VU VRIJE UNIVERSITEIT AMSTERDAM

2023

OMOP CDM-Enabled Genomics Exploration: Unveiling Population Structure and PRS Associations from the Harvard Personal Genome Project

MASTER'S THESIS IN BIOINFORMATICS AND SYSTEMS BIOLOGY

AMSTERDAM, NETHERLANDS

Abstract

With the increasing availability of detailed and diverse data types associated with patient profiles in health data repositories, a great opportunity is presented for combining these data to allow for more comprehensive representations of patient profiles in large scale health research. In particular, the ability to include genomics data in federated analyses utilizing the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) would be of great benefit. Therefore, this thesis project sought to identify modern genomics tools that could integrate well with the OMOP CDM for research studies leveraging both observational and genomic information. The Google Cloud Life Sciences API was identified to be fit for purpose due to its scalability, availability of an intuitive and open-source variant representation tool, integrated compute environment and excellent documentation. To demonstrate the utility of combining the OMOP CDM with GCP Life Sciences API, population structure analyses and a polygenic risk score (PRS) study were carried out on a study population derived from the Harvard Personal Genome Project. Separate ETL pipelines were applied to the electronic health record data and genomics data from the identified study population. An analysis methodology was developed leveraging the ATLAS cohort building tool alongside the transformed health data in the OMOP CDM and genomics data in Google's BigQuery variant schema. Chronic disease prevalence was analyzed alongside occurrences of clinically significant variants with potential phenotype associations. Lastly, a cohort based PRS analysis was carried out to explore associations between polygenic profiles and occurrences of the Gastroesophageal Reflux Disease (GERD) phenotype. While no clinically significant variants with phenotype associations were found, a significant association between PRS and the GERD phenotype was discovered. In light of the growing interest in leveraging EHR data for comprehensive health research, this study contributes to the ongoing movement towards the integration of genetic and observational health data.

Contents

1. Introduction

As healthcare systems evolve and expand globally along with the capabilities of capturing increasingly detailed records across a host of health data modalities, the opportunity for conducting novel patient-oriented research studies has never been greater. However, this opportunity brings great logistical and infrastructural challenges as disparate health databases are not readily comparable and contain data at a massive scale. Every interaction a person has with a healthcare system generates a record, with these records typically being handled by different systems. In 2018, it was estimated that a in the U.S. alone, 30 billion messages were generated, each linked with a patient record [1]. Historically, these large-scale health data repositories have benefited healthcare research, but have been limited by their distinct data representation, monolithic structure and privacy restrictions [2].

To solve these limitations and allow for large scale research endeavors that utilize healthcare data across multiple sites, common data models have been introduced to standardize the health data repositories. One data model that has seen adoption across the world by large institutions such as the UK Biobank is the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [3][4]. This data model specializes in representing observational medical data (demographics, medical condition occurrences, drug exposure, observation periods, etc.), as well as widely used medical vocabularies. The structure of the CDM is presented in Figure 1.

The OMOP CDM offers a relational table structure that allows for optimized analyses of extremely large databases and is organized in a way that can be quickly understood by a novice researcher. It also is also flexible in its ability to accommodate new data that are not yet explicitly modeled [5]. In addition to the table structure to store the medical data, a foundational component of the model are the standardized vocabularies used to represent all health-related data. There are 111 vocabularies currently supported, such as ICD-10 and SNOMED, 78 of which are adopted from external sources [6].

This model has proven to be effective in enabling large scale federated analyses across multiple data sites and there are a number of recent research studies who have utilized the CDM in population-level and patient-level predictions as well as cancer research [7]. The general process of federated analyses utilizing the OMOP CDM-transformed repositories is displayed in Figure 2.

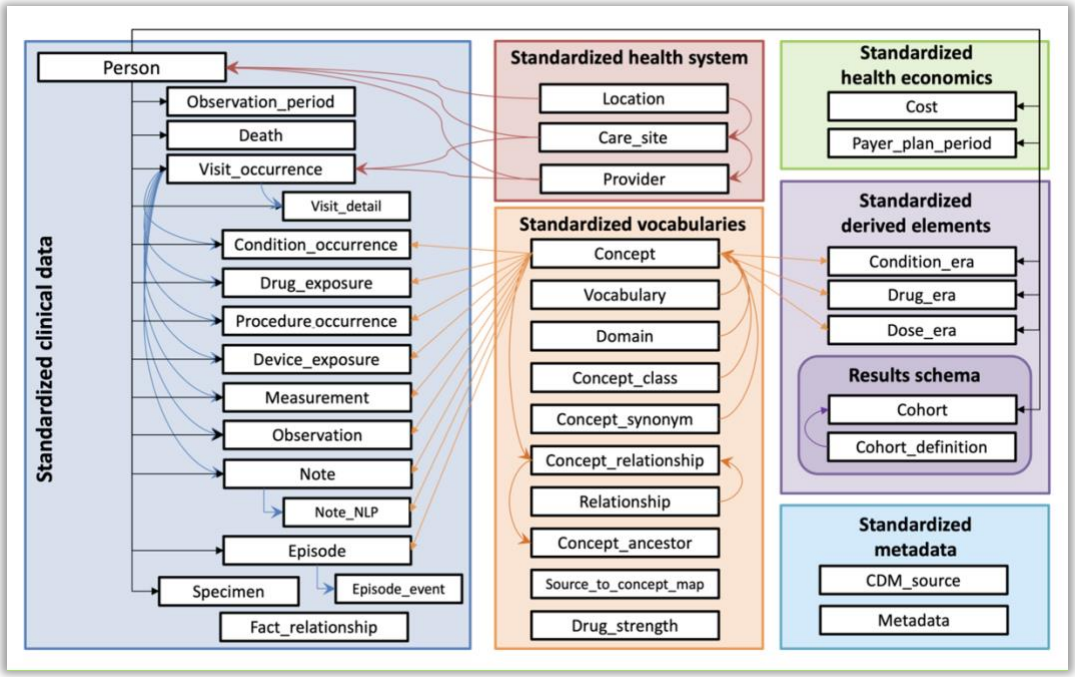


Figure 1. OMOP CDM Table Schema [6]. Note that not all relations are shown.

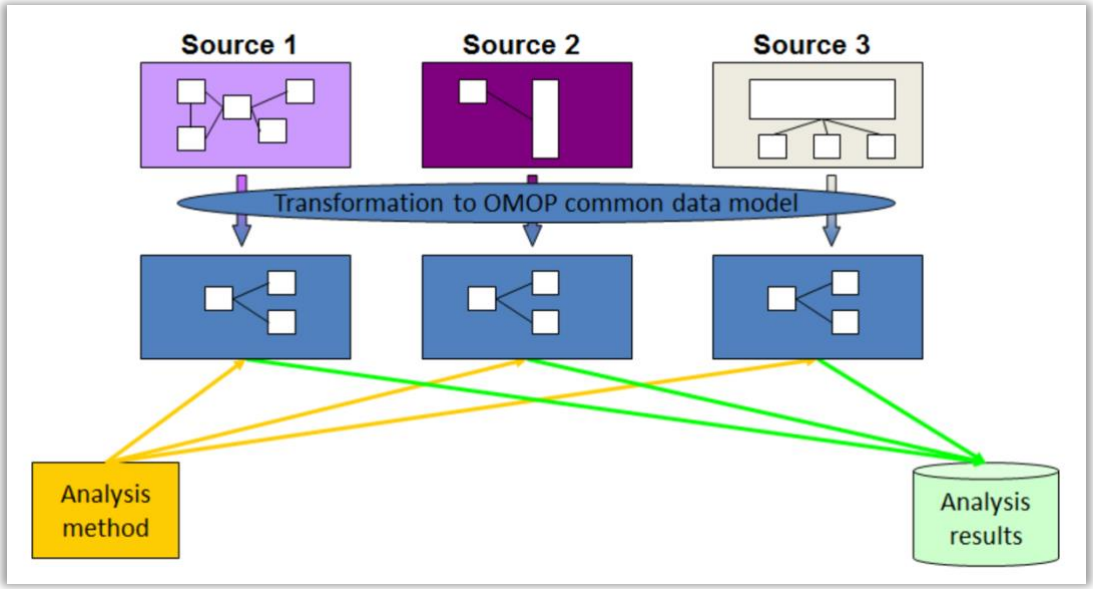


Figure 2. OMOP CDM and Federated Analysis [2]. Example of how standardized research can be conducted using data repositories from multiple sites. Sensitive data can be kept behind secure firewall while analysis results are summarized, extracted and condensed.

While the OMOP CDM has proven its utility in federated analysis for studies concerning patient cohort monitoring, patient level prediction and population-level estimation, it has been less frequently incorporated in studies using the other data modalities commonly available in health record databases - despite its aptness to do so. The ability to integrate these diverse set of data modalities in research is an important effort in the interest of advancing personalized medicine. In particular, genomics data encompassing an individual's complete set of genes makes available, to both clinicians and researchers, insights into inherited traits, response to treatments and disease susceptibilities. By integrating genomics data with observational health data, researchers can build comprehensive models of a patient's profile and discover new patterns between phenotypes and genotypes in cohort populations. With these discoveries in hand, clinicians can provide more tailored medical treatments to patients, leveraging their unique profiles. For example, to develop tailor-made therapies for cancer patients, researchers must have access to genetic variants and their associated pathways together with the clinical information [7]. While there is no common standard for the representation of patient linked genomics data, there are tools and pipelines available that can transform and represent these data in logical and accessible way that could aid federated studies.

Aims

Therefore, this thesis project sought to identify modern genomics tools that could be used in tandem with the OMOP CDM and OHDSI tools to carry out integrative research linking phenotypes to genotypes in a study population. Like the OMOP CDM and associated tools used to transform observational data to the common data model, a suitable software tool for the representation genomics data should have the properties of being open source, scalable, patient-centric, well documented and maintained, and usable by researchers without highly specialized knowledge. To illustrate the utility of combining the selected tool with the OMOP CDM, population structure analyses concerning: 1. Chronic condition prevalence and 2. Pathogenic variant prevalence and their associations to genetic conditions would be carried out. Additionally, polygenic risk score (PRS) analysis of a selected condition prevalent in the study population would be performed. These types of analyses are becoming common practice as genomics data are made more available in health databases and should therefore be achievable with the integration of these tools.

