OHDSI Gold Standard Phenotype Library
Requirements Document

Contents

Objective

The objective of the OHDSI Gold Standard Phenotype Library is to enable members of the OHDSI community to find, evaluate, utilize, and contribute community-validated cohort definitions for research and other activities.

Summary

The objective is to be achieved by establishing and maintaining a repository of community-validated cohort definitions. A user interface will be built to facilitate searching, examining, and downloading cohort definitions from this repository. Additionally, the ability to submit new definitions, validate existing definitions, and a cite usages of definitions will be possible. Data submission will encourage and/or require certain documentation standards. After data are submitted, they are subject to verification by Librarians (volunteer OHDSI subject matter experts) prior to becoming sanctioned as part of the library's official record.

The library will allow for the following features/activities:

- A **user interface** that facilitates the submission and retrieval of data pertaining to the definition, validation, and citation of cohort definitions
  A **staging area** where librarians can review/edit proposed library submissions and to reorganize the library's structure, as needed
- A **repository** which represents the official record of the library
- **Provenance,** such that linkages can be made between disparate cohort definitions to associate them with each other
- The ability to **save and load** cohort definitions to enable working intermittently
- **Authentication** for non-read-only activities (i.e. submitting data or saving a submission draft)
- All authors listed on a cohort definition will receive **an automated confirmation e-mail** when their definition has been submitted
- **Structured forms** for submitting a **cohort definition**, **validation set**, or **citation usage**, with **data integrity checks** where possible

Library Analogy

Throughout the brainstorming process, we've frequently benefited by drawing analogies between the phenotype library and a brick-and-mortar library.

Book – Phenotype

A library is filled with books. Our "books" are phenotypes, which we have defined as the following:

Phenotype - as it pertains to observational research, an observable set of characteristics in health data about an organism. A phenotype's purpose is the desired intent to identify members in a health dataset with the observed set of characteristics of interest. The observable set of characteristics can include conditions, procedures, exposures, devices, observations, etc.

A book is merely a container – it is a bucket that holds related chapters.

Chapter – Cohort Definition

Within each book is a collection of chapters. Our “chapters” are cohort definitions (phenotype algorithms) that are intended to approximate the phenotype of the book in which they are contained. We have adopted the following definition of what is meant of a cohort definition:

Cohort Definition - a coded set of instructions for best approximating the desired intent of identifying members of a phenotype. Defines a set of members in health data who satisfy one or more criteria for a duration of time. Each phenotype could have one or more phenotype algorithms (e.g. T2DM broad, T2DM narrow). The instructions could be rule-based (heuristic) or computable (probabilistic). Heuristic based phenotype algorithms consist of rules and one or more concepts sets. Probabilistic phenotypes are implemented using a predictive model.

In this sense, each chapter is a distinct cohort definition. The existence of multiple chapters within the same book implies that there could be many variations of the same definition. This is because there is not a one-size-fits-all cohort definition for every phenotype. For example, there may be a need for a sensitive and specific definition of the same phenotype.

Cohort

When a cohort definition is selected and executed on a database, a cohort results. Due to differences in data, two people running an identical definition on separate databases will naturally end up with different cohorts even though they followed the same instruction set.

Librarian

Libraries are run by librarians. In our context, we view librarians as volunteer members of the OHDSI community. Librarians have elevated access to the library’s staging area where reorganization and edits can occur, and they have the authority to pull definitions from the staging area into the official record. They are responsible for maintaining the integrity of the library by ensuring that documentation is complete and sufficient and that the library is properly organized. It is worth noting that librarians don’t pass judgements about the books/chapters themselves; for instance, it would be inappropriate for a librarian to prevent a cohort definition from entering the library on the basis of the validation metric values.

User Personas

To envision how the library might function, we’ve constructed some user personas of types of individuals who may be likely to use the library.

Jane, MD Clinical Researcher

Jane is a rheumatologist who works at an academic medical center. In addition to regularly treating patients with rheumatic diseases, she also performs research as part of her appointment. Her background is primarily clinical, though she has taken coursework in basic statistics in medical school. She has heard of terminologies like ICD9/10 and CPT codes but is not deeply familiar with medical ontologies. She is not experience with software development or health IT, but her medical center has an Atlas installation through which she can run OMOP-based analyses.

Tom, PhD Health Services Researcher

Tom recently completed a PhD in population health sciences with a minor in biostatistics. He is now a postdoctoral researcher at a large research university, with the goal of becoming a professor after he completes his postdoctoral studies. His primary research interest is in improving health outcomes for patients with diabetes mellitus. Tom is well-versed in medical terminologies and has a solid statistical background. He often works with clinicians as part of his research. Tom's lab has an Atlas instance running which he uses sometimes but also does direct SQL and R-based analyses with OMOP.

