FAIR Phenotyping with APHRODITE

Juan M. Bandaa,[[1]](#footnote-1), Andrew Williams *b*, Mehr Kashyap *c*, Martin G. Seneviratne *c*, Aaron Potvien *d*, Jon Duke *d*, Nigam H. Shahc

a Department of Computer Science, Georgia State University, Atlanta GA 30303

b Tufts Medical Center, Boston MA 02111

c Center for Biomedical Informatics Research, Stanford University, Stanford CA 94305

d Georgia Tech Research Institute, Atlanta GA 30308

**Abstract.** Electronic phenotyping over the years has been evolving from simple to complex rule-based definitions, and 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.

**Keywords.** Electronic Phenotyping, Machine Learning, Electronic Health Records, OHDSI, FAIR

# Introduction

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. Open science principles have been adopted by the Observational Health Data Science and Informatics (OHDSI) community to promote the reproducibility of its members’ research.

At that same time, a movement to make scientific work that uses big data more efficient has emerged. It seeks to standardize data and other digital artifacts in order to facilitate the consistency and ease with which these can be used by a community of scientists. The acronym FAIR stands for the set of principles that make this possible: making digital scientific artifacts Findable, Accessible, Interoperable and Reusable.

The OHDSI community implements these principles in several ways by standardizing heterogeneous clinical data to conform to a public common data model (CDM) with metadata provided by controlled medical terminologies and by ensuring interoperability through the mapping of semantically equivalent concepts. For example, different versions of the same lab code are all mapped to the same standard concept. OHDSI develops and publishes publicly available tools that allow concept sets to be searched for and used to define patient cohorts with data that can be analyzed using standard open source libraries. This allows analytic code written at one site to be reused while producing consistent results at other sites regardless of differences in their source data. The definitions of cohorts used in such analyses are referred to as computable phenotypes. A phenotype definition, in this sense, is any combination of healthcare data values that are used to classify patients based on one or more defining attributes.

In recent times, biomedical and health informatics have been facing reproducibility and replicability issues [1–3] as previously published studies could not be replicated, making their findings not quite useful. Unsurprisingly, with the exponential growth of publications, methods, datasets and resources, researchers going back to 2013 were asking for ways to aid scientific reproducibility [4]. Machine learning and artificial intelligence approaches also suffer their own reproducibility crisis [5], based on the fact that most algorithms need a good amount of tuning parameters and are highly sensible to the underlying training data, both of which are rarely publicly available for other researchers. We opted to make FAIR [6] phenotype definitions with APHRODITE for two main reasons: APHRODITE is built on top of the OMOP common data model (CDM) and supported by the OHDSI initiative, which means that all OHDSI sites will have the same underlying data representation and vocabulary, allowing for easier interoperability between multiple sites. Second: all tools, data models, software and vocabularies are publicly available, released with open-source licenses to ensure maximum transparency and availability for all researchers. Taking advantage of the OHDSI initiative philosophy and the availability of its tools and members, we believe that creating FAIR APHRODITE phenotype definitions can drive the community and then the field to adopt FAIR principles for phenotyping and enabling higher reproducibility, availability and replicability of research results and studies.

# Background and related work

As defined in [7], electronic phenotyping can be seen having two main approaches: rule-based methods and machine learning methods. While this work describes how to create FAIR phenotypes with APHRODITE, a machine learning method, we present the advances rule-based phenotyping researchers have made in creating reproducible and accessible phenotype definitions. We also discuss how the Observational Health Data Sciences and Informatics (OHDSI) initiative and its gold standard phenotype library will get us closer to fully reproducible and accessible phenotypes.

## Rule-based phenotyping

Involving one or more clinicians to specify inclusion and exclusion criteria based on structured data elements, available on electronic health records (EHR), such as diagnosis codes, procedure codes, laboratory values and medications. Often drawn from guidelines around diagnosis and treatment of the particular phenotype [8], these criteria are usually agreed upon by more than a few experts in the field. For example, in order to build a rule-based phenotype definition to identify patients with type 2 diabetes, such a definition would include a laboratory measurement of HbA1c above a certain threshold, one hypoglycemic medication and one or more mentions of the actual type 2 diabetes code [9].

