Derived Study Elements

Building an Analytic Dataset in N3C

N3C LOGIC LIAISONS:
Johanna Loomba, ME
Andrea Zhou, ME

Learning Objectives

What does it mean to build an analytic dataset for observational research?

What are important considerations I should take into account?

What are some available N3C tools that can help me?

Background

Observational Research: **An N3C Study Workflow Perspective**

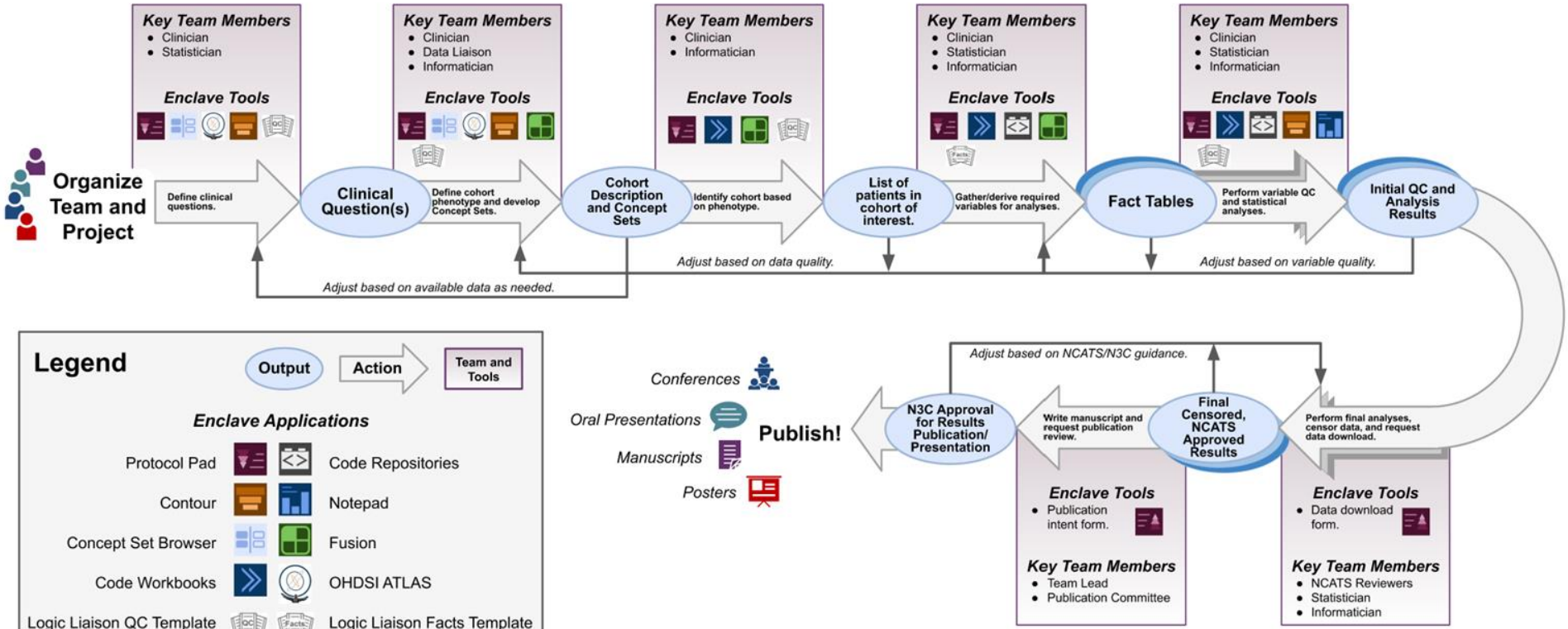
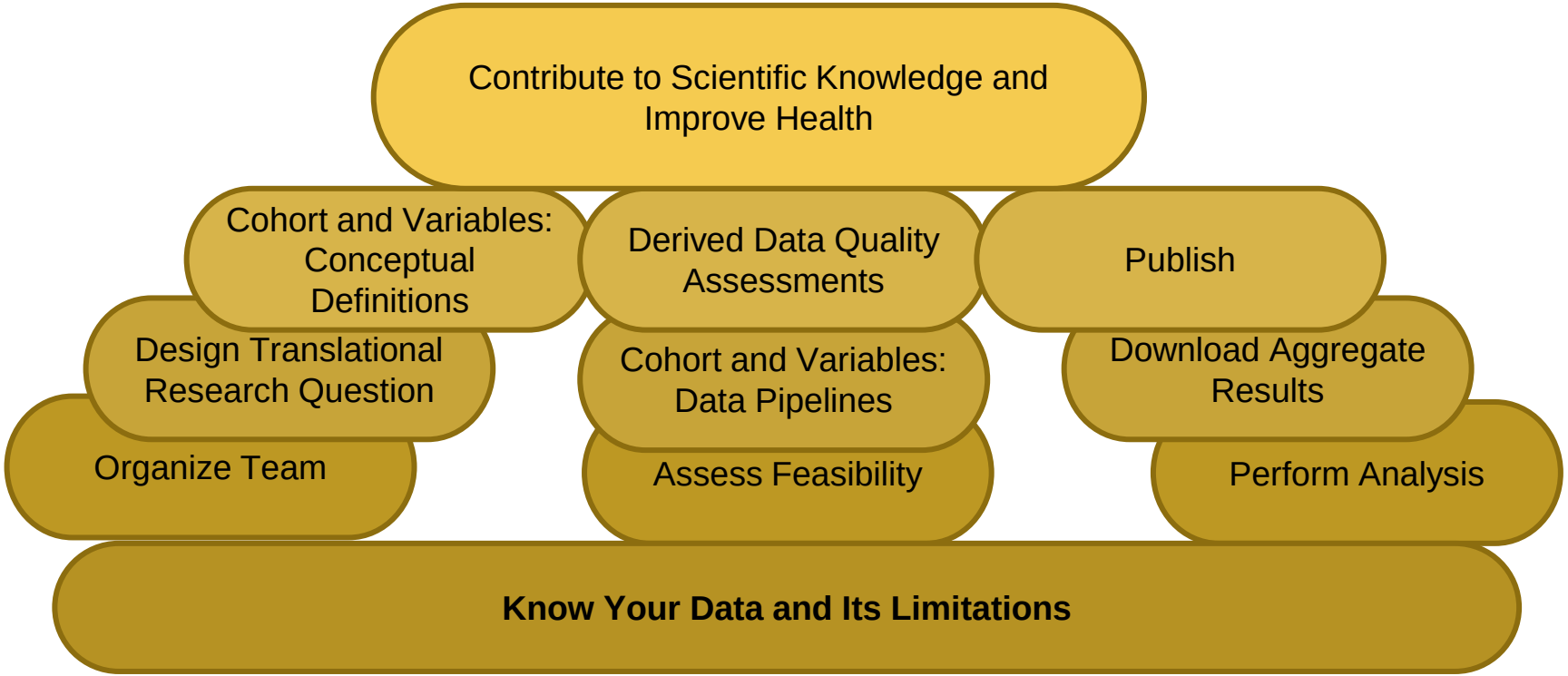
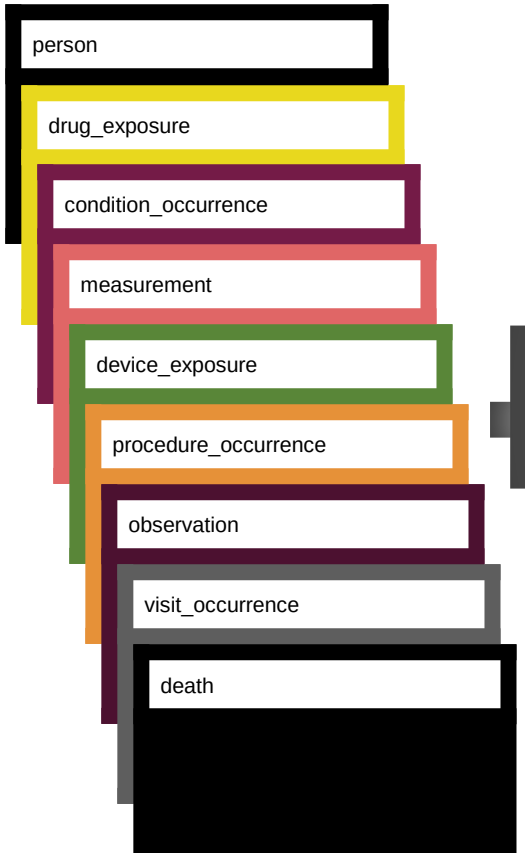


Image Source: The Researchers Guide to N3C https://national-covid-cohort-collaborative.github.io/guide-to-n3c-v1/chapters/tools.html

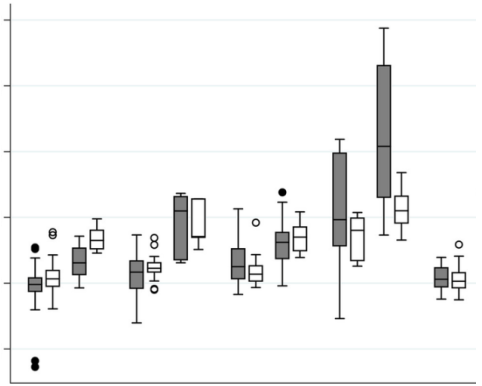
Observational Research: Foundations Matter!