The dataset acquired for this study was derived from the Harvard Personal Genome Project website. First, the phenotypes present in the population were explored, with a focus on chronic conditions. The prevalence of chronic conditions in the study population were compared to rates seen in other larger populations, such as

those reported in the National Health and Nutrition Examinations Survey (NHANES), a US based population. Similarly, the prevalence of pathogenic and likely pathogenic variants was also explored and compared to findings in related works. These population structure explorations would answer the question as to whether the study population possesses different disease patterns and predispositions than expected. Lastly, a polygenic risk score study was undertaken to determine association between genotype profile and gastroesophageal reflux disorder (GERD) occurrence, a condition well represented in the study population. Association tests and a predictive model were developed to indicate whether genotypes profiles are connected to the GERD phenotype and whether such an association is strong enough to be utilized in a predictive model.

Related Work

In a systematic review screening 248 articles for analyses that utilized a combination of OMOP CDM-harmonized data and predictive cancer modelling, only 5 articles were found to have made use of cancer modelling algorithms that extensively took OMOP-transformed data as input [7]. Of the 5 studies matching the search criteria, only two used the OHDSI webtool ATLAS to define patient cohorts used in the study. Given that predictive modelling of cancers is an actively researcher field, it may be extrapolated that the OMOP CDM and OHDSI tools have not been used with great frequency in predictive modelling of disease despite their utility, and that further work should be done in this area.

In a recent paper surveying genome-wide disease-associated genes in adults who possessed highly detailed phenotype profiles, the prevalence of pathogenic (P) and likely pathogenic (LP) variants in their study population were reported along with associations to chronic diseases [8]. ClinVar defines P variants a those that are known or likely to cause a particular disease or condition, whereas LP variants have substantial evidence for association with specific diseases (~90% likely) but may not be as strongly tied as P variants [9]. In the paper, it was found that 11.5% of individuals had associations between their genotype and phenotype profiles, 17.3% of individuals had at least one medically significant genetic finding tied to their variant profile and 43% of individuals with P/LP variants associated with cancer predisposition did not have corresponding family history or phenotypes. In addition, the population was assessed for chronic disease enrichment and no enrichment was reported. Such analyses are useful in elucidating disease frequencies and association patterns in populations. These discoveries are insightful in their own right but can also elucidate potential confounding factors in other research endeavors, especially those involving patient cohort building. Although the Harvard PGP study population acquired for this project does not possess as richly detailed phenotypic profiles as those used in the study survey, it was determined to still be of benefit to reproduce the chronic disease prevalence and P/LP variant by disease category analyses conducted.

The massive genome wide association studies (GWAS) performed over the last decade have revealed the contribution of inherited genetic variants to common complex disorders [10]. It is understood that the underlying genetic associations with these disorders are highly polygenic, with hundreds of thousands of variants exerting minor but cumulative influences on disease risk. While every genetic variant linked to a disease provides valuable insights into relevant genes or biological pathways connected to the disorder, there is also the hope that this genetic data can be employed to estimate disease risk, offering potential benefits in clinical practice. The most common mode of deriving insight from the array of genetic variants that and individual possesses is through the generation of polygenic risk scores (PRS), where a sum of risk alleles for a particular phenotype is calculated using the weighted effect size estimates from GWAS for the phenotype. The PRS can then offer an overall measure of risk for a disorder based on one's genetic variants. The PRS score is acquired by satisfying the following equation:

$$PRS_i = \sum_j \beta_j \times X_{ij}$$

Where the PRS of an individual (i) is equal to the sum of the effect sizes (β) of all variants (j) multiplied by the genotypes (X) of the individuals (i) at variant (j). The beta coefficients are estimated from GWAS summary statistics for the phenotype of interest. X represents the individual genotypes in the target data.

While there is a growing body of research that demonstrates strong correlations between PRS and disease occurrence, the practical usefulness of PRS in a clinical context is yet to be widely established [11]. The strongest evidence for clinical relevance of PRS currently comes from cardiovascular diseases and breast cancer, where risk stratification of those at high polygenic risk has clinical utility [12].

Gastroesophageal reflux disease (GERD) was identified early on as one of a handful of well represented conditions in the Harvard PGP dataset that could be analyses in a PRS context. GERD is caused by gastric acid entering the esophagus and will affect 15-30% of the U.S. population in their lifetime [13]. It is also the major risk factor for Barrett's esophagus (BE) and esophageal adenocarcinoma (EA). In a recent GERD GWAS meta-analysis, 25 independent genome-wide significant loci were identified [13]. Of the alleles discovered to increase GERD risk, 91% were also discovered to increase risk in BE and EA, indicating these alleles play an important role in esophageal related conditions. Despite these known loci, no literature was found regarding predictive modelling or stratified risk modelling of this disease based on its risk alleles, further motivating the PRS analysis undertaken in this project.

2. Methods

Study Data Acquisition – Harvard Personal Genome Project

In order to demonstrate the utility of combining the OMOP CDM with modern genomics software tools to carry out integrative patient studies, a dataset containing patient electronic health (EHR) record information linked to associated genomics data was required. After evaluating different open-source datasets, the Harvard Personal Genome Project (Harvard PGP) site was determined to be fit for purpose. The Harvard PGP is a unique dataset consisting of consenting individuals who elected to make their electronic health record information, and in many cases genomics data, public. The contains thousands of individuals with unique identifiers, hereby referred to as profiles. Each profile may contain some survey data, health record data, and various forms of sequencing or genotyping data. Using custom Python scripts, the Harvard PGP site was scraped and a dataset of 798 individuals was created based on the following criteria:

Genomics Data Providers – Direct-to-consumer (DTC) genotyping data, whole genome sequences and variant called sequencing files are linked to many profile pages. A variety of genomics data providers and file formats were noted. 23andMe and Complete Genomics were two of the most common sources of comprehensive genomics files with complete chromosome sets. Therefore, only profiles with a file produced by one of these providers were included as a part of the study dataset.

Electronic Health Records and Survey Data – Some profile pages contained EHR data, survey data, both or neither. For inclusion in the study dataset, a profile was required to contain at least some EHR data or some survey data.

Harvard PGP Genomics Data

Of the 798 profiles in the study dataset, 202 profiles had their genomes sequenced by Complete Genomics while the other 595 profiles contained DTC genotyping files from 23andMe. These two providers use different approaches for providing genomics data to their customers. Complete Genomics uses a proprietary technology that relies on DNA nanoball sequencing which results in files ranging from 1-2 GB that reports all variations from a reference sequence and indications of stretch regions that are homozygous with the reference, effectively resulting in a compressed whole genome sequence. 23andMe provides genotyping files of predetermined variants based on a genotyping chip array, and are typically around 15 - 30 MB in size [14]. Due to these fundamental differences in genomic data types, unique challenges and considerations were presented that resulted in separate preprocessing strategies being required to make these data comparable for the objectives of this research - later described. All valid genomics files for identified study profiles were scraped, downloaded and checked for integrity using the Python scripts in 'Scraping PGP Harvard' available in section *8. Code Availability*. It was not uncommon for profiles to have both a 23andMe genotyping file and

Complete Genomics sequencing file. In these cases, the 23andMe file was selected over the Complete Genomics file for the given profile. In Figure 3 for an overview of the acquisition and quality assurance process for 23andMe and Complete Genomics files identified on the Harvard PGP site is presented.

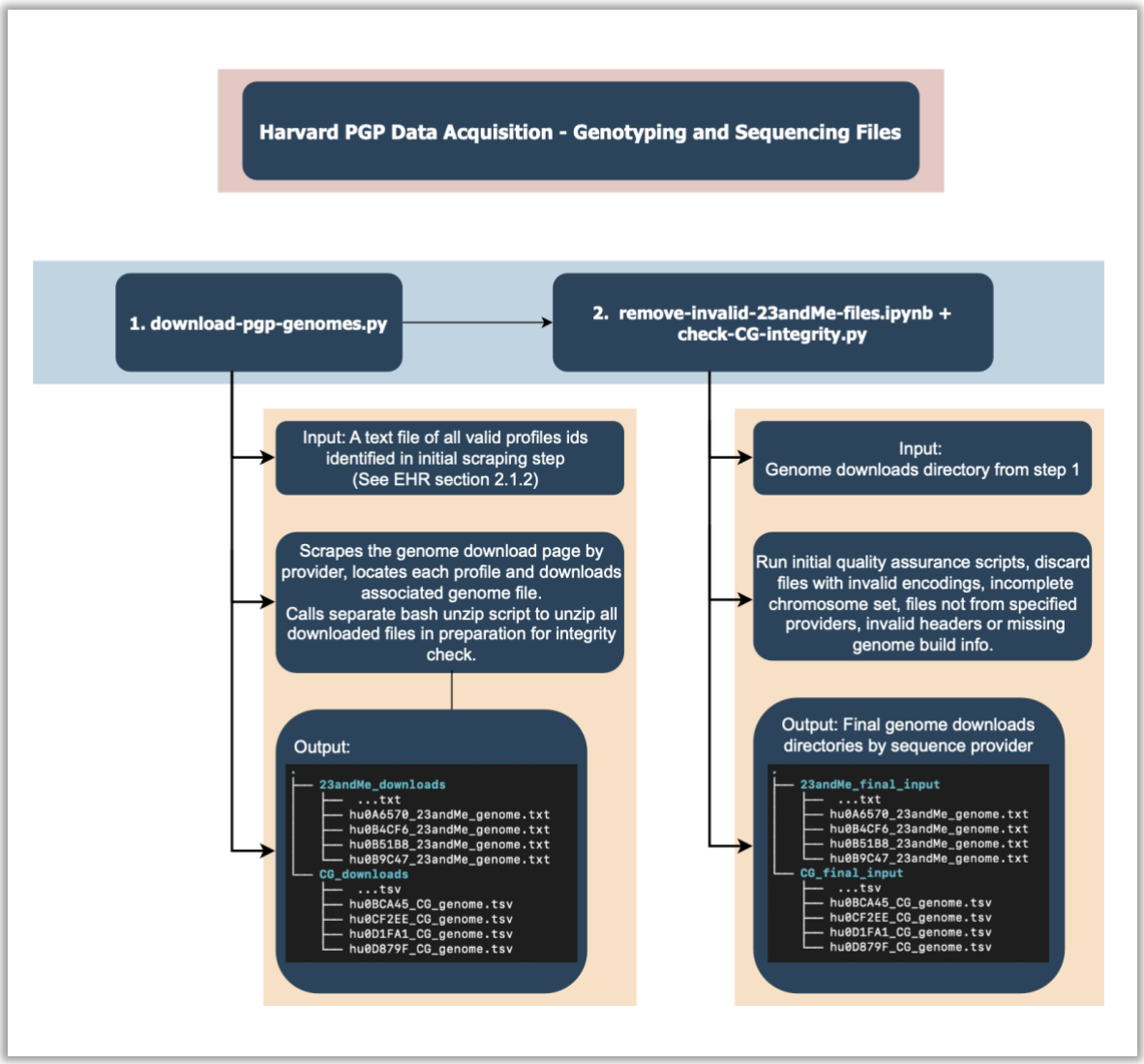
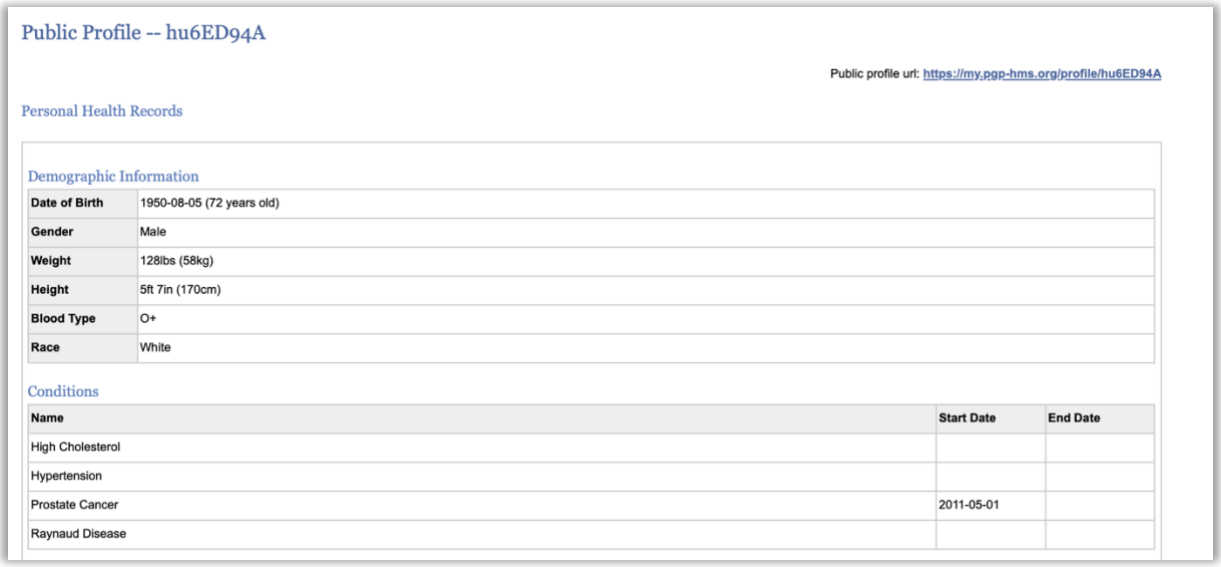