With any research group or institution able to create their own phenotype definitions, it was not until 2013 that the Electronic Medical Records and Genomics (eMERGE) Network was established as a consortium of academic medical centers, with the purpose of developing generalizable EHR phenotype definitions intended to conduct genome wide association studies across multiple clinical datasets [10]. The eMERGE consortium is responsible for a large catalog of phenotypes including hypothyroidism, type 2 diabetes, atrial fibrillation and multiple sclerosis [9,11,12]. Two important lessons were that developing phenotype criteria is an iterative process that benefits from input from multiple sites [10], and it requires around six to eight months for the development of a single definition [13]. These elements are highly relevant as to why it is necessary to have control over definition versioning, provenance, and versions of the underlying data and software tools used for the construction of the phenotype. For the wide dissemination of phenotype definitions, the Phenotype Knowledgebase (PheKB) was created [13], which currently provides 53 publicly-available phenotypes. The vast majority of the phenotype definitions available here, at the time of writing, are rule-based using structured data. More recently, researchers at the Mayo clinic have introduced: Phenotype Execution and Modeling Architecture (PhEMA) [14] which takes the eMERGE phenotype algorithms stored in PheKB and makes them easier to implement across multiple institutions in an automated way. However, PhEMA does not provide enough metadata elements to ensure versioning and provenance while trying to address the generalizability of definitions. The problem of how to define rule-based phenotypes has already been addressed, with even design patterns available for their construction [15], but how to properly share them, store them and ensure maximum replicability is still a wide-open research problem. While the eMERGE network established important pioneering practices for externally validating phenotype definitions. There are still latent challenges to doing such validations efficiently and maximizing the reproducibility of phenotypes across sites. The same patient attributes will often be represented by different coding systems or units across sites are treated as semantically equivalent.

## Phenotyping with Machine Learning

Over the years rule-based definitions have initially been preferred as they provide an easily interpretable (rules) as to why a person belongs (or not) to a certain phenotype. While some machine learning algorithms are viewed as black-boxes (support vector machines (SVM), gradient boosted trees, etc.), phenotyping has slowly been moving into the machine learning realm. Li et al. used demographics, lab measurements and coded billing data alongside distributional association rule mining (ARM) to model phenotypes [16], by generating interpretable and exhaustive rules. Since ARM is a supervised learning algorithm, the authors used labels created by implementing and applying the eMERGE type 2 diabetes phenotype definition. Having a labeled set of patients, ARM which utilizes the Apriori algorithm, discovered combinations of rules through exhaustive enumeration. When compared against standard ML algorithms (logistic regression, decision-trees, and SVM) the authors showed better performance against an expertly curated gold-standard. The expertly curated gold-standards used with both rule-based definitions (to evaluate performance) and with the machine learning definition (to train algorithms) are sets of patients that have been validated by clinicians to have (cases) or not have (controls) the given phenotype at hand. This is the biggest bottleneck in phenotyping tasks as these sets are expensive to create, both in terms of time and resources. Additionally, gold-standard sets are not portable across institutions and they cannot be easily shared due to HIPAA regulations. In order to alleviate the need for gold-standards, both Halpern et al. [17] and Agarwal et al. [18] proposed methods that take advantage of clinical notes and other structured EHR data. The intuition underlying both approaches is that by using a large amount of imperfectly labeled training data, we can still learn good phenotype classifiers.

## The Observational Health Data Sciences and Informatics (OHDSI)

In order to gain widespread distribution for our probabilistic phenotype definitions, we have built our phenotyping software and definitions within the OHDSI community. OHDSI is a world-wide collaborative featuring over 160 collaborators in 18 countries comprised of healthcare industry leaders, clinical researchers, computer scientists and biostatisticians. With a vision to improve health by empowering a community to collaboratively generate evidence that promotes better health decisions and better care [19], we believe that our FAIR phenotype definitions can be of great impact here. In order to provide standardization within the community, OHDSI features both a common data model (CDM), and a standard vocabulary for consistent representation of EHR data across sites. Approximately 90 sites have converted their local data into the common OHDSI CDM including both clinical and claims, totaling over 600 million patients, allowing us to reach a good amount of institutions and a large amount of patient data. Our probabilistic phenotyping software is called APHRODITE [20], which is an open source R software package for building phenotype models. The output of this package is the FAIR phenotype definitions that can be shared and ported across institutions as shown in [21].

Our FAIR phenotype definitions are a product of the R package APHRODITE, which implements phenotyping methods developed by Halpern et al. [17] and Agarwal et al. [18], for the OHDSI community. Both of these methods which focus on shifting the paradigm of needing gold-standards and take advantage of clinical notes and structured EHR data. Both of these methods rely on using a large amount of imperfectly labeled training data, in order to learn good phenotype classifiers. This has been shown to be true and work as expected for datasets converted to the Observational Medical Outcomes Partnership (OMOP) CDM in [20].

