision the subsequent patient cohorts within secure, trusted research environments.

Methods

PhenoEx provides both a web-based graphical user interface to build phenotype definitions and a backend API to validate the definition and run it against a clinical data source.

We designed the frontend UI with input from both subject matter experts and researchers who are not programmers or data scientists so that, while providing the building blocks to design complicated phenotype definitions, it is also accessible to a general user. Phenotype definitions are saved in a structured, JSON format and can be shared with teammates or applied to different CDM-based clinical datasets in different *runs*. Running a phenotype will generate the cohort from the clinical dataset. Summary statistics disaggregated by demographics and domain-specific metrics are also reported alongside a feasibility analysis that allows users to consider whether the resulting cohort shows selection skew or disproportionalities that may threaten future work.

The frontend is just the intuitive interface to the backend, which runs a Python FastAPI⁶ server to validate the phenotype definitions and run them against the clinical data. Data scientists on the

team may integrate their programmatic tools directly with the API through R or a Python Jupyter Notebook. Larger projects with in-house programs can also connect to the backend API; for example, teams that use Airflow to automate data pipelines could automate the regeneration of a phenotype cohort after clinical data is updated.

Results

PhenoEx is being developed directly for N3C, which has over 90 participating institutions contributing data that become deidentified, Limited Data Sets (LDS) and are stored in the secure enclave for researchers. In PhenoEx, the synthetic datasets that are used for development have up to 400GB of clinical data, allowing us to model performance for reasonably large data loads.

The frontend UI balances simplicity with the need to represent complex relationships between clinical concepts. The user adds a clinical domain entity to the canvas and connects it to either a target Concept or Concept Set (see Figure 1). Separate domain entities are connected through nested, boolean groupings (AND, OR, XOR, with the option for a negate/NOT modifier).

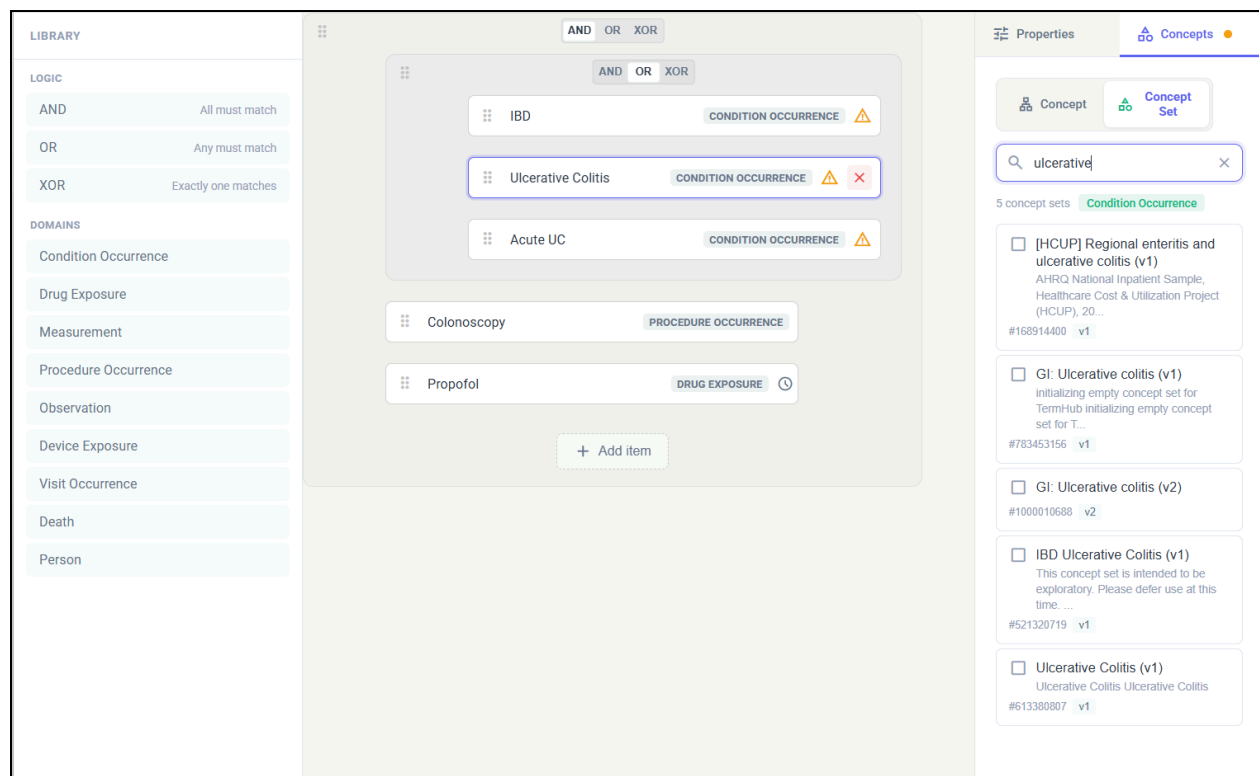


Figure 1. Frontend UI, demonstrating Concept selection for a new Condition Occurrence.