Figure 3. Genomics Data Acquisition Workflow. Source code can be viewed in section 8. Code Availability

Harvard PGP Electronic Health Record Data

Each qualifying study profile contained some medical record or survey data on their Harvard PGP page, as is illustrated in Figure 4. At a bare minimum, demographics information such as age, gender, race/ethnicity and blood type could be acquired for all participants in the study, derived from survey data if a health record was not present. To acquire the data present in these profiles, a two-step scraping process was carried out where viable profiles were first identified, their metadata saved and passed to the second step where the profile pages then had their data downloaded and converted to tables in .csv format. Additional postprocessing steps were also taken to consolidate all demographic and health data into central comprehensive datasets by data type, compatible with the ETL process which would later follow. See Figure 6 for more details.



Public Profile -- hu6ED94A

Public profile url: https://my.pgp-hms.org/profile/hu6ED94A

Personal Health Records

Demographic Information

Date of Birth	1950-08-05 (72 years old)
Gender	Male
Weight	128lbs (58kg)
Height	5ft 7in (170cm)
Blood Type	O+
Race	White

Conditions

Name	Start Date	End Date
High Cholesterol		
Hypertension		
Prostate Cancer	2011-05-01	
Raynaud Disease		

Figure 4. Example participant profile on the Harvard Personal Genome Project Site. Note that only the first two tables in the electronic health record are shown.

PGP Trait & Disease Survey 2012: Skin and Subcutaneous Tissue	Responses submitted 1/18/2013 11:29:15. Hide responses
Timestamp	1/18/2013 11:29:15
Have you ever been diagnosed with any of the following conditions?	Dandruff, Skin tags

Figure 5. Example of PGP Trait and Disease Survey found on a Harvard PGP participant's page. A variety of other survey types are also common.

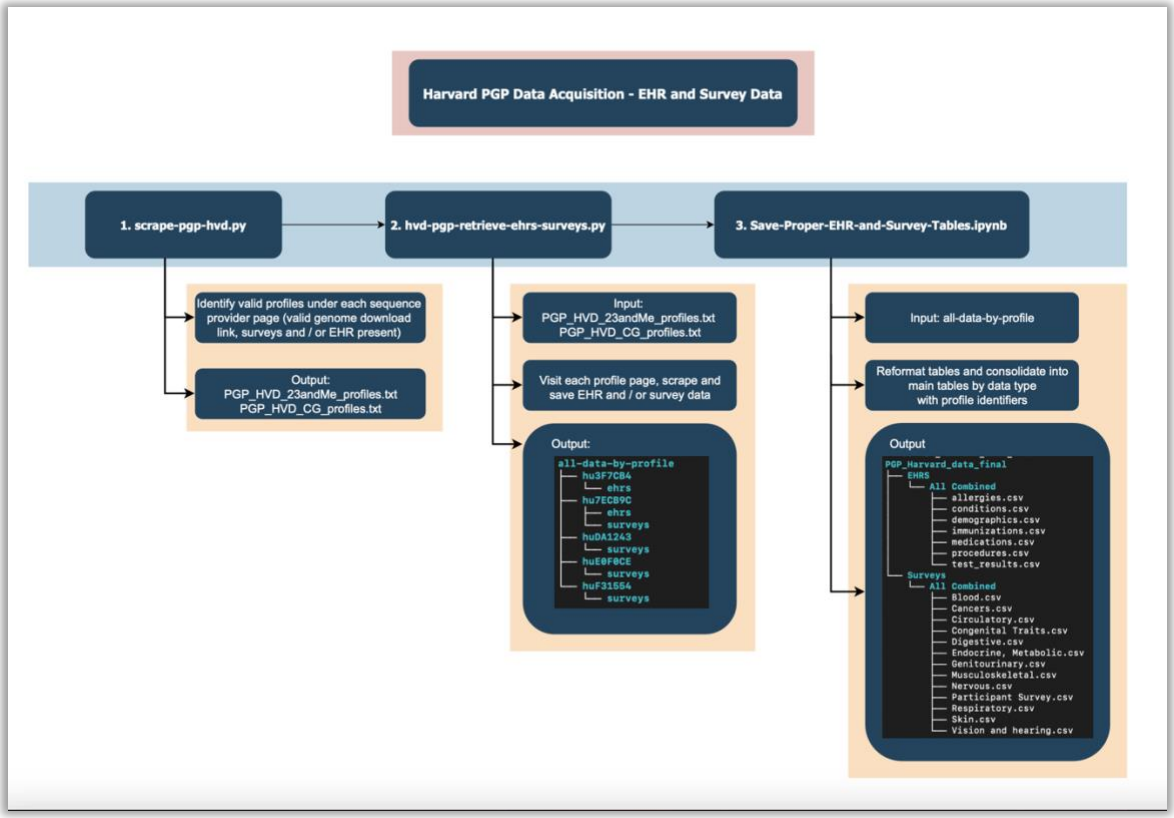


Figure 6. Harvard PGP study participant and health record data acquisition workflow.

Assessment of Variant Analysis Tools

Alongside the initial explorations of the Harvard PGP study data, an investigation was also undertaken to assess and select a software tool for the representation, storage and querying of genomic variants in the study population. An appropriate tool needed to be compatible with VCF files, open source, scalable, well documented and maintained, easy for researchers to use, linkable to individual patients and not constrained by particular classes of variant types, such as strictly cancer variants. The assessment metrics for these criteria are displayed in Table 1. Team members from Open Targets, cBioPortal and Fairspace from The Hyve were consulted for their expertise on associated tools. Independent research on and experimentation with non-OHDSI/Hyve tools such as OpenCGA and Google Variant Transform Tool was also carried out.

✓	**Fulfills criteria**
~	**Partially fulfills criteria or wasn't assessed**
X	**Does not fulfil criteria**

Table 1. Genomics Research Tool Assessment Criteria. See Table 2 for their application.

Study Data Transformation

To prepare the genomics and observational health data for analysis, respective transformation pipelines were applied. The observational health data underwent transformation via the source data to OMOP CDM ETL process; a pipeline involving a suite of open-source tools for generating an ETL outline, mapping source vocabulary to standard vocabularies, applying the transformations and targeting a standardized database schema. The resulting database of transformed observational data can finally be accessed and queried via the open-source web-based analytics platform – ATLAS. The acquired genomics data required a more bespoke transformation process involving a sequence of Python and bash scripts utilizing common bioinformatics packages such as plink, BCFtools, SAMtools, and Picard – run in a distributed workload environment. The resulting genomics data were made available in a way that would allow for easy querying of genomic variants that could be related back to the participant profiles. The processed VCF files were also identifiable via Harvard PGP profile IDs in a way that would be compatible with the cohort based approach used for the PRS analysis.

EHR Transformation: OMOP CDM

The general process for the transformation of the observational health data acquired from the Harvard PGP site followed the schematic illustrated in Figure 7.

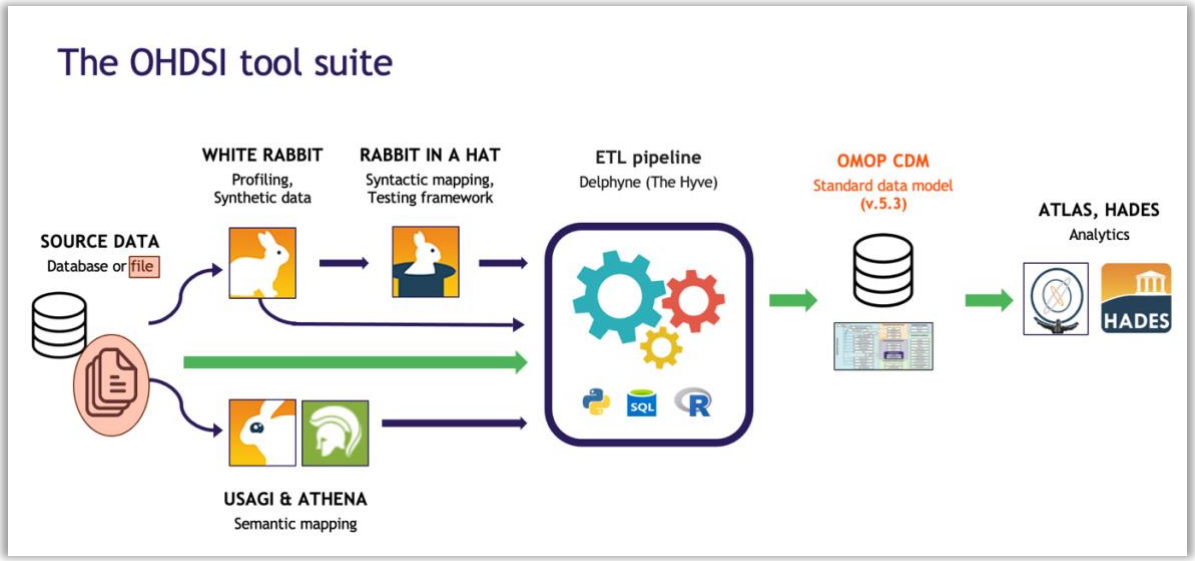


Figure 7. ETL Pipeline process for transforming observational health data to the OMOP CDM using OHDSI tools and Delphyne.