Harmonized Clinical Records to Analytic Results



	v1	v2	…	vX
Patient 1	0	0	…	1
Patient 2	0	1	…	0
Patient n	1	0	…	1



Creating Cohorts and Variables

Apply additional logic

Apply relevant logical operators and time relationships between concepts

Example: Patient severity at the time of COVID is defined by logical relationships between six concept sets and applies two temporal tests.

3

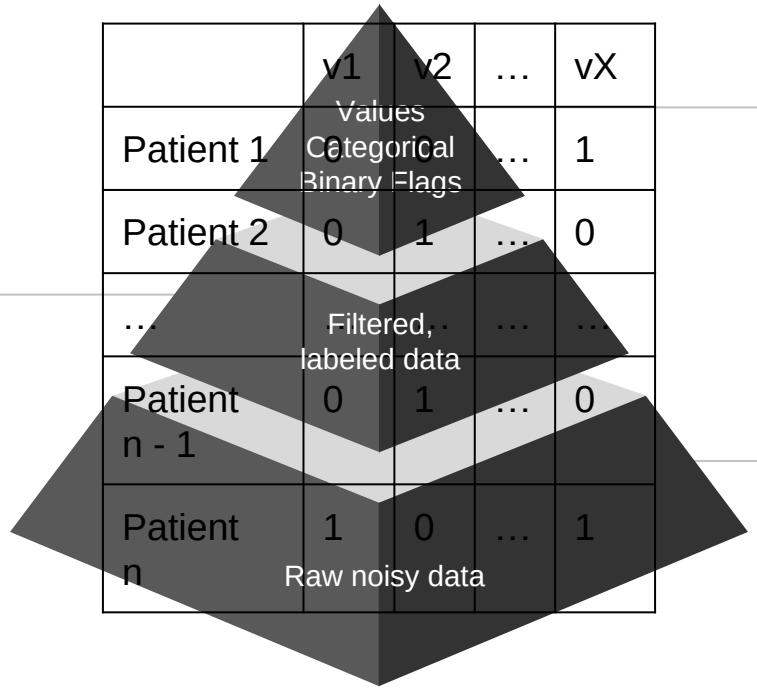
Apply Concept Sets

Use to collapse raw data.

Leverage community built concept sets.

Evaluate specificity and sensitivity in light of you team's specific aims.

2



Source Data

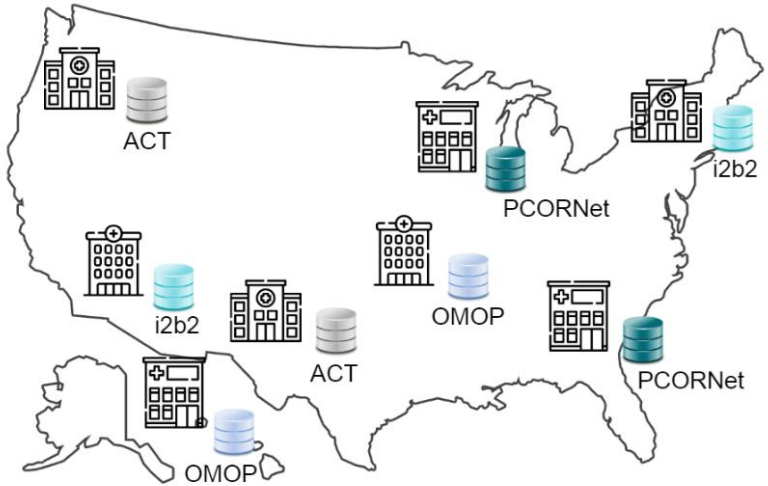
Find and evaluate relevant tables and fields in your source data.

Note limitations.

