Compositional Public Health Approaches to Observational Health Research

Jacob S. Zelko
JuliaHealth

Nathaniel Osgood
University of Saskatchewan

1 Background

Given complex and evolving public health threats, the need for robust analysis pipelines that can rapidly adapt to new conditions is paramount. However, harmonizing and leveraging disparate data sources in these scenarios (such as electronic health records, census survey microdata, and environmental data) can be labor-intensive and error-prone. As a potential framework, research methods rooted in compositional public health may provide the flexibility and rigor needed to efficiently and effectively address such situations.

Compositional public health is an emerging research field seeking to address the complexity of public health responses. [6, 1] The field lies at the intersection of category theory, epidemiology, systems science, and engineering, and utilizes tools from applied category theory for public health applications. Core to compositional public health is category theory, a branch of pure mathematics whose area of study can be encapsulated as the slogan, "how things relate to things". This intentional vagueness allows tremendous flexibility while studying mathematical structures, while approaching those structures with a common set of mechanisms that allows transporting mathematical constructions, techniques and findings from one area to another. Applied category theory then expands the study of such structures from strictly pure mathematics to other domains such as dynamical systems, systems engineering, and metaphysics models. [4, 9]

Using approaches within compositional public health, we present preliminary work on how one can interpret the OMOP CDM as mathematical structures, translate them into computational data structures called ACSets (i.e. attributed **C**-Sets), and how to perform versatile analyses on top of this framing. As a proof of concept example, we demonstrate how such a framework could be used in the analysis of climate-impacted diseases, such as heart attacks or stroke, by integrating patient data with environmental and census data.

2 Methods

Within category theory, the fundamental mathematical structure is the category. By referencing our earlier slogan, a category comes equipped with two pieces of data: objects (the "things") and morphisms (the "relations"). We define a specific category called the category, **Schema**, where the objects are tables of a database schema and morphisms are the foreign key relationships that exist between tables in a database schema [11]. Using this language, we can interpret an OMOP CDM schema as an instance of this category. While the OMOP CDM has now been matched to a mathematical object, further steps are needed to link it to data and to compute upon it.

To put this category to use, we leveraged methods from AlgebraicJulia, an open-source software ecosystem for applied category theory written in Julia. [7] The central data structure within AlgebraicJulia is the ACSet, which builds on top of categories and additionally captures column information for tables (referred to as attributes). [8] Using the package, ACSets.jl, we can express the OMOP CDM as an ACSet to allow for representation of data and computation while maintaining the mathematical integrity of the OMOP CDM as a **Schema** category. 1

While ACSets provide a computational and mathematical foundation for understanding the OMOP CDM, it is restricted by memory constraints of the client machine due to operating in-memory. Additionally, there has historically been limited support to harmonize one ACSet with ACSets representing other data sources and handling large schemas such as the OMOP CDM. To remedy this, we explore the use of a "data fabric" developed within AlgebraicJulia. [10] While borrowing from other industry language, a data fabric in this case is a diagram of ACSets, whose nodes are data sources and edges are foreign key constraints between data sources, and which supports database reflection and virtualization (see for example 2). The data fabric supports the following approaches:

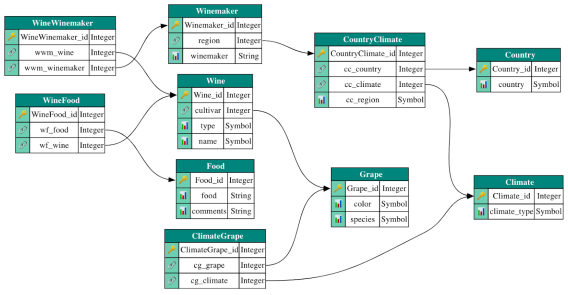
1. Quickly specify large database schemas.

```
@present SchOMOPCDM(FreeSchema) begin
  label::AttrType
  numerical::AttrType

  # Objects
  concept::Ob
  person::Ob

  # Homomorphisms
  gender_concept_id::Hom(person, concept)
  gender_source_concept_id::Hom(person, concept)
  race_concept_id::Hom(person, concept)
  race_source_concept_id::Hom(person, concept)
  ethnicity_concept_id::Hom(person, concept)
  ethnicity_source_concept_id::Hom(person,
concept)
  # concept Attributes
  concept_id::Attr(concept, numerical)
  concept_name::Attr(concept, label)
  standard_concept::Attr(concept, label)
  concept_code::Attr(concept, label)
  valid_start_date::Attr(concept, label)
  valid_end_date::Attr(concept, label)
  invalid_reason::Attr(concept, label)

  # person Attributes
  person_id::Attr(person, numerical)
  person_source_value::Attr(person, label)
  gender_source_value::Attr(person, label)
  year_of_birth::Attr(person, numerical)
  month_of_birth::Attr(person, numerical)
  day_of_birth::Attr(person, numerical)
  birth_datetime::Attr(person, numerical)
  race_source_value::Attr(person, label)
  ethnicity_source_value::Attr(person, label)
end
```

Figure 1: Subset of the OMOP CDM as an ACSet. This contains 2 objects (the `person` and `concept` tables), 6 morphisms (the foreign key relations between the tables), and 9 attributes (the columns that are not key relations across tables).

