FAIR Phenotyping with APHRODITE

PRESENTER: **Juan M. Banda**
EMAIL: **jbanda@gsu.edu**
TWITTER: **@drjmbanda**

- Electronic phenotyping has been evolving from simple to complex rule-based definitions over the years, more recently entering the machine learning age with probabilistic phenotype models.
- With the added complexity comes the additional need to have consistent and reproducible phenotype definitions for maintenance, replicability, and community sharing.
- In this work we introduce how to construct probabilistic phenotype definitions with Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) that follow the FAIR principles to improve their reproducibility and quality.
- By using a centralized repository and creating a standard list of meta-data elements, we aim to guide probabilistic phenotype definition developers with a FAIR-compatible standard.
- **By developing this standard within the Observational Health Data Sciences (OHDSI) initiative, we aim to ensure community wide compatibility and maximum reproducibility.**

Findable Accessible Interoperable Reusable

The common failure to reproduce published results has created an atmosphere of crisis even in disciplines where precise measurement and tight experimental control are the norm. There is even more reason for vigilance in disciplines that must manage lower degrees of measurement accuracy and experimental control. Observational health research on large secondary data is a case in point. One response to this crisis has been the emergence of open science principles that publicly expose the process of defining hypotheses, data selection and development, study design and analytic choices.

Anatomy of a FAIR phenotype definition

A phenotype definition will be **Findable**

To address the need to have a persistent global unique resource identifier (URI) for each phenotype definition version, we have utilized GitHub unique commit hash value to identify each individual phenotype definition version. While this approach is rather simplistic, it achieves the goal and simplifies the process for authors. The OHDSI Gold Standard Phenotype Library workgroup has defined and created an additional abstraction layer over the phenotype definitions available as a R Shiny App.

A phenotype definition will be **Accessible**

The phenotype definition, generation script, and metadata will be retrievable by their identifier using any regular web browser or the application layer of the phenotype library. By using a publicly and freely available resource such as GitHub, we offer better accessibility than placing the definitions on an institutional server.

A phenotype definition will be **Interoperable**

We will leverage the OMOP CDM and associated vocabularies to solve the major obstacle to interoperability across sites. Our phenotype definitions' metadata will use JSON for knowledge representation and ease of machine readability. When developing phenotyping definitions based on prior publications, or when a publication is generated from a definition generated from our pipeline, we will include all proper URI's to the publications in question.

A phenotype definition will be **Re-usable**

Currently APHRODITE definitions are easily shareable and re-usable for other sites. We have added meta-data elements related to software, CDM, and vocabulary versions, as well as a plurality of accurate and relevant attributes to guarantee re-usability. All the publicly available phenotypes will be released under relevant open source licenses, details of which will be attached to the definition's meta-data. Site and researcher information will be recorded as well as relevant publications in allowing fully traceable provenance for each definition.

Take a picture to **download** the **full paper**

Table 1. Meta-data elements to for re-usability. All elements with a * are required and will be auto-populated using system calls and the APHRODITE configuration file.

Meta-data element	Description
Generating Institution*	Generating institution name
Generator Name*	Maintainer and responsible individual name
Generator ORCID*	Maintainer and responsible individual ORCID
Date Generated*	System recorded phenotype definition generation date
Validating Institution	(If available) Name of validating institution
Date Validated	(If available) Date the phenotype definition was validated
Validator Name	(If available) Name of validator
Validator ORCID	(If available) Validator ORCID
License*	Licensing information under which the definition was released
Aphrodite Version*	Which version of the APHRODITE package was used
CDM Version*	Version of the OMOP CDM utilized
Vocabulary Version*	OHDSI Vocabulary version
Vocabularies Included*	Included vocabularies list from the generating site
R Version*	R Statistical software version used
R Dependencies and Versions*	APHRODITE package dependencies used and their versions
Database Used*	Database server used
Publication Source	Identifier of publication the phenotype is based on
Published In	Identifier of publication the phenotype was released under
Previous Location	GitHub URL of the source phenotype (if being reused)

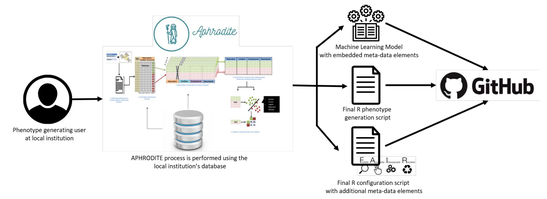

Figure 1. FAIR APHRODITE phenotype generation process.

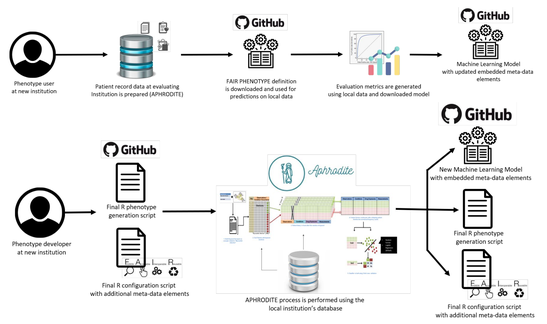

Figure 2. Usage of a FAIR APHRODITE phenotype definition.

Juan M. Banda[1] , Andrew Williams[2], Mehr Kashyap[3], Martin G. Seneviratne[3], Aaron Potvien[4], Jon Duke[4], Nigam H. Shah[3]

[1]Department of Computer Science, Georgia State University, Atlanta GA 30303
[2]Tufts Medical Center, Boston MA 02111
[3]Center for Biomedical Informatics Research, Stanford University, Stanford CA 94305
[4]Georgia Tech Research Institute, Atlanta GA 30308

Georgia State University

OHDSI