## A gold standard phenotype library

In order to enable members of the OHDSI community to find, evaluate, and utilize community-validated cohort definitions for research and other activities, OHDSI is currently building a Gold Standard Phenotype Library. This repository will centrally house phenotype definitions which meet certain “Gold Standard” development, validation, and documentation standards held by the community. The library supports probabilistic phenotype definitions, such as those created using APHRODITE, as well as rule-based (heuristic) definitions.

The foundation of the library relies on the following interactive roles: Users, Authors, Validators, and Librarians. Under this framework, an APHRODITE definition could be submitted by an Author, vetted and inducted into the library by its Librarians, tested at various sites by independent users acting as Validators, and utilized for its intended purpose by any number of Users. This provides a natural framework for sharing APHRODITE and other phenotype definitions in an open, FAIR-aligned manner.

# Limitations of this work

The main limitation of this work is that we do not address (or intend to) the sharing of the underlying patient data used to build the phenotype definition models. Due to HIPAA regulations and countless other internal institutional regulations, this is something we cannot incorporate to our FAIR definitions. However, our APHRODITE models provide the next best option as they are objects resulting from the training of classifiers using the actual patient data. This is considerably better than just providing a list of instructions or rules to build a phenotype, and letting the potential users interpret them at will. By leveraging the OHDSI Vocabulary and CDM we are steps closer to provide universal (with limitations) definitions and the APHRODITE models are a result of this. While we provide all the needed facilities to make the phenotypes available and all the meta-data elements required to have both provenance and reproducibility, we are aware that there might be versioning issues within the OHDSI tools, CDM, and R packages that might hinder this. However, under the best-case scenario, all the needed elements are present on our definitions.

# Anatomy of a FAIR phenotype definition

In this section we define how our probabilistic phenotype definitions incorporate the basic FAIR [6] principles. Most of the additional meta-data elements are automatically generated by the APHRODITE package, and others are part of a configuration file. To enforce compliance with the required elements, APHRODITE does not allow users to export their phenotype definition models and predictors without having completed all the proper configuration file requirements. This enforces researchers to fill in all needed information as it will be embedded into the phenotype definitions automatically.

## 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.

By utilizing public GitHub repositories we are ensuring that the phenotype definitions will be findable (GitHub is indexed by all major search engines). Additionally, the OHDSI Gold Standard Phenotype Library workgroup has defined and created an additional abstraction layer over the phenotype definitions available in GitHub. Available as an R Shiny App [22], users are able to browse through available phenotype definitions in a user friendly manner. We have described this gold standard phenotype library in the background section.

## 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. For the case of restricted use phenotype definitions, we will handle all access control and authorization using private repositories on GitHub. We will store all metadata in alternative locations (phenotype library) to maintain its accessibility even if GitHub becomes inaccessible.

## 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 either JSON or GitHub Markdown 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.

There will be different versions of each phenotype at generating institutions, for this we have identified and added the following meta-data elements to be attached to the phenotype model binary object and an additional JSON object (for automated readability):

**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 validate |
| 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 Versions\* | APHRODITE package dependencies used and their versions  |
| Database Used\* | Database server used |
| Publication Source | Identifier of publication the phenotype is based from |
| Published In | Identifier of publication the phenotype was released under |
| Previous Location | GitHub URL of the source phenotype (if being reused)  |

Table 1 provides three types of meta-data elements: Generator and validator items (light blue cells), technical elements (light red cells), and provenance elements (green cells). While all the generator elements are fully mandatory, a phenotype can be produced but not fully validated, thus the validator fields are not required to be posted. Once the phenotype is validated either locally, or at a different institution, they can be populated. For all technical elements, APHRODITE has an auto-populate function to automatically fetch all the required fields. This function will be called whenever the save phenotype model function is invoked. The provenance elements are intended for each version of a phenotype to be released by the corresponding authors or modifying team. This is the reason we have a source and a ‘published in’ elements, to keep track of provenance chain, and with the previous location element to contain the source phenotype GitHub URL is being reused/modified by other authors.

# Generation and Usage of FAIR APHRODITE phenotype definitions

In Figure 1, we describe the generation of a FAIR APHRODITE phenotype definition. It starts with a user generating a phenotype at their local institution. The user then follows all the APHRODITE steps to build the machine learning model using their own local institution data, for more details on this process please read [20]. Once the definition is stable and performs desirably, APHRODITE will save the phenotype definition model by automatically attaching all meta-data used in the configuration process as well as all local versioning details. Lastly, the definition is uploaded to be shared with others.

**Figure 1.** FAIR APHRODITE phenotype generation process.

Once the definition is uploaded on GitHub, researchers can download it for two different purposes: validation and improvement. In Figure 2 we show the flow of these two purposes and how these definitions further incorporate the validated meta-data elements and the provenance elements (improvement).

