DATA QUALITY BRIEF 0003

How do I define my study sample (computable phenotype)?

This brief focuses on a common process faced by researchers using real-world evidence: the development of a computable phenotype. It is designed as an introduction, not as an exhaustive guide and is best used in complement with additional resources, particularly the guidance of an informatician.

USE CASE

Susan is leading a study focused on **asthma in patients with obesity**. She needs a **case definition** that she can use to identify the sample she will study from **EHR data**.

BROAD TOPIC AREAS

Concept Introduction

PaTH Network
patient empowered research

KEY CONSIDERATIONS

1. Developing a Computable Phenotype:

- The right level of specificity depends on the research question. Considering the full range of options helps you choose appropriately. Use a **funnel metaphor**: start with a broad population definition and narrow it by adding criteria such as diagnoses, service use, medications, or lab values.

2. Develop a clear research question that can be answered with structured EHR data. Follow these steps:

- Articulate how broad or narrow you would like the computable phenotype to be. For example, do you need to focus on specific demographic groups or a specific degree of severity? Or perhaps you are interested only in patients who receive a specific type of care?
- Start the computable phenotypes with demographic data (e.g., age, sex as appropriate for the study, potentially other factors) along with diagnosis (and/or sometimes procedure) codes.
- Be clear about the inclusion and exclusion criteria (e.g., based on ICD-10 codes, laboratory results, etc.).

USEFUL TIPS

1. Define a timeframe of interest.

- Your definition will vary whether you are interested in incidence (new cases) vs. prevalence (all cases in a given timeframe)

2. Before creating a code list, think of who you want to include. Make sure you are capturing that group.

- The role your computable phonotype plays within your design (e.g. cohort definition, exposure, outcome) is important for understanding tradeoffs between sensitivity and specificity.
- Should you use a clinical definition (e.g., ICD-10 codes) and/or a laboratory definition (e.g., pulmonary function tests for asthma, or serum glucose +/- glycosolated hemoglobin for diabetes)? Requiring laboratory data provides greater certainty, but is likely to also impose limits on the size of your sample.
- Do you need to identify only patients who have been diagnosed with the condition of interest or would you also like to include patients who likely have the condition, but do not have a diagnosis code in the chart?
- Decide which set of codes or ontology you intend to use (eg SNOMED, ICD, etc) and how the data are mapped to the existing standard (ICD for PCORnet) in the data source. When possible work with an informatician and clinician.

3. Keep in mind that the presence of a diagnosis code does not always mean a condition is present.

- Providers may use a code when they order a test to potentially roll out a diagnosis. That code can stay in the chart after the diagnosis is ruled out.
- Consider requiring an ICD 10 code to be used more than once before using it to indicate a diagnosis. How many encounters with an ICD 10 code will you require for identification? Example: An asthma researcher requires 2 diagnosis of asthma coded in a given time frame.

4. Not all conditions are adequately identified using diagnoses codes

- For example to identify patients with obesity, ICD-10 diagnosis codes could be used, but could miss patients as obesity is often not recorded as a diagnosis. You could consider using a body mass index (BMI) value for obesity status or use height and weight to calculate BMI for those also missing BMI. If reasonable, carry height forward to utilize weight values without a height measurement. You could also employ multiple methods using an “or” function.

5. Consider whether disease severity is important to the research question.

- Disease severity can impact treatment which can lead to confounding by indication.
- In 2015 ICD-9 codes were mapped onto ICD-10. Older diagnosis values, entered as ICD-9 codes may lack the specificity of ICD-10 codes now available because ICD-9 codes did not capture disease severity. In such cases complementary information such as the number of ED visits in a certain time frame may be helpful for establishing severity.

6. Symptoms can apply to multiple diagnose and may be under- or over-coded.

- E.g. Not everyone who wheezes has asthma, so the symptom alone could lead to too broad a list of patients. In such a case, you might want to consider combining the symptom (e.g., wheezing) with the presence of asthma-specific medications.
- Symptoms (e.g. shortness of breath) are most commonly recorded in free text narrative so they may be under coded and require access to free text, such as notes.
- Even symptoms such as pain that may have a standard scale may not be captured consistently.