White Rabbit & Rabbit In A Hat

To begin the process, the source data was provided to White Rabbit's scan report tool. This created an Excel file outlining the structure of the source tables together with some summary statistics, in order to understand the dataset in the context of the OMOP CDM. This report was then imported in the Rabbit in a Hat application. This tool is used to create a visual mapping on how to populate the target tables in the OMOP CDM from the source tables accompanied by source field to target field specifications between tables. The specifications can then be extracted and used as the basis for the ETL documentation assisting in the writing of specific transformations in the ETL pipeline step using the Delphyne Python program. Some examples of the results of this process are presented in Figure 8 and Figure 9.

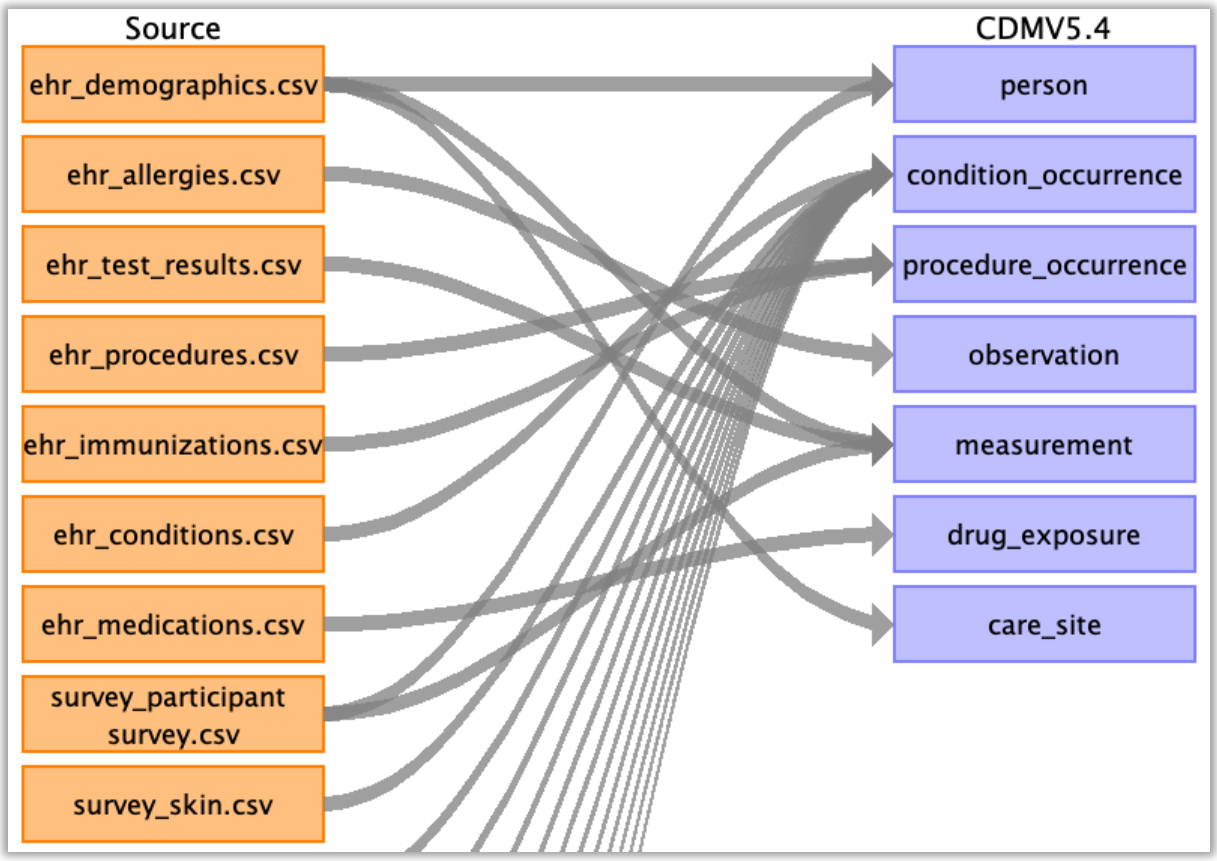
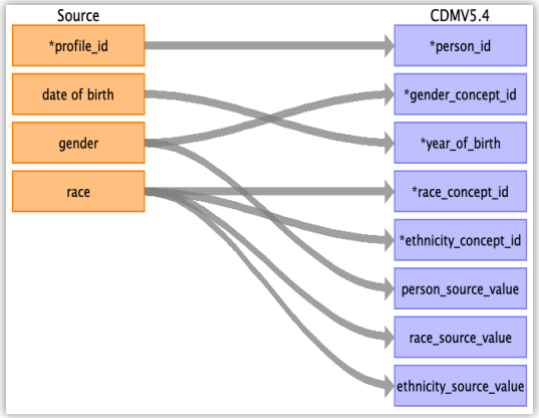


Figure 8. Subsection of Harvard PGP source table to target table (OMOP CDM) mappings generated by White Rabbit and Rabbit In A Hat OHDSI Tools. Note that not all mappings are shown.



Destination Field	Source Field	Logic	Comment
person_id	profile_id		
gender_concept_id	gender		
year_of_birth	date of birth		Contains age and exact date of birth in one string. Need to split
month_of_birth			
day_of_birth			
birth_datetime			
race_concept_id	race		
ethnicity_concept_id	race		
location_id			
provider_id			
care_site_id			
person_source_value	gender		
gender_source_value			
gender_source_concept_id			
race_source_value	race		
race_source_concept_id			
ethnicity_source_value	race		
ethnicity_source_concept_id			

Figure 9. (Left) Source value mappings from Harvard PGP demographics table to target values in OMOP CDM. (Right) Standard target value fields in OMOP CDM person table with transformation notes added via Rabbit In A Hat.

Usagi Vocabulary Mappings

Following the schema mapping process, the Usagi tool was utilized to create semantic vocabulary mappings between the source vocabularies used to describe the Harvard PGP health record information to the standard vocabularies accepted by the OMOP CDM. This required downloading the latest version of OHDSI vocabularies from Athena (see *9. Supplementary links)*. Vocabulary occurrence data for each source table was

derived from the White Rabbit Scan report and used as input in the Usagi program. An example vocabulary transformation is displayed in Figure 10.

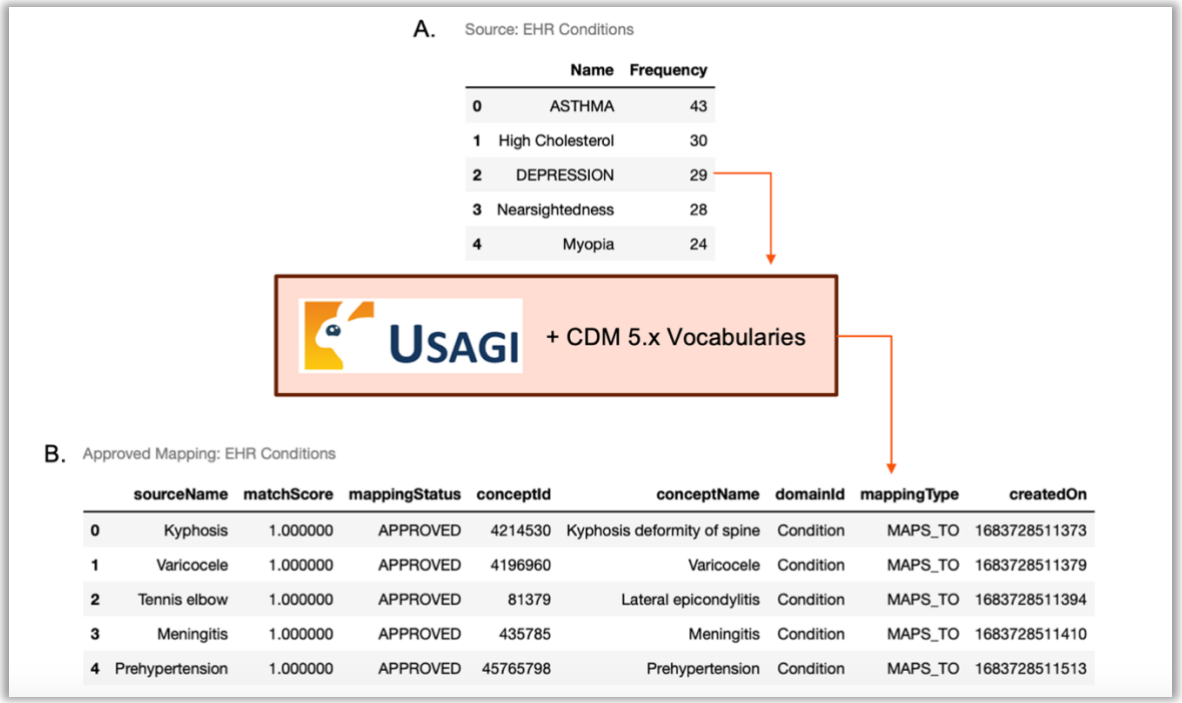


Figure 10. Small-scale example of using Usagi to map standard vocabularies to EHR condition occurrences from Harvard PGP source data.