Marla, PhD Epidemiologist

Marla completed her PhD in Pharmacoepidemiology ten years ago and is a researcher on a pharmaceutical company real-world evidence team. Marla works routinely with phenotype creation and evaluation including review of the medical literature and analyzing performance using the datasets purchased by her company. She is very proficient with epidemiology, statistics, and numerous software packages including SAS, R, SQL, and OHDSI tools such as Atlas.

Juan, PhD Data Scientist

Juan has a PhD in Computer Science and focuses on machine learning algorithms in healthcare. He utilizes the OMOP common data model and vocabularies for his research. He has developed several automated techniques for phenotype generation using OMOP. As a methodologist, he is well-versed with mathematical, computational, and statistical concepts, but he isn't quite as familiar with the medical aspects. Accordingly, he often collaborates with medical professionals for guidance.

Elena MD, PhD, Regulator at FDA

Elena has a background in internal medicine and has been working at the FDA for 20 years. She is supportive of advancing the quality of real-world evidence-based analytics to improve health safety. She must ensure an extremely high level of rigor in the studies that she uses as evidence in her regulatory work. Elena is interested in the potential of research networks like OHDSI.

Brenda, MD Chief Quality and Patient Safety Officer

Brenda is a pediatric oncologist at a children's medical center, where she now leads a clinical quality improvement team. She and her team look to data-driven solutions to improve the operations of their facility, including the reduction of response times, readmissions rates, and frequency of adverse events. Brenda's hematology background gives her thorough expertise in laboratory tests and procedures. She relies on her team to execute analytic tasks, which they perform using OHDSI tools.

Chan, MD, OHDSI Researcher

Chan is a cardiologist and is studying OHDSI in Korea. He conducted multinational retrospective research through OHDSI for comparing first-line combination treatment in hypertension. He is familiar with OHDSI tools such as ATLAS, OHDSI R packages and has been developing some of OHDSI R tools, too. He can validate defined golden standard phenotypes with Korean databases.

Rebecca, BS, Research Coordinator

Rebecca is a research coordinator working with a liquid tumor group at an academic medical center. She is working on her MS in biostatistics and is somewhat computationally inclined, but more familiar with clinical trials than with data from claims or the electronic health record. The principal investigator of her lab has suggested that she familiarize herself with the resources available through the medical center's participation in the OHDSI consortium, but doesn't have the time or energy to provide more detailed guidance. Her goals include both boosting clinical trial accrual by developing phenotypes that accord with trial inclusion and exclusion criteria and assessing the feasibility of new investigator-initiated retrospective observational studies by translating a clinician's thoughts into a computable phenotype and determining how many patients potentially meet the criteria.

Andrew, PhD, Research Faculty

Andrew is a health services researcher and informatician at an academic medical center. He conducts and contributes to original research in several areas, advises on and helps oversee research informatics and analytics development, and represents research interests in governance of health system IT. He has a strong interest in advancing the state of the art in phenotyping within and across institutions for trial cohort identification, observational CER, predictive modelling, and for analyses of data and knowledge bases that span clinical, omics and other basics biological process domains.

User Stories

With the above user personas, we consider some user stories in how they might interact with the library.

Selecting a Rule-Based Phenotype

(Tom, PhD Health Services Researcher)

In his latest research effort, Tom has constructed a pamphlet and other educational materials that he believes will help patients overcome the 10 most common challenges for achieving adequate self-management. An earlier pilot study showed promising results, and now Tom wants to assess more rigorously whether his materials have a measurable impact on care. He has just received IRB approval to carry out a system-wide delivery of his intervention. Tom wants to begin identifying patients who are eligible for his intervention by using his institution's electronic health record. He knows that conceptually he wants to target "diabetic patients" but is not immediately sure about how to systematically identify such patients in his institution's EHR.

As a member of the OHDSI community, Tom is looking to identify a robust diabetes phenotype definition. He navigates to the OHDSI webpage and spots an eye-catching link to the OHDSI gold standard phenotype library, where he navigates to next. While browsing the library, he finds many variants of acceptable diabetes definitions and metrics associated with each choice.

Since the delivery of Tom's intervention to an individual who does not actually have diabetes would be fundamentally harmless, Tom is more concerned about false negatives (missing someone who has diabetes) than he is with false positives (delivering the brochure to someone who does not have diabetes). Hence, he chooses a gold standard diabetes definition that has been shown to have a high sensitivity and a moderate specificity. Tom is pleased to see that this definition has been validated on several sites much like his own and uses a rule-based approach he is familiar with. Using the provided phenotype definition and instructions, he successfully obtains a group of patients to whom he will direct his intervention to in his study.