7. Determine whether your computable phenotype incorporates your exposure variable.

- Consider a study on use of a bronchodilator among people with asthma. If bronchodilator use is part of the computable phenotype for asthma, you may be biasing your ability to assess bronchodilator use as an exposure or outcome. e.g. people with asthma who use a bronchodilator, for a study of the effects of long-term bronchodilator exposure. Consider your hypotheses and planned analyses carefully to make sure you are capturing the population needed to answer the question as stated.

8. At times, there is value in having a "core" computable phenotype, and then a couple of variations that are addressed in sensitivity analyses.

- Consider the fact that the "core" phenotype might be more sensitive and the sensitivity analysis phenotype might be more specific (or the other way around depending on your research question). For example, if you were looking at the impact of GLP-1 inhibitors you might have your primary group be people with diabetes taking those GLP-1 inhibitors that are FDA-approved for diabetes. As a sensitivity analysis, you could expand the group to patients with diabetes taking any GLP-1 inhibitor.

9. Work iteratively with your local informatics team and consult a clinician when possible.

- Informaticians are experts at how the EHR organizes and stores clinical data.
  - They have unique experience with workflow and processes that influence the way data is captured, stored, and interpreted.
  - They also have knowledge of standard terminology that can support research collaborations across sites. For example, clinicians may not be aware of the shift from ICD-9 to ICD-10 in 2015. This must be accounted for when making code lists for a study spanning this time period.

- Likewise, local shifts in data storage or institutional changes to EHR workflows may impact data and not be apparent to non-informaticians. As another example, in some systems when patients are admitted from the emergency department, an ordered procedure may be linked with emergency department admission even if it occurs under a separate admission record, which was generated at the time that the patient shifted to an in-patient bed. To link the procedure appropriately with the illness, both admissions should be accounted for.

- Clinicians can clarify nuances in clinical coding, which may vary across health systems—codes that seem unusual at one site may be essential at another to identify your population of interest. Clinicians may also be aware of policy changes in diagnosis or treatment that could affect how well your phenotype identifies your population of interest over time.
  - For example, hypertension treatment guidelines in 2017 lowered the diagnostic threshold for high blood pressure from 140/90 to 130/80, so the number of people with uncontrolled blood pressure shifted abruptly. Analyzing the data without an understanding of the shift in clinical practice would lead to confusion.
  - For instance, medications used to manage a condition prior to 2020 may differ substantially from those used today. Incorporating clinician knowledge helps ensure that phenotype definitions remain valid across longitudinal datasets and accurately capture the intended patient population despite temporal changes in clinical practice.

10. Consider using, updating, or modifying a previously validated computable phenotype.

- Investigators may carry out chart reviews, and use them to define positive predictive value and negative predictive value for a given computable phenotype within a broader population of interest.
- Some potential sources of previously deployed computable phenotypes include:
  - CMS Chronic Conditions Warehouse
  - AHRQ Clinical Classifications Software
  - eMERGE Network
  - FDA Sentinel Initiative
  - CMS QualityNet
  - PheKB Phenotype Knowledgebase
  - HDRUK Phenotpye Library
  - NLM Value Set Authority Center
  - OHDSI Phenotype Library
  - N3C Concept sets and cohorts
  - Primary peer-reviewed literature
  - PEDSNet Data Quality Modules
  - OHDSI/OMOP Phenotype Library

A graphic summary of these tips is included below.

Defining a Study Sample

Computable Phenotype: Key Rules

Possible additional considerations

- Minimum observation period / data history
- Data completeness and missingness
- Cross-site portability / external validation
- Versioning and reproducible documentation

Goal: *create a computable phenotype with the right balance of sensitivity and specificity.*

ADDITIONAL RESOURCES

- Carrell, D. S., et al. (2024). A general framework for developing computable clinical phenotype algorithms in EHRs. *Journal of the American Medical Informatics Association.* (Available via PMC).
- Mo, H., Thompson, W. K., Rasmussen, L. V., et al. (2015). Desiderata for computable representations of EHR-driven phenotype algorithms. *Journal of the American Medical Informatics Association*.
- eMERGE Network. Phenotyping: Cohort discovery using EHR data. eMERGE-Network.org.