Delphyne

The final step in in the ETL process was writing and running the source data to OMOP CDM transformations using the Hyve's proprietary Python package: Delphyne. This package was designed to simplify and standardize the transformation process. A transformation was written for each source data table. The transformation script would load in both the source table and vocabulary mapping table created by Usagi. Source value to target value mappings were written to an SQL Alchemy object relational mapper (ORM) which would subsequently be used to load the transformed data into the target PostgreSQL database. An example transformation can be viewed in Figure 11, full code is not made available. In total, 22 transformations were written as seen in Figure 12.

source_ehr_immunizations_to_drug_exposure.py — OHDSI-ETL-HVD-PGP

source_ehr_immunizations_to_drug_exposure.py

```python
from __future__ import annotations
import numpy as np
import pandas as pd
from typing import List, TYPE_CHECKING
from src.main.python.util import get_datetime, generate_person_id

if TYPE_CHECKING:
    from src.main.python.wrapper import Wrapper

df = pd.read_csv('/Users/jerenolsen/Desktop/Usagi/Approved Mappings/ehr_immunizations_rough.csv')
mapping_dict = pd.Series(df.conceptId.values,index=df.sourceName).to_dict()

def source_ehr_immunizations_to_drug_exposure_trans(wrapper: Wrapper) -> List[Wrapper.cdm.DrugExposure]:

    source = wrapper.source_data.get_source_file('ehr_immunizations.csv')
    df = source.get_csv_as_df(apply_dtypes=False, delimiter = ',')

    records = []
    for _, row in df.iterrows():

        r = wrapper.cdm.DrugExposure(
            person_id = generate_person_id(row['profile_id']),
            drug_concept_id = mapping_dict[row['Name']],
            drug_exposure_start_date = get_datetime(row['Date']),
            drug_exposure_start_datetime = get_datetime(row['Date']),
            drug_exposure_end_date = get_datetime(row['Date']),
            drug_exposure_end_datetime = get_datetime(row['Date']),
            drug_type_concept_id = 32817,
            drug_source_value = str(row['Name'])[0:50],
            drug_source_concept_id = 0
            )

        records.append(r)

    return records
```

Figure 11. Example transformation using the Delphyne program. Records of immunizations from HVD PGP source data are transformed, then target the standard drug exposure table in OMOP CDM.

```
class Wrapper(BaseWrapper):
    cdm = cdm

    def __init__(self, config: MainConfig):
        super().__init__(config, cdm)

    def transform(self):

        # Init Caresite #
        self.execute_transformation(build_care_site_table)

        # Populate cdm.Persons #
        self.execute_transformation(source_table_ehr_to_person_trans)
        self.execute_transformation(source_table_surveys_to_person_trans)

        # EHR transformations #
        self.execute_transformation(source_ehr_allergies_to_observation_trans)
        self.execute_transformation(source_ehr_conditions_to_condition_occurrence_trans)
        self.execute_transformation(source_ehr_immunizations_to_drug_exposure_trans)
        self.execute_transformation(source_ehr_medications_to_drug_exposure_trans)
        self.execute_transformation(source_ehr_test_results_to_measurement_trans)
        self.execute_transformation(source_ehr_procedures_to_procedure_occurrence_trans)
        self.execute_transformation(source_ehr_demographics_to_measurement_trans)

        # Survey Transformations #
        self.execute_transformation(source_survey_blood_to_observation_trans)
        self.execute_transformation(source_survey_cancer_to_observation_trans)
        self.execute_transformation(source_survey_circulatory_to_observation_trans)
        self.execute_transformation(source_survey_congenital_traits_to_observation_trans)
        self.execute_transformation(source_survey_digestive_to_observation_trans)
        self.execute_transformation(source_survey_endocrine_metabolic_to_observation_trans)
        self.execute_transformation(source_survey_genitourinary_to_observation_trans)
        self.execute_transformation(source_survey_musculoskeletal_to_observation_trans)
        self.execute_transformation(source_survey_nervous_to_observation_trans)
        self.execute_transformation(source_survey_respiratory_to_observation_trans)
        self.execute_transformation(source_survey_skin_to_observation_trans)
        self.execute_transformation(source_survey_vision_hearing_to_observation_trans)
```

Figure 12. All Transformations between source Harvard PGP observational data and OMOP CDM.

Genomics Data Transformation: Pipeline Process

As previously noted, the genomics dataset acquired from the Harvard PGP site consisted of 202 assembly variations files sequenced by Complete Genomics and 595 genotyping files produced by 23andMe. To make these files comparable, a pipeline process was developed to convert the files to the same file format and genome build, perform genome imputation using a common reference panel, and apply data filtering where appropriate. Quality checks, such as concordance measures, were also made to give an indication of reliability of the chosen processing strategy. See Figure 13 for general workflow of the genomics pipeline described in the following sections.

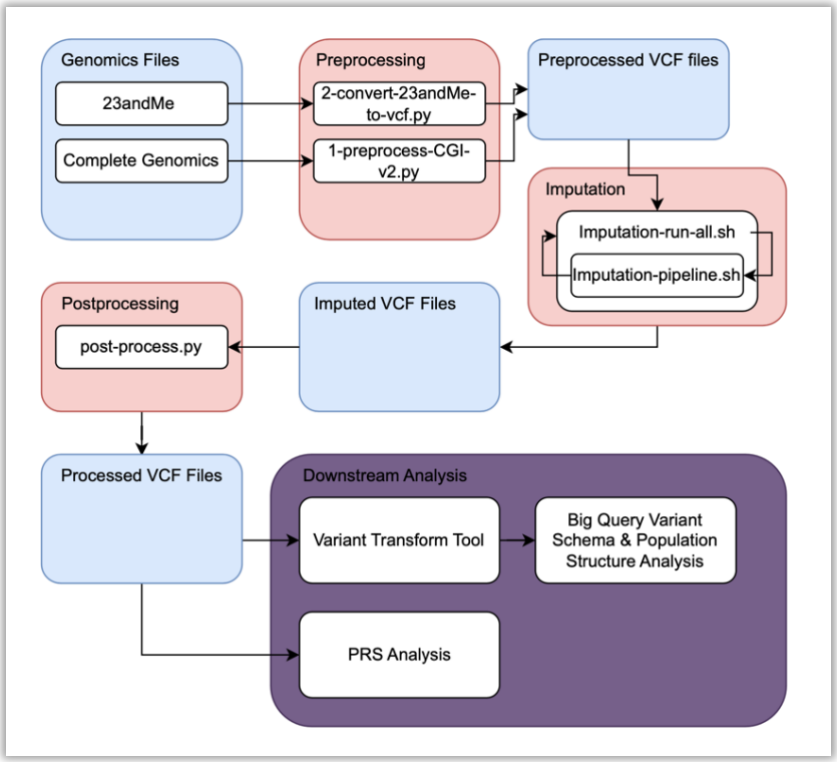


Figure 13. General steps of the pipeline used to process all genomics files from Harvard PGP. Downstream analyses also included.

Preprocessing

Two separate preprocessing scripts were utilized to transform 23andMe and Complete Genomics files into variant call format (VCF). While SNPs of MAF >= 5% would be the basis of the PRS study and variant analyses, this filter was not applied during preprocessing in order to avoid imputation information loss in the following imputation step [15]. Following the execution of these Python scripts, all 798 genomics files were compatible with the remaining pipeline steps.

Preprocessing - 23andMe

The 23andMe files used in this study were represented in tab-separated (tsv) format, possessing external identifier information (rsid), chromosome number, position and genotype. To make these files usable in downstream analyses concerned with the genomic variants, a Python script was used to convert the files to VCF, perform filtering and genome build lift-over where necessary. This process is outlined in Figure 14. For the conversion of the tsv files to VCF, appropriate reference genome FASTA files were used depending on the build of the file. Of the 595 23andMe files, 210 were of build GRCh36, with the remaining being GRCh37.

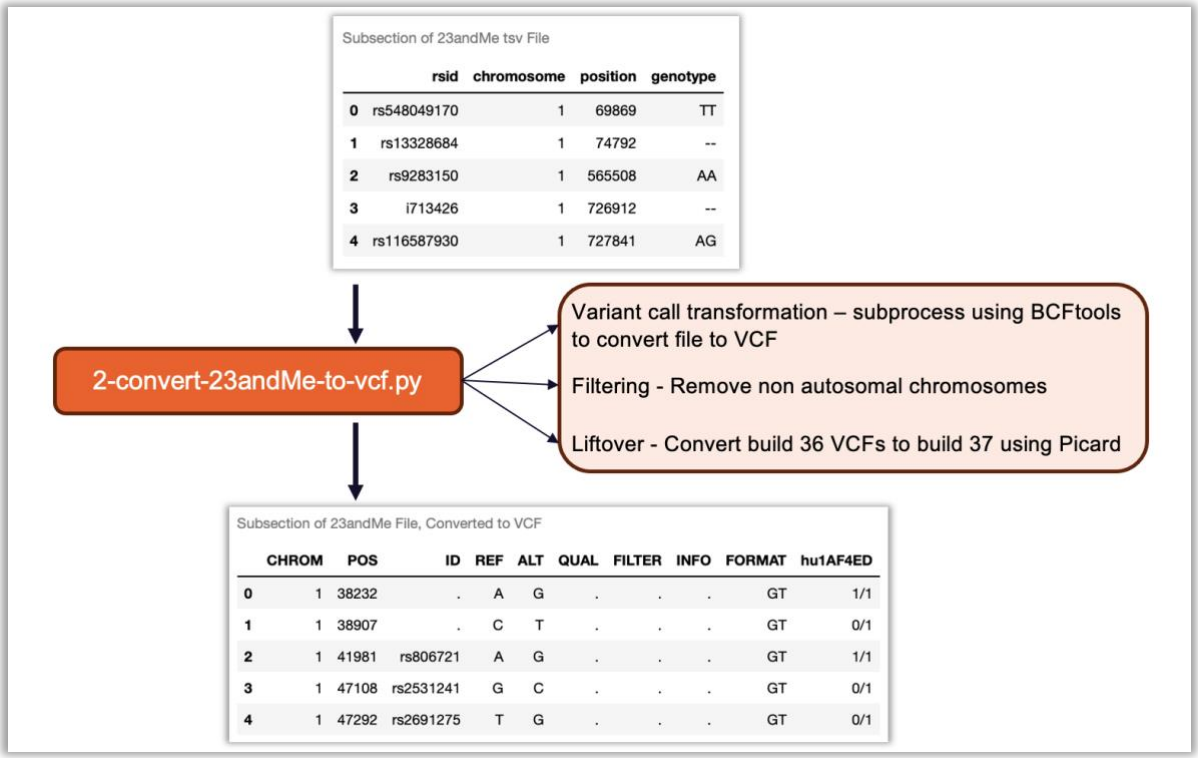


Figure 14. Summary of the Python script '2-convert-23andMe-to-vcf.py' used to preprocess and convert 23andMe files from the Harvard PGP dataset to VCF.