1

Important Considerations

Observational Health Data Limitations: Heterogeneity



Geographic Effects

Clinical Site Features

Electronic Health Record Systems

Clinical Data Models, Vocabularies, and Versions

Clinical Workflows, Care Practices, and Charting

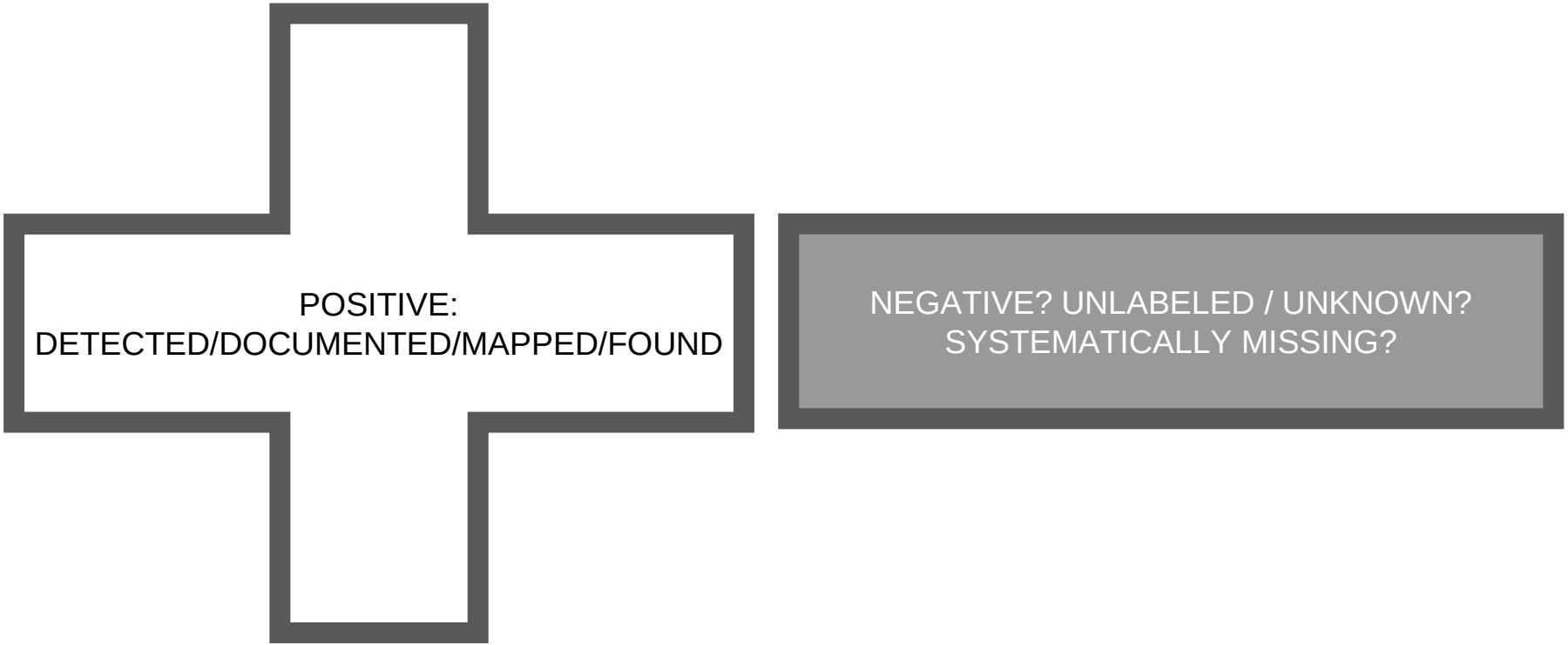
Patient Access to Care, Trust, Transparency, Compliance, and Context

Privacy Preserving Record Linkage (PPRL): Data Augmentation Limited to certain patient populations and sites (i.e. Medicare, Death, Cross-Site Patient Linkage)