Selecting a Computational Phenotype

(Brenda, MD Chief Quality and Patient Safety Officer)

As part of her team's quality improvement initiative, Brenda is interested in reducing 30-day readmission rates at her facility. She wants to begin by identifying those who are at highest risk of readmission for the purposes of intervening early on in the index admission. However, designating a patient as "high risk" is usually too complex a situation for standard rule-based phenotype derivations.

When browsing the OHDSI gold standard phenotype library, Brenda is pleased to see that there are computational phenotypes derived through APHRODITE that offer ways to find at-risk patients. Brenda finds a relevant phenotype definition that was derived from a machine learning model trained on one site and validated at 2 others; this model identifies patients who fall in the top decile of readmission risk. Brenda reviews the variables required to construct this definition and finds that the requirements would be satisfied via her facility's Atlas instance. Additionally, she reviews the 4 key metrics (sensitivity, specificity, PPV, and NPV) and finds their values to be suitable for her needs. After receiving IRB approval, Brenda instructs her team to apply the computable phenotype definition to determine how many patients would potentially be eligible for a readmission-reducing intervention.

Adding a New Phenotype

(Jane, MD Clinical Researcher)

Jane is interested in studying the effects of disease-modifying anti-rheumatic drugs (DMARDs) on patients with rheumatoid arthritis. Specifically, she is interested in whether they perform better than nonsteroidal anti-inflammatory medications (NSAIDs) in terms of adverse patient health outcomes. Jane decides to write a proposal for a study.

As she is designing her study, Jane searches for a rheumatoid arthritis phenotype definition in the OHDSI gold standard phenotype library but finds none listed. After some research, she devises a cohort definition that she believes would suit her and the larger community when attempting to identify patients diagnosed with rheumatoid arthritis. Accordingly, she would like to see this definition added to the OHDSI gold standard phenotype library.

Jane submits her new cohort definition using the OHDSI phenotype submission form. After submission of the form, it goes to the OHDSI phenotype working group for review, where it is reviewed at a monthly meeting, along with other submissions. The group consults with other rheumatology experts, who suggest some revisions to the definition. Through an iterative feedback loop, the proposal eventually transitions from its initial submission to the staging area, and finally becomes part of the official record after collective agreement of the revisions.

Validating an Existing Phenotype

(Chan, MD OHDSI Researcher)

After Jane's rheumatoid arthritis cohort definition becomes part of the official record (see above), Jane reaches out to the OHDSI community to see if others can validate the definition at their sites.

Chan sees this and agrees to validate it at his site. After performing a chart review, Chan navigates to the validation submission form. He selects Jane's cohort definition from the list and proceeds to fill out the form. After he is finished, he submits his data, where it goes to the staging area. It is reviewed by a librarian and inducted into the official record. The next time anyone view's Jane's phenotype, they will see that Chan validated it and the results he obtained at his site.

Phenotype Admissibility

(Marla, PhD Epidemiologist)

Marla is conducting a study at her site to evaluate adverse drug response rates on a subset of patients within one of her site's purchased datasets. One aspect of her cohort definition relies on patients having adequate blood function. As a familiar user of the OHDSI gold standard phenotype tool, she uses the search tool to find phenotype definitions related to "blood". She finds a relevant definition based on hematocrit, corrected platelet count, blood pressure, and few other hematologic metrics.

On the phenotype evaluation page, Marla finds a warning that the definition should only be applied if test completeness rates for the corrected platelet laboratory test are high in the sample population, as the integrity of this particular phenotype definition heavily relies on the completeness of this lab test in the data. Marla investigates her source dataset and finds that the test completeness rates are relatively low for the test in question; hence, the definition is not admissible for her data. In order to avoid incorrect application of a gold standard phenotype definition, she cautiously decides against using this particular phenotype definition and instead finds an alternative one that she believes is conceptually similar and more viable for her site.

Adding a Computational Phenotype

(Juan, PhD Data Scientist)

Using Aphrodite, Juan has derived a random forest classifier that can designate individuals as being high or low risk for having a breast cancer diagnosis within 5 years. He has validated the algorithm at his own site and would like to see it introduced into the OHDSI gold standard phenotype library.

Juan navigates to the OHDSI phenotype submission form, where he fills out the documentation for his definition and the validation results he obtained. A nuance of the evaluation in this context is that the validation should specify whether the classifier was used “as is” and/or whether a derived site-specific classifier was trained based on the submitted phenotype keywords.