Following the completion of the filtering process where variants occurring in non-autosomal chromosome regions were removed, the GRCh36 files were lifted over to CRCh37 using the GATK lift-over tool, Picard. The reference genome file and chain files used in this step can be found in section *7. Data Availability*. To view the implementation of these steps, see script '2-convert-23andMe-to-vcf.py' in *8. Code Availability.*

Preprocessing - Complete Genomics

The 202 genome files in the assembly variations format produced by Complete Genomics required a separate preprocessing approach. For full documentation of the assembly variations format, see *9. Supplementary links*. As displayed in Figure 15, the assembly variations format represents the entire sequence of each chromosome. Variants are called for both alleles at a given locus, along with call quality information and external references to classify the variation. The regions that are homozygous with the reference sequence are denoted by a 'begin' and 'end' range, and '=' reference symbol. These homozygous reference regions were ultimately discarded as the process required to decompress and fully represent them in a VCF file was too computationally intensive and would have resulted in problematic file sizes. Instead, the imputation step described in the following section was relied upon to accurately impute these missing homozygous reference regions. For the called variants, only SNPs were retained (indels, transversions, etc. removed), ambiguous calls, low-quality calls, non-calls and calls in non-autosomal chromosomes were removed. The call information for both alleles at a single locus were combined into a single row. External reference cells that listed multiple external references were discarded as it was ambiguous which reference should be kept. The following imputation step would repopulate these empty external reference values with a single rsid. The processed variant call data was then written to a VCF file using a custom VCF writer function. To view the implementation of these transformation steps, see script '1-preprocess-CGI-v2.py' in *8. Code Availability*.

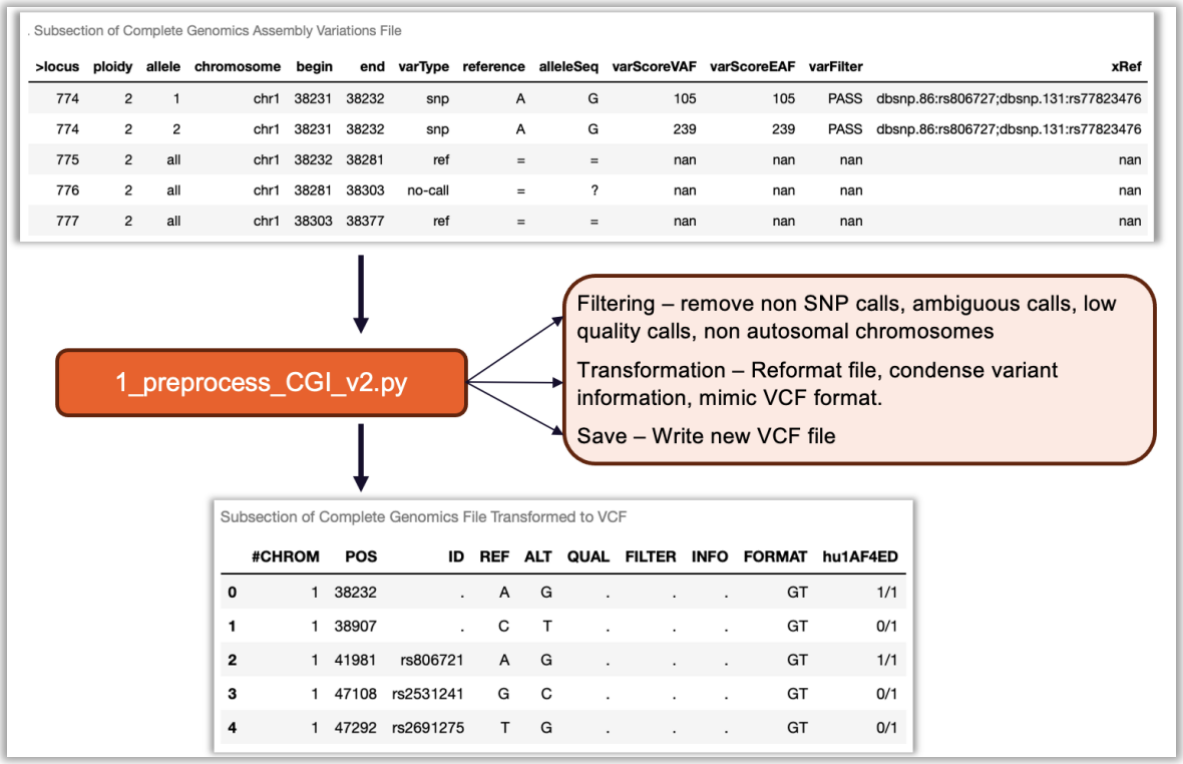


Figure 15. Summary of the Python script '1-preprocess-CGI-v2.py' used to parse and reformat all Complete Genomics assembly variation files from the Harvard PGP dataset to VCF.

Imputation

In order to make the 23andMe and Complete Genomics files comparable, genome imputation using the Beagle 5.4 program along with the 1000 Genomes human genome reference panel (n=1000) was carried out [16][17]. Default Beagle 5.4 default setting were used. This process resulted in a dataset of VCF files that all shared the same genetic markers present in the reference panel. The imputation program Beagle 5.4 was selected as it is one of the faster imputation programs available and is highly accurate in imputing common variants (MAF >= 5%) compared to other programs [18]. Variants of MAF >= 5% were to be used in the later PRS analysis, making this an appropriate choice.

Beagle 5.4 was run on each of the preprocessed VCF files, resulting in an imputed VCF file with an additional column containing imputation labels and estimated dosage information for each variant. Reference panel version can be found in *7. Data Availability*. The shell scripts used to execute the imputation process can be viewed in *8. Code Availability*.

Two small-scale assessments of imputation reliability were also carried out. The first assessment involved imputing a preprocessed 23andMe file and Complete Genomics file, both belonging to the same Harvard PGP profile. A concordance measure was then taken between the two imputed files using a shell script that utilized the SnpSift program to count positions between the two files that differed, as well as the specific call type changes. The concordance by sample results produced by the SnpSift program were then analyzed.

A second assessment was carried specifically to assess how well Beagle 5.4 could impute homozygous reference regions which were discarded in the preprocessing of Complete Genomics files. To accomplish this, a data frame of all positions in a subregion of chromosome 1 of a single Complete Genomics file was created with the help of Samtools' faidx command. This command extracted all reference positions occurring in the specified subregion from the GRCh37 reference genome assembly that was originally used to generate the file. The resulting data frame would function as the golden reference in which the imputation of the same region could be compared. Subsequently, an imputed version of the same region was generated by only using

heterozygous and homozygous variant calls as input into the imputation program. The two regions were then compared and the number of positions with imputed non-homozygous reference calls were measured. The sequence decompression process for this experiment was very time and memory intensive, therefore constraining the extent of the chromosome subregion analysed.

All imputation experiments can be found in 'Imputation Validation Experiments' from *8. Code Availability*.

Postprocessing

Following the genome imputation step, a final postprocessing step was carried out to reformat the imputed outputs and filter for variants that this study concerns. A single Python script, 'postprocess.py', was used to read the imputed VCF files, remove non-SNPs imputed by Beagle 5.4, filter out variants with MAF < 0.05, remove variants with missing external references and reorder all variants based on chromosome number and position.

Beagle 5.4 provided a dosage R-squared value (DR2) in the 'INFO' field of each imputed genotype, giving an estimate of the squared correlation between the estimated allele dosage and true allele dosage, estimating how well the imputed genotypes match the actual genotypes [16]. An experiment was carried out to assess if this measure should be used as a quality control filter in the postprocessing step. A 23andMe file was imputed and compared to a complete genomics file belonging to the same individual, checking the imputed variants against the canon variants in the later file. The relationship between DR2 score and correctly/incorrectly imputed genotypes was explored, resulting in the decision not to use this measure as a filter. See *Imputation Quality Assessment* for further reasoning.

The resulting dataset of VCF files were then ready to be loaded into the BigQuery variant schema via Google Variant Transform Tool and used in the population structure and PRS analysis.

GCP Workload Distribution

The genomics pipeline described up to this point was first developed locally and then refactored to run in distributed virtual machines using Google Cloud's Compute Engine. Refer to Figure 16 for an illustration of how Google's compute services were utilized to execute the processing of the Harvard PGP genomics dataset. See section *8. Code Availability* for VM configurations and all code used to distribute files and orchestrate workload across VMs. This was all completed under GCP's free trail tier, where ~€360 of credit was afforded. The resource quota limits influenced the compute strategy taken. Following the execution of the cloud pipeline, the final output files were assessed for quality. Some files that mistakenly passed earlier quality control steps were identified by their abnormal file sizes and excluded.