Constraints are applied to the attributes of each domain entity, such as the `start_date` of a `drug_exposure` or the `gender` of a `person`. Constraints can be absolute (`start_date > 2020-03-16`) or relative to other entities included in the model. Figure 2 shows the user setting a relative constraint in which the date of drug exposure to Propofol happens during the Colonoscopy procedure.

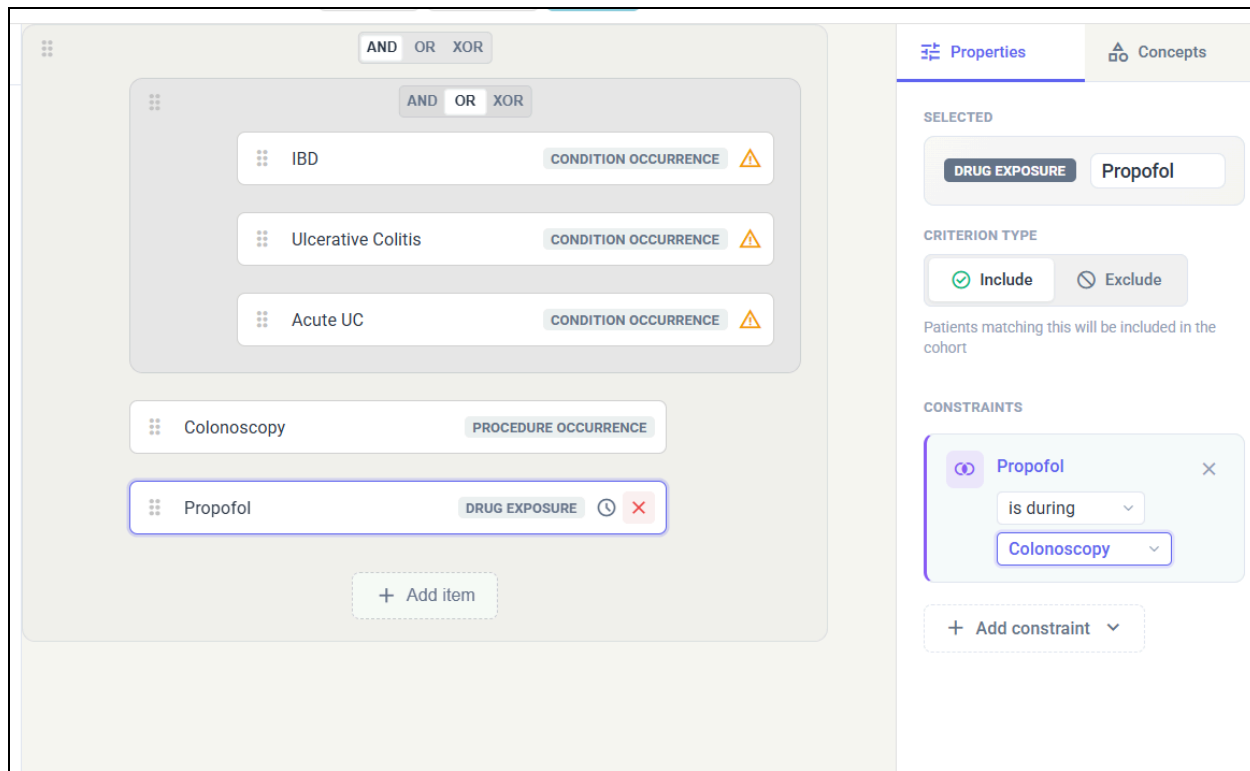


Figure 2. Frontend UI demonstrating the constraint of a specific drug exposure during a procedure.

PhenoEx also provides an Advanced mode that allows users to craft custom logic (in SQL at present) when they have more specialized requirements than the standard operators and constraint options. In the Advanced mode, the generated SQL clause is provided as a template, and the user can edit it to meet their specific need.

The backend API provides the powerhouse for PhenoEx’s operations and validates definitions, executes runs, and builds summary reports. Our current development connects to both Postgres and Palantir for data storage, which has led to a general interface that could easily extend to other CDM-compliant databases, data warehouses, and novel data stores. Common RDBMSes with python sqlalchemy support may already be supported.