After submission, the data goes to the staging area where it is reviewed and accepted by a librarian to become part of the official record.

Accessing Phenotypes with Known Performance for Predictive Modeling

(Andrew, PhD Research Faculty)

An analytics platform Andrew is helping to develop will guide investigators to associated areas at different levels of explanation (clinical, omics, chemical). The analytics supported by the platform suggest hypotheses at different levels of explanation than the area where the investigation began. The platform will be used by investigators at multiple institutions. The clinical level entities that serve as the start or end point in an analysis need to be machine readable, clinically coherent, reproducible across institutions, and validated by clinical experts. The clinical data operated on by the analytics are in an OMOP data store.

An investigator has characterized biological pathways associated with mitochondrial metabolism changes in response to tyrosine kinase inhibitors (TKIs). She wishes to use the platform to gain insight into the potential clinical implications of these pathway differences by examining patterns in the clinical profiles of individuals who receive TKI-containing treatments.

To support this analysis, the platform uses knowledge bases to search for and access machine-readable phenotypes of indications for TKI-containing treatments in the OHDSI gold standard phenotype library. The metadata in the library associated with the retrieved phenotypes enables diagnostics on the analytic results that account for the phenotypes’ quantified validity (performance metrics) and the extent to which associated data quality requirements are met in the clinical data store.

Disseminating Phenotype Usage in a Publication

(Chan, MD OHDSI Researcher)

Chan has just learned that his research article was accepted for publication in a prestigious journal. In his article, Chan utilized a cohort definition from the Gold Standard Phenotype Library. He wants others to be aware of the documented use of this cohort definition.

Chan navigates to the OHDSI Gold Standard Phenotype Library. He navigates to the Citation Submission page and selects the phenotype he used. He then adds a citation for his publication and clicks a button to submit the data.

The data go to a staging area for librarian review. A few days later, a librarian reviews it, accepts it after verification, and it becomes part of the official record. Anyone who looks up this particular cohort definition in the future will see that it has been used in Chan's publication.

Restoring a Draft Generically

(Andrew, PhD Research Faculty)

Andrew is working on submitting a cohort definition to the Gold Standard Phenotype Library. He is leaving soon and wants to resume his work tomorrow. Andrew clicks a "save" button on the interface, and he receives confirmation that his work has been saved.

The next day, Andrew logs in and sees that his draft is available for editing. He clicks on it and resumes working on the submission. Andrew repeats this process until he finishes and submits his definition a few days later.

Restoring a Draft via a Sharable Link

(Andrew, PhD Research Faculty)

Andrew is working on submitting a cohort definition to the Gold Standard Phenotype Library. He will soon be going on vacation for two weeks and realizes he will not be able to complete the submission until he gets back.

Andrew messages Juan, who is a co-author on the submission to see if he can help to complete the submission. Juan agrees, and Andrew saves his draft, receiving a sharable link. Andrew sends the link to Juan, and Juan uses it to load the in-progress cohort definition submission. Juan finishes the definition and submits it to the librarians.

Architecture

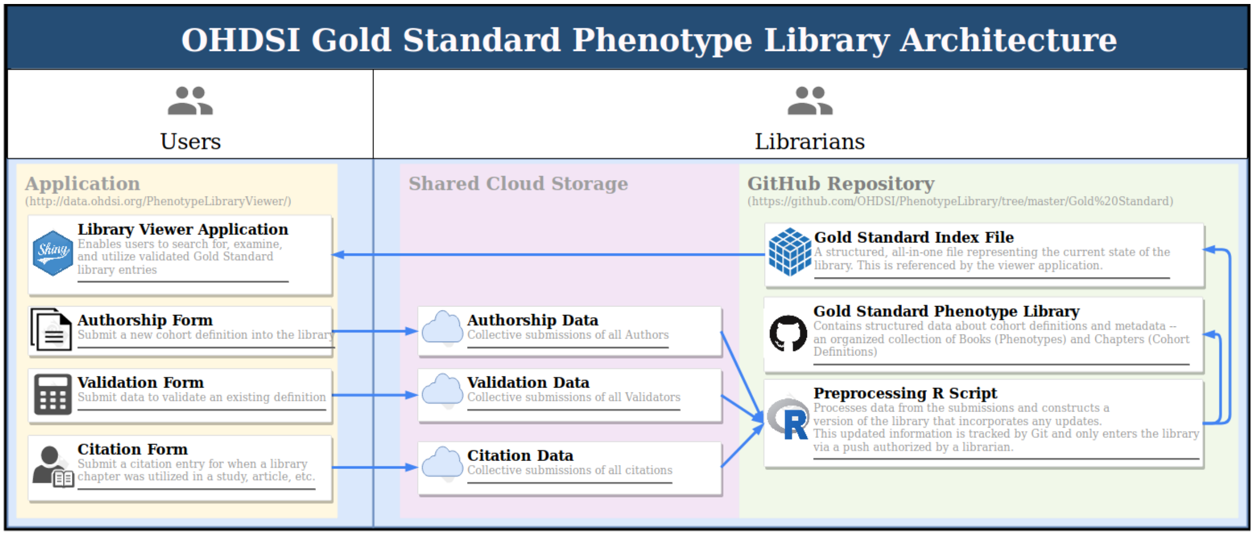
The library consists of Users, Librarians, Authors, and Validators:

- **User**: Searches for, examines, downloads, and utilizes Gold Standard cohort definitions
- **Librarian**: Maintains the organization of the library and completeness of incoming data
- **Author**: Submits a new cohort definition to the library
- **Validator**: Submits a new validation set of an existing definition to the library

An individual can assume any or all of these roles. We envision different levels of involvement and interaction for each of these actors depending on the circumstances.

Overview

We anticipate using a structure as diagrammed below:


OHDSI Gold Standard Phenotype Library Architecture
Users
Librarians
Application
(http://data.ohdsi.org/PhenotypeLibraryViewer/)
Library Viewer Application
Enables users to search for, examine, and utilize validated Gold Standard library entries
Authorship Form
Submit a new cohort definition into the library
Validation Form
Submit data to validate an existing definition
Citation Form
Submit a citation entry for when a library chapter was utilized in a study, article, etc.
Shared Cloud Storage
Authorship Data
Collective submissions of all Authors
Validation Data
Collective submissions of all Validators
Citation Data
Collective submissions of all citations
GitHub Repository
(https://github.com/OHDSI/PhenotypeLibrary/tree/master/Gold%20Standard)
Gold Standard Index File
A structured, all-in-one file representing the current state of the library. This is referenced by the viewer application.
Gold Standard Phenotype Library
Contains structured data about cohort definitions and metadata -- an organized collection of Books (Phenotypes) and Chapters (Cohort Definitions)
Preprocessing R Script
Processes data from the submissions and constructs a version of the library that incorporates any updates.
This updated information is tracked by Git and only enters the library via a push authorized by a librarian.

In this framework, there is a circular pattern to the flow of data. Users submit new cohort definitions, validation sets, and citation events to the librarians to become part of the official record. The proposed additions to the library are reviewed by the librarians, and if approved, are incorporated as part of the official library.

Application

This application is the interface that allows users to find and submit data. They can search, filter, sort, and otherwise manipulate data to locate cohort definitions they're looking for. After selecting a book and a chapter, they can view all information about the definition in an organized fashion and download it if they choose. The application also allows for submission of a new cohort definition, validation set, or citation. Each of these options has a distinct form with required data elements.

Shared Cloud Storage

After data are submitted, it goes to a "staging area" where it can be reviewed by the librarians. This is also where unfinished submissions are saved as drafts. Here, the library's contents can be edited or rearranged as necessary, but the librarian is expected to keep modifications to a minimum and to actively work with the submission author(s) if sweeping changes are required.

GitHub Repository

After the contents of a submission are reviewed, the librarian will run a script to construct the current state of the shared cloud storage as a Git repository. The librarian can then commit the changes they are sanctioning to the library. This repository is the official record of peer-reviewed library data.

Common Data Elements

This section details exactly what data are to be collected for each mode of submission into the library. They are part of what make the library "Gold Standard", in that they ensure that every entry meets a certain standard of documentation.

Elements marked as "Supplied" are expected to be input by the user. Depending on the context, this input could be in the form of a raw text field, or it could be guided such as a drop down list, upload box, or similar input widget. Elements marked as "Derived" are data elements required for record-keeping in the library but the user needn't specify them directly because they are derived "behind the scenes". For example, the user doesn't need to input the date something was submitted because a timestamp can simply be calculated at the time of submission.

Cohort Definition Submission (New or Existing Book and New Chapter)

The following data would be submitted as a proposal for a new chapter in the library. Since every chapter must reside within a book, the author also has the opportunity to choose from an existing book, or to propose a new book. Books cannot be created on their own since an empty book would have no use.