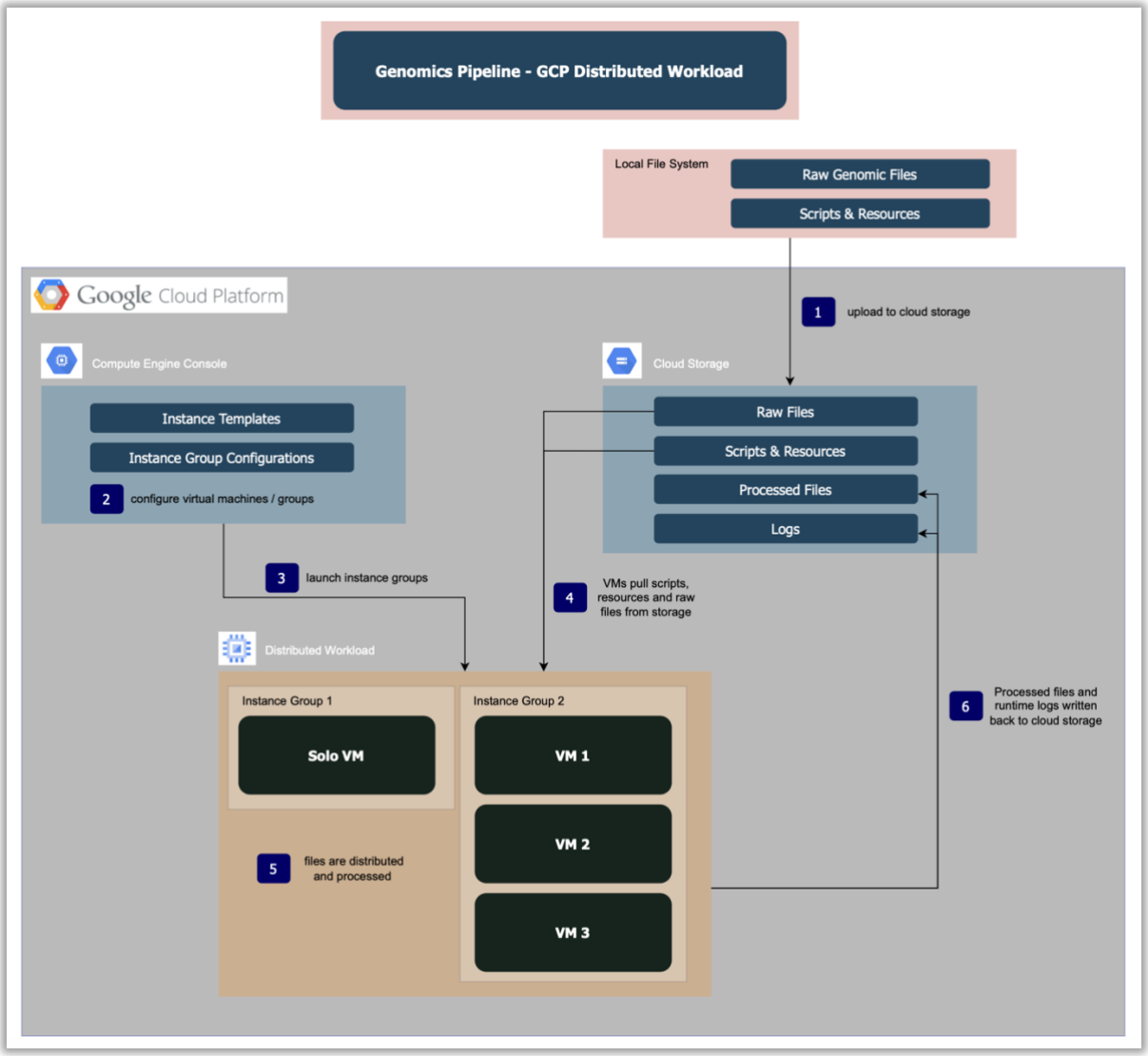


Figure 16. Diagram illustrating the execution of the custom genomics pipeline in a distributed manner using GCP. Details of virtual machine configurations, file structures and execution scripts can be viewed in the GitHub repository accompanying this paper.

Google Variant Transform Tool

The preprocessed VCF files were then loaded into Google BigQuery tables via the Google Variant Transform Tool. The Variant Transform Tool, combined with the BigQuery variant schema displayed in Figure 17, is an open-source tool that allows for scalable transformation and loading of VCF files into a queryable format. The transform tool extracts all sample information across the VCF files and subsequently populates tables by chromosome in BigQuery, a process referred to as sharding per chromosome (see *9. Supplementary links*). This representation would allow for cost and time efficient exploration of genetic variants in the study population and could also be applied to much larger datasets. Docker was used to run the tool, and the configuration used to do so can be found in *8. Code Availability*.

Field name	Type	Mode	Description
reference_name	STRING	NULLABLE	Reference name.
start_position	INTEGER	NULLABLE	Start position (0-based). Corresponds to the first base of the string of reference bases.
end_position	INTEGER	NULLABLE	End position (0-based). Corresponds to the first base after the last base in the reference allele.
reference_bases	STRING	NULLABLE	Reference bases.
alternate_bases	RECORD	REPEATED	One record for each alternate base (if any). See Additional **alternate_bases** record information.
alternate_bases.alt	STRING	NULLABLE	Alternate base.
names	STRING	REPEATED	Variant names (for example, RefSNP ID).
quality	FLOAT	NULLABLE	Phred-scaled quality score (-10log10 prob(call is wrong)). Higher values imply better quality.
filter	STRING	REPEATED	List of failed filters (if any) or "PASS" indicating the variant has passed all filters.
call	RECORD	REPEATED	One record for each call.

Figure 17. Details of the BigQuery variant schema used to represent genomic variants by chromosome. More details about the schema can be found in 9. Supplementary links.

Analysis

Population structure analyses and PRS analysis were carried out using the OMOP CDM-transformed health record information and processed genomics data from Harvard PGP. The OHDSI web tool ATLAS was used extensively for retrieving health data from the OMOP CDM and defining cohorts. This tool was setup via the OHDSI WebAPI and ATLAS setup guide GitHub tutorials (*9. Supplementary links*).

Population Structure - Chronic Diseases

To get a sense for the prevalence of chronic diseases in the Harvard PGP dataset, the 'observation' and 'conditon_occurence' OMOP CDM tables were queried for occurrences of 12 conditions. The counts for each condition were divided by the size of the study population to get its prevalence. Next, the NHANES 2017-2018 chronic disease questionnaire was acquired from the Center for Disease Control and Prevention's website (*7. Data Availability*) and the same 12 diseases were analyses for prevalence in their U.S. based population. Chronic condition prevalence between the two datasets were the compared. Any enrichments in the Harvard PGP dataset vs the NHANES dataset were noted.

Population Structure - Pathogenic and Likely Pathogenic Variants

The variants in the Harvard PGP population were analyzed with the objectives of noting variants that were of pathogenic (P) or likely pathogenic (LP) clinical significance. These variants were then linked to corresponding genes, and the genes linked to disease categories which were used to make associations between an individual's genotype and associated phenotypes in their medical record, if present. For this, an external dataset of variant information was acquired from ClinVar. After filtering for autosomal P/LP SNVs of genome build GRCh37 with valid external identifiers (rsid), the dataset yielded 54,018 unique clinically significant variants relating to 4640 unique genes. See 'Create-Clinvar-Ref-Variants.ipynb' for filtering implementation. Next, a Python script was used to construct a data frame of clinically significant variants occurring in the Harvard PGP study population. For this, the BigQuery variant schema with all loaded Harvard PGP variants was queried by chromosome, extracting all entries for individuals who possessed genotypes that were heterozygous or homozygous for the P/LP variants in the filtered ClinVar dataset. Next, disease categories

were created consisting of genes defined in a recent paper performing similar genotype-phenotype association analyses [8]. All categories and genes can be viewed in Figure 18.

Dyslipidemia	Cardiomyopathy Arrythmia	Diabetes Endocrine	Chronic Liver Disease	Cancer	Immunological Neurological Other	Hematological Diseases	Inborn Errors of Metabolism
ANGPTL4	ANK2	ABCC8	HFE	APC	CFHR5	CPOX	BTD
APOB	DSC2	FAAH		ATM	COL9A1	EGLN1	FMO3
APOC3	ENG	GLMN		BARD1	CYP21A2	F11	DMGDH
LDLR	GPD1L	HNF1A		BRCA1	DNAH5	G6PD	SLC22A5
LPL	KCNE2	HNF1B		BRCA2	FCGR1A	HBB	
MEF2A	KCNQ1	INSR		BRIP1	FLG	ITGB3	
NPC1L1	LMNA	MC4R		CHEK2	GJB2	PROC	
PAFAH1B2	MYBPC3	NEUROD1		CYLD	GRN		
PCSK9	MYH7	NOBOX		EPCAM	KIDINS220		
	MYL2	NR3C1		FAM175A	LRP5		
	MYLK	PCSK1		FH	LRRK2		
	PKP2	PPPIR3A		GEN1	MATN3		
	RBM20	PROK2		GLMN	MS4A2		
	RYR2	TBC1D4		HOXB13	MY015A		
	SCN1B			LZTR1	NMNAT1		
	SMAD6			MSH6	PKD1		
	TCAP			NBN	PRPF31		
	TNT2			NF1	RAPSN		
	TTN			PALB2	R1		
	TTR			PMS1	RYR1		
	VCL			PMS2	SCN1A		
				RAD50	SERPINA1		
				RAD51C	SGCE		
				RB1	SNCA		
				RECQL	TGM6		
				TP53	TNFRSF13B		
					WFS1		

Figure 18. Genes lists corresponding to disease categories acquired from Y.-C. C. Hou et al. 2020 reference paper [8].

For each disease category, the prevalence of heterozygous and homozygous genotype occurrences for the P/LP variants associated with each gene were visualized and compared to the those found in the aforementioned paper.

Finally, associations between variant occurrence and conditions related to the disease categories were made. For this, the OHDSI OMOP data visualization and cohort building tool, ATLAS, was used to generate cohorts of individuals who had an occurrence of at least one condition falling under the disease categories in their medical record, see 'ATLAS Disease Categories to P-LP Genotypes.ipynb' for implementation. The percent of P/LP variants of each disease category that also had a related medical record phenotype was measured. See Figure 19 for a visual overview of how the Harvard PGP study data was matched with the ClinVar, disease categories and ATLAS cohorts for these analyses.

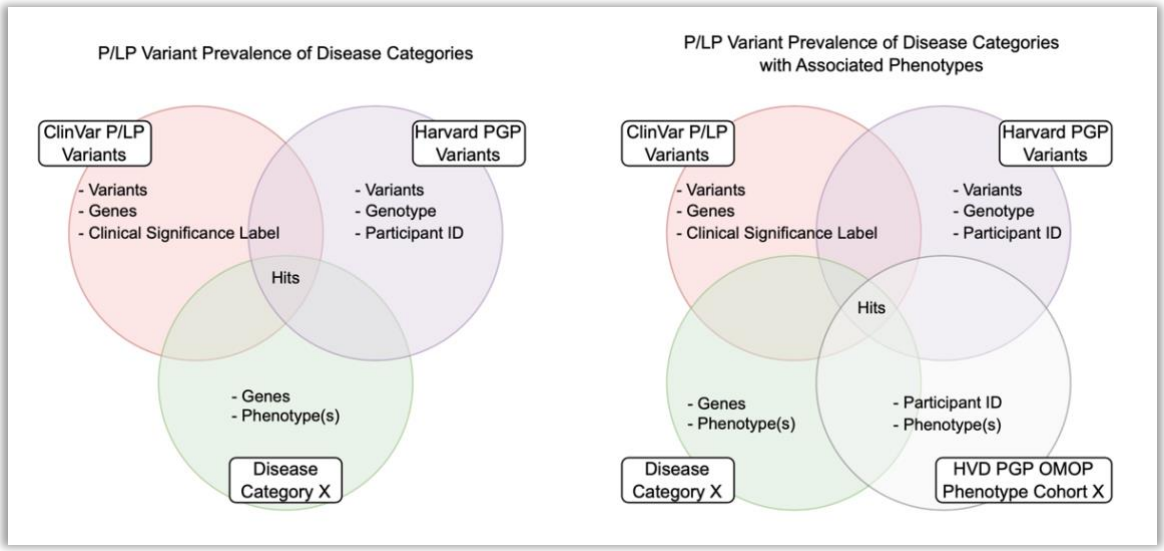


Figure 19. Abstract illustration of the pathogenic variant analysis by disease category (left) and pathogenic variant analysis with phenotype manifestations by disease category (right). Complete intersections between datasets are shown to be the objects of study.

Gastroesophageal Reflux Disease PRS Analysis

A PRS analysis was carried out to test if there was an association between GERD PRS and the phenotype. If the PRS were associative with the GERD phenotype, a predictive model would be constructed to test if the association was strong enough to be predictive. The steps taken in this analysis were guided by the PRS tutorial from Nature Protocols [19].