2. Chunk large schemas into smaller schemas populated by differing data sources.

3. Query schemas and manipulate them through a common access protocol.



Identifier	Name	PrimaryKey	Type	Fields	Indegree	Outdegree
1	Climate	Climate_id	InMemory	2	1	0
2	Grape	Grape_id	InMemory	3	1	0
3	ClimateGrape	ClimateGrape_id	InMemory	3	0	2
4	Country	Country_id	InMemory	2	1	0
5	CountryClimate	CountryClimate_id	InMemory	4	1	1
6	Wine	Wine_id	InMemory	4	1	0
7	Winemaker	Winemaker_id	DBSource{DB}	3	1	1
8	WineWinemaker	WineWinemaker_id	InMemory	3	0	2
9	Food	Food_id	InMemory	3	1	0
10	WineFood	WineFood_id	InMemory	3	0	1

Figure 2: Data fabric example. This simplified example shows how a data fabric can be created from a database schema. On the left is a database schema representing the relationships between wineries, wines, and other information. The blocks represent tables, the arrows to key emojis represent foreign key constraints, and the chart emoji represents other columns within these respective tables. By using the ACSet interface, we can load data from this schema regardless of where the data is housed. In this case, the `Winemaker` table is loaded from a SQLite database and the other tables are housed in-memory.

3 Results

For an initial proof of concept of this work, we loaded the Eunomia database into an ACSet presentation alongside publicly available data from the IPUMS Current Population Survey and NCEI weather information. By using the common interface afforded by ACSets, we can then create queries in the language of ACSets to simultaneously retrieve all patient data harmonized against these other datasets (see the code in for a simple example 1). As a result, we can reduce the work needed to manually harmonize and query multiple data sources together by grounding the approach in applied category theoretic methods.

```
person_query = @relation (Person=p_id, PersonDOB=p_dob, Visit=v_id, Condition=c_id) begin
    Person(_id=p_id, dob=p_dob)
    VisitOccurrence(_id=v_id, visit_person=p_id)
    ConditionOccurrence(_id=c_id, condition_visit=v_id, condition_person=p_id)
end

# Get all patient related information
query(omop_acs, person_query)

# Get patient 146's information
query(omop_acs, person_query, (p_id=146,))
```

Listing 1: Query patient information using a conjunctive query. The `person_query` defines a small query object that can pull patient information, condition information, and visit information for a specific patient from Eunomia. Then, we can query information for all patients or individual patients as needed using the ACSet interface.

4 Conclusions

As a field, compositional public health is rapidly emerging and evolving to adapt to the myriad uses that can exist across public health settings. Early work within compositional public health suggests great promise across various applications such as in System Dynamics modeling [3], agent-based modeling [2], and dynamical systems modeling [5]. The field of observational health research is no exception. In the preliminary work demonstrated here, we show how such methods can be applied in the context of observational health research concerning the OMOP CDM and other data sources.

As we continue exploring how to bridge the world of pure mathematics to applications across public health, we have a number of research directions planned based on these efforts. For example, we are actively exploring how to formalize phenotype definitions within category theoretic language, conduct an observational health study using data fabrics, and couching traditional data science methodology within categorical approaches. Finally, future work here will encompass addressing our long-term vision of using these formalisms of analyses in mathematical language to reduce the time cost of understanding and repurposing old analyses, so as to support more robust and quicker insight generation within observational health research.

References

[1] "An Introduction To Compositional Public Health". In: *New York City Category Theory Seminar* (Feb. 2025). Available at: https://www.youtube.com/watch?v=60GHdzetEqI.

[2] JOHN C BAEZ et al. "SOFTWARE FOR COMPOSITIONAL MODELING IN EPIDEMIOLOGY". In: ().

[3] John C Baez et al. "A Categorical Framework for Modeling with Stock and Flow Diagrams". In: *Mathematics of Public Health: Mathematical Modelling from the Next Generation*. Springer, 2023, pp. 175–207.

[4] Yujun Huang, Marius Furter, and Gioele Zardini. "On Composable and Parametric Uncertainty in Systems Co-Design". In: *arXiv preprint arXiv:2504.02766* (2025).

[5] Sophie Libkind et al. "An algebraic framework for structured epidemic modelling". In: *Philosophical Transactions of the Royal Society A* 380.2233 (2022), p. 20210309.

[6] Nathaniel Osgood, Jacob S. Zelko, and Nastaran Jamali. "Applied Category Theory for Public Health". In: *Transactions in Category Theory*. 2025. URL: https://jademaster.xyz/TACT25.html.

[7] Evan Patterson. *AlgebraicJulia: A Compositional Approach to Technical Computing*. 2022. URL: https://www.algebraicjulia.org.

[8] Evan Patterson, Owen Lynch, and Julia Fairbanks. "Categorical Data Structures for Technical Computing". In: *Compositionality* 4 (2022). DOI: 10.32408/compositionality-4-2. URL: https://www.compositionality-journal.org/articles/10.32408/compositionality-4-2.

[9] Evan Patterson et al. "A diagrammatic view of differential equations in physics". In: *arXiv preprint arXiv:2204.01843* (2022).

[10] Jacob S. Zelko, Matt Cuffaro, and Wu L. Sean. "ACT-Informed Data Science: A Case Study in Public Health Research". In: *8th International Conference on Applied Category Theory (ACT)*. 2025. URL: https://gataslab.org/act2025/act2025.

[11] Patrick Schultz et al. "Algebraic Databases". arXiv preprint arXiv:1602.03501. 2016. arXiv: 1602.03501. URL: https://arxiv.org/abs/1602.03501.