Table 1. Chapter Submission Data

Characteristic Name	Supplied/ Derived	Characteristic Description	Entry Example
ID	Derived	A unique identifier for the cohort definition: 40-digit SHA1 hash of implementation file	86b9edf712c7870615082efa4e08 6e2b2945ee63
File_Upload	Supplied (Upload)*	Raw implementation file uploaded to verify what definition is being validated	cohort_definition.json
File_Link	Derived	Link to download raw implementation file	URL to "raw" file on GitHub: Ex: Cardiovascular Disease
Book_Name	Supplied	Desired location for the chapter	Sjogren's Syndrome
Book_Description	Supplied	Clinical description for the phenotype	
Chapter_Name	Supplied (Text)*	Short, general title of the entry	Sjogren's Syndrome - highly specific (captures chapter-level intent)
Short_ Chapter_ Description	Supplied (Text)*	A concise human-readable description of what the cohort definition is.	1x stroke (ischemic or hemorrhagic) condition record on or the day before an inpatient or ER visit
Long_ Chapter_ Description	Supplied (Text)	An uploaded document. (txt/docx/pdf)	This should include: - Intended Use - Development Methodology - Was IRB approval obtained? - Please cite any references you used in building the cohort definition - Flowchart/Visualization Open question: Should there be a template for this document?

Long_Chapter_Text	Derived (Text)	Derived searchable text, e.g. converted docx/pdf file + JSON code	
Author_Name	Supplied	Author(s)	John Doe
Author_Affiliation	Supplied	Affiliation(s)	Georgia Institute of Technology
ORCID ID	Supplied (Text)	ORCID ID of the author(s)	0000-0001-1234-5678
Date_Of_Submission	Derived	This is the date the phenotype was submitted to the library.	March 21, 2019
CDM_Version	Supplied (Can this be Derived?)	The version of the CDM that was used for developing the phenotype From this, it is an outstanding question whether we can derive the *minimum* version for which the definition is compatible.	v6.0
Vocabulary_Version	Supplied (Open Question: Can this be Derived?)	The version of the vocabulary that was used for developing the phenotype	v5.0
Modality	Supplied (Dropdown)	This identifies the class of phenotype: Rule-Based or Computable	Rule-Based
Therapeutic Area(s)	Supplied (Group Checkbox)	TA Categories from CDISC: https://www.cdisc.org/standards/therapeutic-areas	Diabetic Kidney Disease ; Kidney Transplant
Type	Supplied (Dropdown)	Indication, Outcome, Ingredient, Treatment, Generic	Treatment
Provenance_ID	Derived	ID(s) of any cross-referenced phenotype(s)	cfa99770da20b404a28c3241defec892a7c542c3, 50bd4fe1efb67b3f44c11143b45fa0d18588ddf7
Provenance_Name	Supplied (Dropdown)	Name(s) of any cross-referenced chapters.	
Provenance_Reason	Supplied (Text)	Reason the phenotype(s) above are cross-referenced	E.g. New version, Definition Derivation
Age_Subgroup	Supplied (Dropdown)	Age group that the phenotype algorithm is intended for:	Pediatric

		Pediatric, Adult, Senior, NA	
Tags	Supplied	Prompt with existing tags while typing, but allow for users to enter new ones as well.	e.g. Sensitive, Specific, Incident, Prevalent
Supplemental_ Documentation	Supplied (Upload)	A catch-all category for information the submitter wishes to provide	
Additional_Author_ Comments	Supplied (Text)	A place for the author(s) to add additional comments, if desired	
(All Validation Data Elements)		The author has an opportunity to add a validation set using his/her own site at the time of submission	See the next section for validation data elements.

Validation Submission

Purpose: The following data would be submitted as an evaluation of the performance of a cohort definition already existing in the library. Validation could be done via manual chart or by using software (i.e. PheValuator, Aphrodite). For any methodology, an explanation of the validation process used is required.

It is understood that not all metrics may be available. For example, in the case of a manual chart review, the false negatives may not be provided, since charts are not usually reviewed for those who the algorithm didn't choose. It is also understood that values may not be integral in the metric fields. For example, PheValuator allows for being fractionally right or wrong, based on the predicted probability of being positive/negative, as compared to the ground truth. The submitter should submit what they have, and the viewer will sort out what it can display on its own.

Table 2. Validation Submission Data

Characteristic Name	Supplied/ Derived	Characteristic Description	Entry Example
ID	Derived	ID of the phenotype being validated	86b9edf712c7870615082efa4e086e2b2945ee63
File_Upload	Supplied (Upload)*	Raw implementation file uploaded to verify what definition is being validated	cohort_def.json
ORCID ID	Supplied	ORCID ID of the validator	0000-0001-1234-5678
Validator_Name	Supplied	Validator(s)	John Doe