**Figure 2.** Usage of a FAIR APHRODITE phenotype definition.

# Sample FAIR APHRODITE phenotype definitions

With this paper we are releasing five phenotype definitions for Type-2 Diabetes Mellitus (T2DM), Myocardial Infarction (MI), Peripheral Artery Disease (PAD), Abdominal Aortic Aneurysm (AAA) and Heart Failure (HF). These phenotype models have been developed and trained at Stanford University and are described in more detail in [MEHR’s paper]. The phenotype models were built with the purpose of multi-site evaluation, and we have attached the necessary meta-data and uploaded them to a public GitHub repository to conform with all our FAIR phenotype definition parameters. To our knowledge, this is the first time a research group released a group of trained phenotype models to the public with the purpose of reproducibility, reusability and interoperability.

The definitions can be found at the following URL:

<https://github.com/thepanacealab/FAIR_APHRODITE_phenotypes>

# Conclusions and future work

In this work we have introduced the basic elements needed to have FAIR phenotype definitions using APHRODITE and released five different phenotype definitions with said elements incorporated. By being able to embed meta-data elements directly into the machine learning models, we are providing all the necessary elements to improve reproducibility and efficiency. With a GitHub hosted central phenotype library we also ensure accessibility, permanence, provenance, and availability of our phenotype models. Providing minimal guidelines for the FAIRness of machine learning phenotype models built with APHRODITE will enable us to further extend these guidelines to encompass rule-based definitions (ATLAS cohorts [23]) within the OHDSI community in the near future. Our overall goal is to have all types of phenotype definitions be FAIR within the OHDSI community, making it the first initiative to fully adopt the FAIR principles for its phenotyping efforts.

References

1. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc. 2018;25: 963–968. doi:10.1093/jamia/ocy028
2. Bakken S. The journey to transparency, reproducibility, and replicability. J Am Med Inform Assoc. 2019;26: 185–187. doi:10.1093/jamia/ocz007
3. Glenn Begley C, Ioannidis JPA. Reproducibility in Science. Circ Res. Lippincott Williams & WilkinsHagerstown, MD; 2015; doi:10.1161/CIRCRESAHA.114.303819
4. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ. 2013;1: e148. doi:10.7717/peerj.148
5. Hutson M. Artificial intelligence faces reproducibility crisis. Science. 2018;359: 725–726. doi:10.1126/science.359.6377.725
6. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. doi:10.1038/sdata.2016.18
7. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. Annual Reviews; 2018;1: 53–68. doi:10.1146/annurev-biodatasci-080917-013315
8. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21: 221–230. doi:10.1136/amiajnl-2013-001935
9. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012;19: 212–218. doi:10.1136/amiajnl-2011-000439
10. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013;15: 761–771. doi:10.1038/gim.2013.72
11. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet. 2011;89: 529–542. doi:10.1016/j.ajhg.2011.09.008
12. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010;86: 560–572. doi:10.1016/j.ajhg.2010.03.003
13. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23: 1046–1052. doi:10.1093/jamia/ocv202
14. Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. J Am Med Inform Assoc. 2018;25: 1540–1546. doi:10.1093/jamia/ocy101
15. Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. J Biomed Inform. 2014;51: 280–286. doi:10.1016/j.jbi.2014.06.007
16. Li D, Simon G, Chute CG, Pathak J. Using association rule mining for phenotype extraction from electronic health records. AMIA Jt Summits Transl Sci Proc. 2013;2013: 142–146. Available: https://www.ncbi.nlm.nih.gov/pubmed/24303254
17. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc. 2016;23: 731–740. doi:10.1093/jamia/ocw011
18. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016;23: 1166–1173. doi:10.1093/jamia/ocw028
19. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216: 574–578. Available: https://www.ncbi.nlm.nih.gov/pubmed/26262116
20. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017: 48–57. Available: https://www.ncbi.nlm.nih.gov/pubmed/28815104
21. Banda JM, Sarraju A, Abbasi F, Parizo J, Pariani M, Ison H, et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. npj Digital Medicine. 2019;2: 23. doi:10.1038/s41746-019-0101-5
22. PhenoPipe [Internet]. [cited 9 Jun 2019]. Available: <http://data.ohdsi.org/PhenotypeLibraryViewer/>
23. ATLAS [Internet]. [cited 9 Jun 2019]. Available: <http://www.ohdsi.org/web/atlas/#/home>
1. Juan M. Banda, Department of Computer Science, Georgia State University, 25 Park Place, Atlanta, Georgia, USA; E-mail: jbanda@gsu.edu ORCID: 0000-0001-8499-824X [↑](#footnote-ref-1)