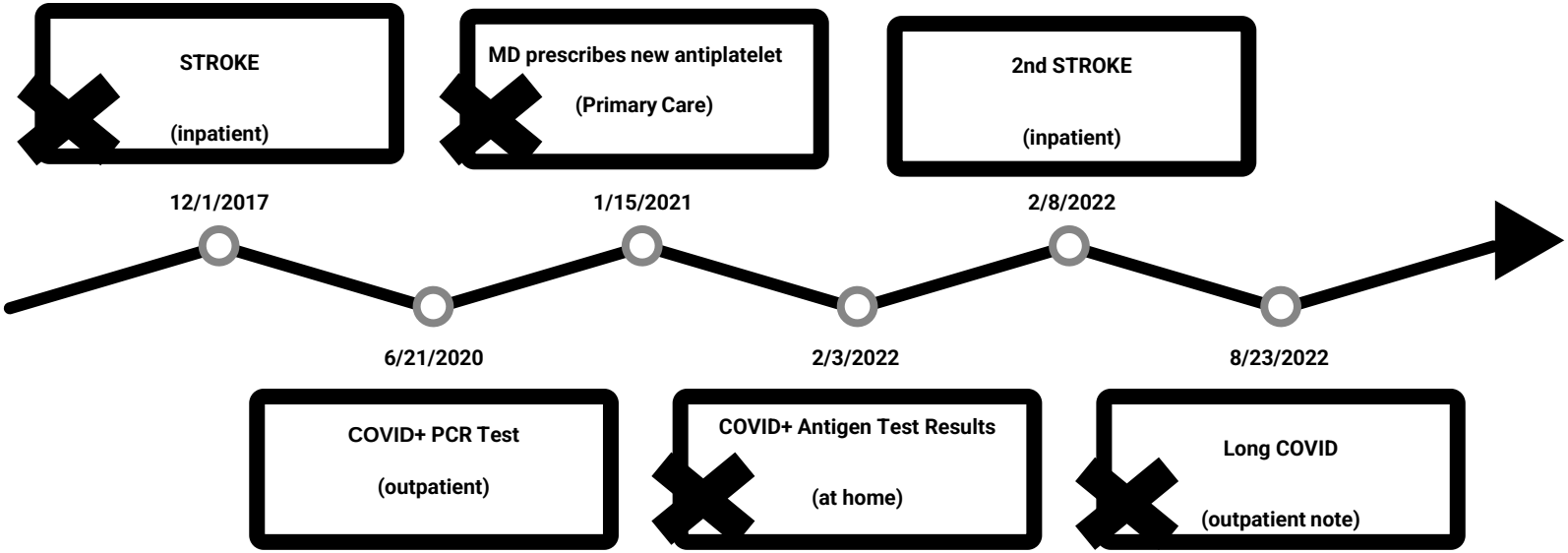
Etc……

Relevant N3C Tables: Manifest, OMOP, PPRL Data

Observational Data Limitations: Positive & Unlabeled Data



Observational Health Data: Time-Related Limitations

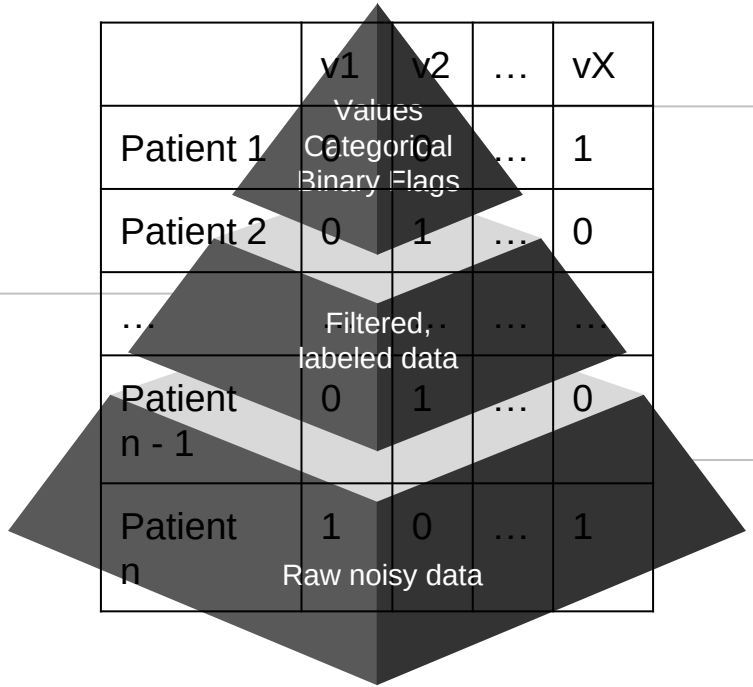


BEWARE of DATE SHIFTING, DATA LOSS, UNRELIABLE START/STOP DATES (chronic issues, patient lost to follow-up), EFFECT OF OBSERVATION PERIODS (data collection periods, data freshness), PHENOTYPE (patient selection effects)