Validator_ Affiliation	Supplied	Validator Affiliation(s)	Georgia Institute of Technology
Method	Supplied	Chart Review, PheValuator, Aphrodite, Other	PheValuator
Data_Description	Supplied	A description of how the measurement error was estimated.	To quote Patrick: "For chart adjudication, we'd like to know what sample of charts were adjudicated, which charts and how many, who were the adjudicators, what information was used for adjudication, etc. For PheValuator, we'd like to know what inputs were used in the process, including specification of the noisy positive and noisy negative labels used to train the probabilistic gold standard."
CDM_Version	Supplied	The version of the CDM that was used for validating the phenotype	v6.0
Vocabulary_ Version	Supplied	The version of the vocabulary that was used for validating the phenotype	v5.0
True_Positives	Supplied	The number of instances where the algorithm predicted phenotype membership, and it was found to be correct	100
True_Negatives	Supplied	The number of instances where the algorithm predicted phenotype non-membership, and it was found to be correct	100
False_Positives	Supplied	The number of instances where the algorithm predicted phenotype membership, but it was found to be incorrect	100
False_Negatives	Supplied	The number of instances where the algorithm predicted phenotype non-membership, but it was found to be incorrect	100

Timing	Supplied	Misclassification of the timing of cohort entry/exit Outstanding question: How to estimate this generally?	

Citation Submission

The purpose of the citation submission is to allow for documenting published use cases of a cohort definition in a study, book, article, etc.

Ideally, we would like to be able to identify and/or specifically link to T, C, O's from other studies. If this could be done automatically (by ORCID ID referencing, for example), that would be preferable. However, automation of citations is not currently a *requirement* but rather a "would be nice to have" feature.

For now, citation events will need to be submitted manually. To prevent this process from being too cumbersome, this form will be kept very simple. It allows for selection of the cohort definition to be cited and allows for a citation (in any standard format) to be added, along with a link to the resource.

As with all submitted data, the cited use data goes to a staging area to be reviewed by a librarian prior to becoming part of the official record. Then, whenever that cohort definition is referenced, the citation data about it become visible.

Table 3. Citation Submission Data

Characteristic Name	Supplied/ Derived	Characteristic Description	Entry Example
ID	Derived	ID of the phenotype being referenced	86b9edf712c7870615082efa4e086e2b2945ee63
Chapter_Name	Supplied (Dropdown)	Chapter (Book) that is being nominated for citation	Type 2 Diabetes Mellitus (Diabetes)
Citation_Data	Supplied	Copy/Pasted citation of the of the resource which has utilized the cohort definition	Doe John, Doe Jane. Example article title. JAMA;100:1-10.
Citation_Link	Supplied	URL to the resource that has utilized	https://www.link_to_paper.com

Risks

Auth0 Freemium Service

For submitting data to the phenotype library, the Auth0 authentication service is being proposed. This service operates under a "freemium" model. Under the free plan, 7,000 logins are allowed per month. This is certainly sufficient for the library's initiation and will be for the foreseeable future, but it is possible that this authentication limit could be reached with increased traffic. However, Auth0 does claim that for open source projects, it is possible to get Auth0 for free with their Open Source Program. This may be an avenue to be pursued if 7,000 logins per month becomes problematic.

Google Drive Upload Service

Similarly, uploads to Google Drive are limited. We anticipate that while submissions will have attachments, most of the data are text, and a large number of submissions (tens of thousands) would need to be submitted to get close to the Google data cap of 15GB. Should we get near this cap, we would either need to subscribe to receive more storage (100 GB is currently priced at $2 a month) or find an alternative shared cloud storage staging area for the librarians.

Vulnerabilities

As with any publically available software, there is a non-zero chance that it could get hacked or attacked. However, some safeguard are automatically in place by using the proposed services. With Google Drive, attachments are automatically scanned for viruses:

https://support.google.com/a/answer/172541?hl=en

Authentication also acts as a deterrent, since we will be able to trace who uploaded each file and when. Auth0 allows for blacklisting bad actors when they are encountered, and the staging area acts as a double check to ensure no inappropriate content ends up as part of the final library.

Human Error

Since data are to be entered into the application by individuals, there is an inherent risk of data being entered in incorrectly into the application. To some extent, this can be mitigated; for example, the ORCID ID can be checked to ensure that it was filled in its expected sixteen-digit format, and similar types of regular expression matching could be useful to validate e-mail addresses and other formatted entries. However, there is always the chance that the data itself were improperly entered. If detected, this can be corrected by the librarians at the time of staging so that it never becomes part of the official record. If an error does become part of the official record, it is still possible to overwrite it with a correction, whereby the correction will be documented via GitHub, as would be the case with any other change to the official record.