PhenoEx generates a phenotype validation schema, which is an OpenAPI JSON schema⁷ that provides type definitions, operator constraints, and linkages between clinical domains and attributes. This schema governs the structure of the phenotype definitions. Figure 3 shows an example phenotype definition for women susceptible to thalidomide-caused birth defects (as in the landmark pharmacological catastrophe^{8,9}). PhenoEx uses its OpenAPI JSON schema to validate the definition, which quickly identifies the locations of any errors and allows users to fix issues before attempting any data pulls.

PhenoEx has an initial production release to the broader N3C community at the end of June. Data scientists can use standard REST libraries (like *requests* in python and *httr2* in R) to

interact directly with the backend. Depending on user feedback, we will add packages to interact more natively with the API and translate output data into pandas or R dataframes.

Conclusion

PhenoEx provides a solution to the need for complex phenotype definition and cohort identification to be transparent, reproducible, and generalizable. The backend API provides a standalone server that data scientists can integrate into their workflows and that can validate phenotypes, execute runs, and retrieve the cohorts identified. The UI provides an intuitive canvas by which researchers, subject matter experts, and other non-programmers can design complex phenotypes and share them with their team. In all, the PhenoEx application allows a new level of engagement with the clinical tables in the OMOP CDM that are intuitive, extensible, and generalizable.

```

{
  "phenotype": {
    "constraints": [
      {"field": "gender", "concept_id": 8532}, // Female
    ],
    "links": [
      {
        "domain": "Condition Occurrence",
        "id": "pregnancy",
        "concept_set_id": 326616819, //Pregnancy
      },
      {
        "domain": "Condition Occurrence",
        "concept_set_id": 726616819, // Morning Sickness(es)
        "constraints": [
          {
            "type": "relationship",
            "column": "condition_start_date",
            "operator": ">=",
            "target_id": "pregnancy",
            "target_column": "condition_start_date"
          },
          {
            "type": "relationship",
            "column": "condition_start_date",
            "operator": "<",
            "target_id": "thalidomide_exposure",
            "target_column": "drug_exposure_start_date"
          }
        ]
      }
    ],
    {
      "domain": "Drug Exposure",
      "id": "thalidomide_exposure",
      "concept_id": 4301375, // Thalidomide
      "constraints": [
        {
          "type": "relationship",
          "column": "drug_exposure_start_date",
          "operator": ">=",
          "modifier": "21 days",
          "target_id": "pregnancy",
          "target_column": "condition_start_date"
        },
        {
          "type": "relationship",
          "column": "drug_exposure_start_date",
          "operator": "<=",
          "modifier": "56 days",
          "target_id": "pregnancy",
          "target_column": "condition_start_date"
        }
      ]
    }
  ]
}

```

Figure 3. Phenotype definition approximating thalidomide-associated birth defect criteria

References

1. Observational Health Data Sciences and Informatics (OHDSI). Who we are [Internet]. New York (NY): Columbia University; 2014 [cited 2026 Jun 5]. Available from: <https://www.ohdsi.org/who-we-are/>
2. Observational Health Data Sciences and Informatics (OHDSI). OMOP Common Data Model [Internet]. New York (NY): Columbia University; 2026 [cited 2026 Jun 5]. Available from: <https://ohdsi.github.io/CommonDataModel/>
3. National Center for Advancing Translational Sciences. National Clinical Cohort Collaborative (N3C): overview [Internet]. Bethesda (MD): National Institutes of Health; 2026 [cited 2026 Jun 5]. Available from: <https://ncats.nih.gov/research/research-activities/n3c/overview>
4. National Center for Advancing Translational Sciences. National Clinical Cohort Collaborative (N3C): translating health data into discovery and knowledge [Internet]. Bethesda (MD): National Institutes of Health; 2025 Jun [cited 2026 Jun 5]. Available from: <https://ncats.nih.gov/sites/default/files/2025-08/NCATS-N3C-Factsheet-August-2025-508.pdf>
5. CNN. Coronavirus pandemic: updates from around the world, June 15, 2020 [Internet]. Atlanta (GA): Cable News Network; 2020 Jun 15 [cited 2026 Jun 5]. Available from: <https://www.cnn.com/world/live-news/coronavirus-pandemic-06-15-20-intl>
6. Ramírez S. FastAPI [Internet]. 2026 [cited 2026 Jun 5]. Available from: <https://fastapi.tiangolo.com/>
7. OpenAPI Initiative. OpenAPI Specification [Internet]. San Francisco (CA): Linux Foundation; 2026 [cited 2026 Jun 5]. Available from: <https://spec.openapis.org/oas/>
8. Mellin GW, Katzenstein M. The saga of thalidomide. *N Engl J Med*. 1962;267(23):1184-93. doi: 10.1056/NEJM196212062672306.
9. Lenz W. A short history of thalidomide embryopathy. *Teratology*. 1988;38(3):203-15. doi: 10.1002/tera.1420380303.