Relevant N3C Tables: Manifest, OMOP, PPRL Data

Community Tools

Eliminating Redundant Efforts: Data and Logic Liaison Tools



Apply additional logic

3

Provide code templates that accelerate the generation of derived data elements

Apply Concept Sets

Generate and maintain a reliable core library of concept sets

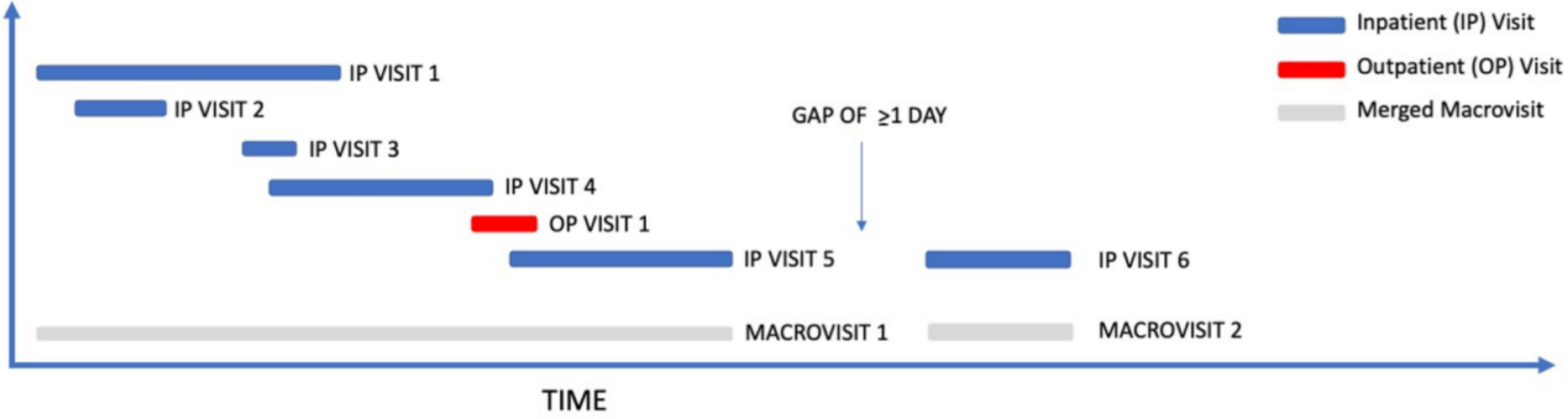
2

Source Data

1

Augment source OMOP tables with additional pre-processing

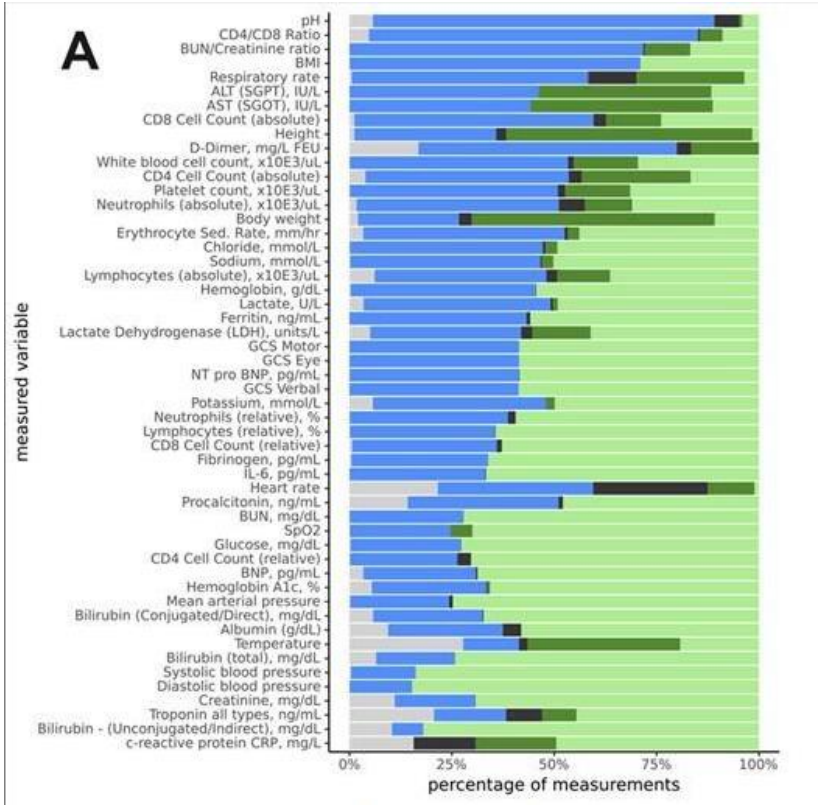
Data Cleaning: N3C Macrovisits



Clinical encounter heterogeneity and methods for resolving in networked EHR data: A study from N3C and RECOVER programs. Peter Leese, Adit Anand, Andrew Girvin, Amin Manna, Saaya Patel, Yun Jae Yoo, Rachel Wong, Melissa Haendel, Christopher G Chute, Tellen Bennett, Janos Hajagos, Emily Pfaff, Richard MoffittmedRxiv 2022.10.14.22281106; doi: https://doi.org/10.1101/2022.10.14.22281106

N3C Microvisit to Macrovisit Map Table: macrovisit_id, macrovisit_start_date, macrovisit_end_date

Data Cleaning: N3C Measurement Harmonization



	Harmonized Unit Present in Source
	Unit Mapped and Value Converted
	Unit Absent but Inferred and Converted