Development and Deployment

Development

We will operate under a "rapid prototyping" framework, where we will debut an application prototype that fulfills the fundamental requirements, release it to the OHDSI community with an invitation for feedback, and then reiterate on the application to address the feedback, continuing this cycle until there is a reasonable consensus that the application's performance is acceptable for use. Along the way, we anticipate releasing an "alpha" and "beta" version of the application prior to its formal deployment.

Currently, the most up-to-date version of the application is only accessible internally at Georgia Tech. The reasons for this are twofold: 1) Iteration is faster in certain aspects (e.g. new R package installations, authentication configurations, and web traffic routing), and 2) With the submission of data, we need to ensure appropriate security measures are implemented before allowing public-facing access.

Our development environment mimics that of data.ohdsi.org (Shiny Server) so that deployment will be easily transferable.

Deployment

All final code will be publically available with the exception of security-related files (e.g. tokens) that will reside privately on the deployment server. We anticipate the official repository will reside here:

https://github.com/OHDSI/PhenotypeLibrary

We also anticipate the application to be hosted alongside OHDSI's existing Shiny applications, here:

https://github.com/OHDSI/ShinyDeploy

Both locations will allow for any community-backed contributions to be made, as needed.

Glossary

As a working group, the Gold Standard Phenotype Library has established what we mean by various definitions. The term "phenotype", for instance, means different things to different researchers. We accept this, but we also put forth the definitions below for the task of the establishing the library.

Phenotype – A phenotype, as it pertains to observational research using health data, is a pattern of observable characteristics for a set of people for a duration of time. These characteristics can include conditions, procedures, drug exposures, devices, observations, visits, cost information, etc.

Cohort Definition – A phenotype algorithm is a coded set of instructions with the desired intent of identifying members of a phenotype in health data. Each phenotype could have one or more phenotype algorithms (e.g. T2DM broad, T2DM narrow). The instructions could be heuristic (rule-based) or probabilistic. A heuristic based phenotype algorithm consists of rules and one or more concepts sets. A probabilistic phenotype algorithm is implemented using a probabilistic model.

Cohort – A cohort instance or phenotype instance is a set of patients for a duration of time which result from the execution of phenotype algorithm instructions against health data.

Concept Set – A concept set is a list of codes used to find records in the Common Data Model.

(Credit: https://forums.ohdsi.org/t/what-do-the-concept-set-and-vocabulary-mean-in-cohort-definition/2441)

Gold Standard Phenotype Algorithm – A "Gold Standard" phenotype algorithm is one that is designed, evaluated, and documented with best practices. The notion of "best practice" refers to the idea that the phenotype algorithm was held to specific standards of design and evaluation and meets all of the requirements OHDSI deems necessary in order to be included into the Gold Standard Phenotype Library (these requirements are under development).

Gold Standard Phenotype Library – A library of publicly available gold standard phenotype algorithms meant to enable members of the OHDSI community to find, evaluate, and utilize community-validated cohort definitions for research and other activities.

Relevant Links

Gold Standard Phenotype Library Working Group Forum

In August of 2018, the requirements development began for the library. In January of 2019, the Gold Standard Phenotype Library Working Group was founded and began meeting regularly. A record of this development is available here:

http://forums.ohdsi.org/t/requirements-development-for-the-ohdsi-gold-standard-phenotype-library/4876

Gold Standard Phenotype Library Working Group Wiki

In addition to a forum thread, this working group has a wiki, which contains links to some presentations given about the library:

http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:gold-library-wg

Documentation Drafts

The Gold Standard Phenotype Library Working Group put together documentation in regularly scheduled group meetings. The first draft of this document was put together here, which is essentially represents a culmination of meeting notes:

https://docs.google.com/document/d/1H_fG94uGhRsY2-aC4j18PTYeIBCXF-tSOSP8m_B4SVE/edit

Phenotype Definition Forum Post

On May 8, 2019, James Weaver posed the question about what a phenotype is in the context of observational research. This discussion guided our development of definitions and book/chapter structure:

http://forums.ohdsi.org/t/what-is-a-phenotype-in-the-context-of-observational-research/6796

APHRODITE

The Gold Standard Phenotype Library houses computable phenotypes. Dr. Juan Banda has developed Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE), a methodology and software for developing computable phenotypes:

https://github.com/OHDSI/Aphrodite/

PheValuator

Validating cohort definitions is an important component of the Gold Standard Phenotype Library. Dr. Joel Swerdel has developed methodology to validate a computable or rule-based phenotype algorithm by checking its performance against extremely specific or sensitive cohorts:

https://github.com/OHDSI/PheValuator/

Interface Prototypes

An early prototype was built and is hosted via a Shiny Server at the following address:

http://data.ohdsi.org/PhenotypeLibraryViewer/