	Unit Missing (Unable to infer)
	Unit Mapped to Null (Nonsensical Value)

Bradwell KR, Wooldridge JT, Amor B, et al. Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (EHR) dataset. *J Am Med Inform Assoc*. 2022;29(7):1172-1182. doi:10.1093/jamia/ocac054

N3C Measurement Dataset Fields:

harmonized_unit_concept_id, harmonized_value_as_number

N3C Recommended Concept Sets

1. *Assemble Team*: N3C Clinical Domain Team Leads (or delegates), N3C Data or Logic Liaison (informaticians and analysts)

2. *Articulate Goals*: Define the scope, intentions, and limitations

3. *Explore*: Use OHDSI tools to explore the CDM and identify candidate codes

4. *Compare*: Use both authoritative and community-built concept sets

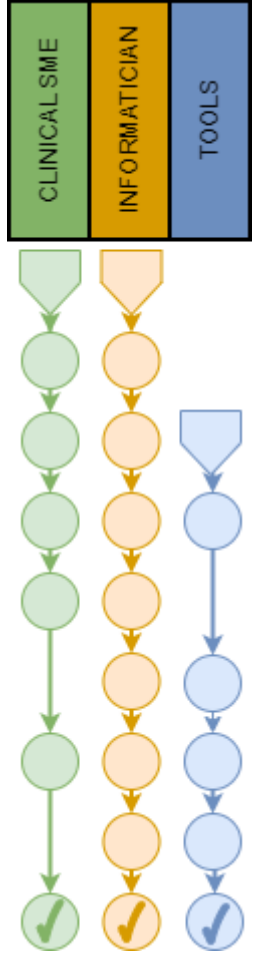
5. *Review*: Present to a broader team of clinical experts

6. *Collapse*: Reduce the intensional concept set expression as parsimonious as possible, retaining all the approved concepts collected in prior steps.

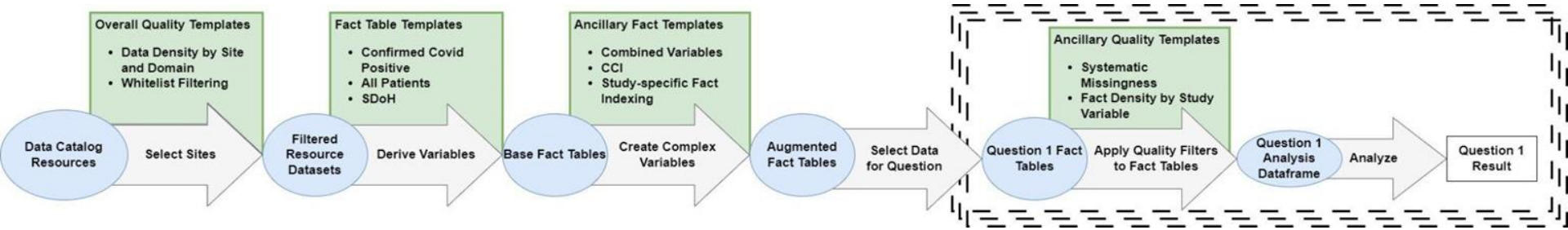
7. *Document*: Intention, Limitations, Provenance, and expert Reviews

8. *Present:* Present for final vetting at the Data Liaison informaticists' meeting.

9. *Publish*: Marked as N3C Recommended and published to Zenodo



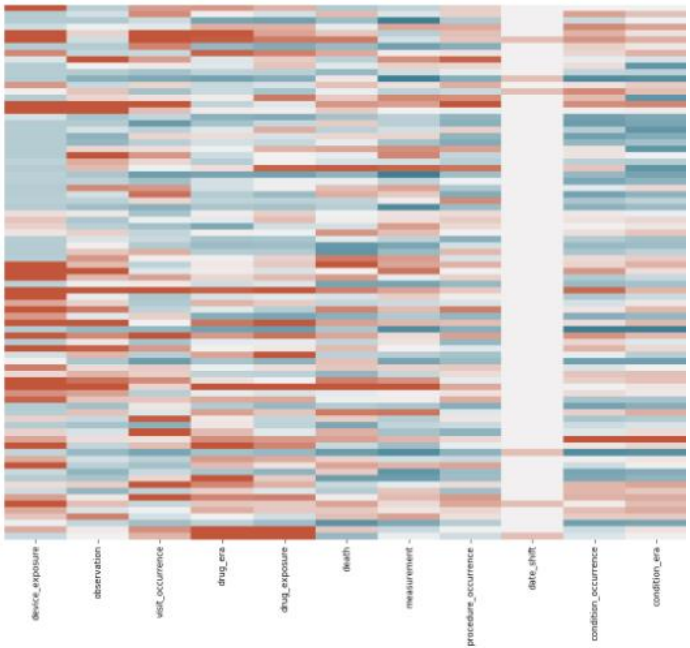
Eliminating Redundant Work: Logic Liaison Templates



Logic Liaison Tables: Minimal Data Cleaning/Imputation

- BMI/Obesity:
  - Unreasonable height/weight/BMI thresholds are applied and can be configured by the user
  - BMI computed when not reported and obesity imputed using BMI>30
- Unreasonable date detection:
  - DOB
  - Visits
  - Death

Logic Liaison Quality Templates:
Domain Density and Data Partner “Whitelist Filtering”

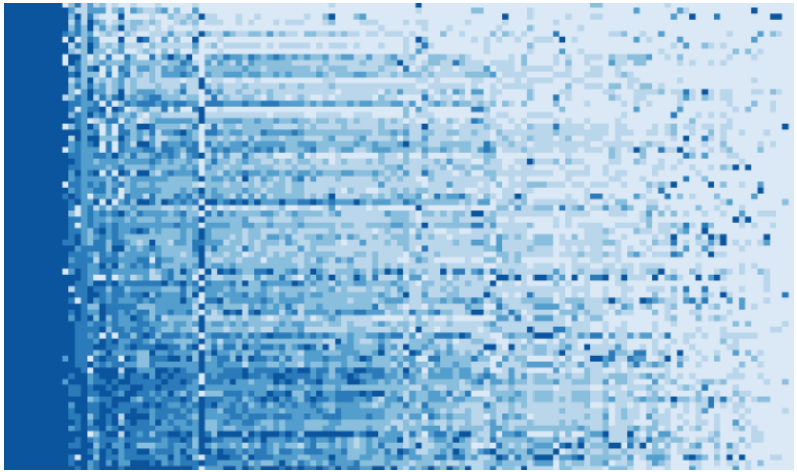


- How rich is each site’s data by OMOP Domain?
- Which sites meet customizable data quality standards

N3C Logic Liaison Quality Templates: Domain Density, Whitelist Filtering

Logic Liaison Quality Templates: Fact Density and Systemic Missingness

Quality must be assessed again
AFTER creating your derived variables
and PRIOR to analysis



N3C Logic Liaison Quality Templates: Fact Density by Site Visualization

Key Takeaways

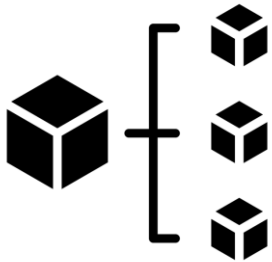
Work with a breadth of domain experts

Don’t recreate the wheel

Interrogate your raw data and derived dataframes

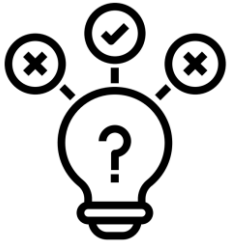
Leverage metadata

Understand and disclose limitations



Carefully derive features

Test your assumptions



Patience and humility