Base Data

The GWAS summary statistics file for this study were acquired from UK Biobank's Amazon s3 GWAS repository, labeled 'categorical-20002-both_sexes-1330.tsv'. The base data had a sample size of n = 361,194. The summary statistics underwent quality control and reformatting using the 'Process-GWAS-SumStats' Jupyter notebook. The data was filtered for autosomal SNPs corresponding to UK Biobanks's European based population. Duplicate and ambiguous SNPs were also removed. The effect sizes in the summary statistics were given as beta coefficients, therefore no transformation was required for compatibility with PRS generation of a binary trait [20].

ATLAS Cohort Definition and Target Data

For this study, ATLAS was utilized to create case and control cohorts. The cases cohort consisted of all white individuals with an age greater than 20 years in the OMOP database who had a reported condition of GERD in their medical record (n=94). The race constraint of the cohort definition fit best with the European based summary statistics, in addition to allowing for the GERD phenotype to have been observed in the selected individuals as it is known to be most prevalent in the age group between 20-29 years [21]. The control cohort used the same demographics constraint and consisted of two-times as many individuals who had no medical history of GERD (n=188). The control cohort was doubled in size due to the smallness of the dataset. Sex was not directly controlled for in the cohort definition, but its potential effect was accounted for as a covariate during the PRS generation process.

The information for each cohort defined using ATLAS were stored in the OMOP CDM. The PRS script 'main-run-prs-generation_postgres.py' was used to load the cohort information from the CDM and acquire the VCF files corresponding to each participant. The filters applied during the genomics pipeline made these files ready for PRS analysis.

PRSice2 PRS Generation

Each VCF file received a new header with updated contig coordinates and were subsequently merged into a single larger sample VCF file. Plink was then used to generate .bed, .bim and .fam format files which were the target data input into the RRSice2 PRS program. A phenotype file was created with the case and control labels for each sample. A covariates file containing sex information for each of the samples was created, followed by an eigenvector file of the first 6 principal components for all samples. The covariate and eigenvector files were then merged into a single covariates file. The PRSice script was then run with the above inputs.

PRSice2 automatically tested a range of P-value thresholds between 0.001 and 1 for selecting SNPs. In addition, linkage disequilibrium was automatically handled using clumping to thin SNPs in high linkage disequilibrium with each other. The combination of clumping and thresholding employed by PRSice2 is known as the C+T method [22]. Strand flipping was also automatically handled.

The arguments used to run the PRSice2 program are shown in Figure 20, and the complete PRS Python script is available in 'Analysis' of section *8. Code Availability*.

```
/Users/jerenolsen/Desktop/All_Tests/PRSice_Testing/Test_Run/PRSice_mac/PRSice_mac \
    --a1 alt \
    --a2 ref \
    --bar-levels 0.001,0.05,0.1,0.2,0.3,0.4,0.5,1 \
    --base /Users/jerenolsen/Desktop/All_Tests/PRSice_Testing/Test_Run/base_data/UKB_GWAS_SumStats_GERD_processed.txt \
    --beta  \
    --binary-target T \
    --bp BP \
    --chr CHR \
    --chr-id c:1-ab \
    --clump-kb 250kb \
    --clump-p 1.000000 \
    --clump-r2 0.100000 \
    --cov /Users/jerenolsen/Desktop/All_Tests/PRSice_Testing/Test_Run/cohort_covariates.covariate \
    --interval 5e-05 \
    --lower 5e-08 \
    --num-auto 22 \
    --out /Users/jerenolsen/Desktop/All_Tests/PRSice_Testing/Test_Run/PRS_output/cohort_prs \
    --pheno /Users/jerenolsen/Desktop/All_Tests/PRSice_Testing/Test_Run/cohort_phenotypes.pheno \
    --pvalue P \
    --seed 3333882480 \
    --stat BETA \
    --target /Users/jerenolsen/Desktop/All_Tests/PRSice_Testing/Test_Run/plink_output/cohort_plinked \
    --thread 8 \
    --upper 0.5
```

Figure 20. Configuration and input arguments used to run RPSice2 PRS program.

Association Testing and Predictive Model

With the best PRS results generated by PRSice2, association testing and predictive modelling were performed to assess the degree to which the GERD PRS generated in the study cohort associated with the phenotype. The PRS results and cohort information were loaded into the 'PRS Analysis Standardised' jupyter notebook. Because the PRS were to be used as a continuous variable in regression models, the risk scores were first standardized to a normal distribution with a mean of 0 and standard deviation of 1 [20]. A distribution of standardized PRS scores for cases and controls were then generated, and a standard t-test was performed to compare the means of both groups. A regression plot was generated to visualize the estimated relationship between an individual's assigned score and likelihood of having the binary GERD trait.

To further explore the associations and effects between the continuous standardized PRS scores and the GERD binary outcomes, a logistic regression analysis was carried out using the Python statsmodel API library. The log odds ratio coefficient of the standardized PRS scores were calculated, in addition to other model measures such as pseudo R^2, z-score, modelled and null log-likelihoods and associated P values. The change in odds of having the GERD condition for a one-unit increase in standardized PRS score was also calculated.

Finally, to assess the predictive capabilities of the PRS for the GERD phenotype, a new regression model was constructed using Sklearn's logistic regression model. The PRS dataset was split into 70/15/15 train/validate/test groups. The model was trained using the training dataset and subsequently underwent 3-fold GridSearchCV cross validation with the validation set in order to find optimal regression penalty strength (c) and regression penalty type (l1 or l2). The model's performance was finally measured against the test set. Accuracy, precision, recall and f1-score were reported. Additionally, a receiver operator characteristic (ROC) plot was generated to visualize the performance. Feature importance coefficients were also measured for the continuous standardized PRS variable and binary sex variable used as input in the model.

3. Results

Assessment of Genomics Tools

The Google Variant Transform Tool, part of the Google Life Sciences API, was found to be most suitable for the research needs. It proved easy to understand and deploy via docker, was well documented and resulted in an array of tables by chromosome in BigQuery, each possessing an intuitive schema, that allowed for time and cost-efficient querying of variants that would also scale with datasets of much greater size. The tables possessed variant-tool generated IDs for each sample (Harvard PGP participant) that were easily relatable to a metadata table which allowed for all query results to be related back to the original profile IDs, satisfying the patient-centricity requirement. The tool is also opensource and can be adapted. For these reasons, the

Google Variant Transform Tool was determined initially to fulfill all the needs of this research, saving for a Google Life Sciences API deprecation announcement on July 19th; see discussion.

Regarding the other tools considered, Open Targets Genomics and cBioPortal lacked patient-centricity - they rather focused on the variants themselves and their connection to research studies. Fairspace was noted to be a generally applicable tool to a variety of research data needs but would not scale well as it requires representation of each variant for an individual using the resource description framework (RDF) model. OpenCGA was promising as a large-scale variant storage and analysis framework that hosted tools specifically for VCF representation but proved exceedingly complicated to setup and possessed poor and outdated documentation. Application of assessment criteria to each tool can be seen in Table 2.

	Open Source	Scalable	Well Documented & Maintained	Ease of Use	Patient Centric	Not constrained by variant data types or vocabularies
Open Targets Genomics	✓	~	✓	X	X	X
cBioPortal	✓	~	✓	~	X	X
Fairspace	✓	X	✓	✓	✓	✓
OpenCGA	✓	✓	X	X	✓	✓
GCP Variants Transform Tool	✓	✓	✓~	✓	✓	✓

Table 2. All genomics research tools assessed for variant storage and querying use case. Note that the technologies passed or failed criteria based on the perspective of suitability for hosting large scale genomics datasets.

Imputation Quality Assessment

DR2 Score

The DR2 score generated by Beagle 5.4 for imputed variants was not used as a quality control filter in the genomics pipeline due to the results shown in Figure 21. In the experiment, a DR2 score of 0 was shown to have been assigned to a much greater portion of correctly imputed variants relative to incorrectly imputed. Additionally, the number of incorrectly imputed variants with DR2 scores between 0.1 and 0.3 were determined to be negligible. See Figure 21 for DR2 comparisons.

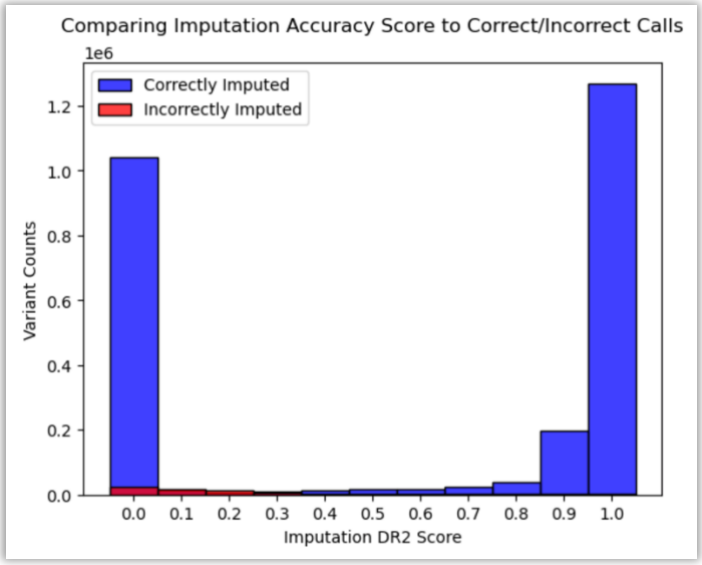


Figure 21. Comparison of correctly imputed and incorrectly imputed variants to DR2 imputation score.

SnpSift Concordance Test

The concordance test results generated by SnpSift where an imputed 23andMe and imputed Complete Genomics file of the same Harvard PGP profile were compared and can be viewed in Table 3 and Table 4. The two imputed files were measured have an overall 98.55% similarity. The most common difference in genotypes were changes from homozygous alternative to homozygous reference (0.62%) and the reverse, homozygous reference to homozygous alternative (0.48%).

Position Similarities	Measure
Total Positions Count	3.07615e+07
Changed Positions Count	445761
Changed Percent	1.45

Table 3. Overall difference between the two imputed files (23andMe and Complete Genomics of same individual) that underwent concordance measure. The total positions are given along with changed position count, indicating how many positions had different calls between the two files.

Allele Change Type	Count	Percentage
Heterozygous ALT to Homozygous REF	189788	0.62%
Heterozygous ALT to Homozygous ALT	49242	0.16%
Homozygous ALT to Heterozygous REF	44843	0.15%
Homozygous ALT to Homozygous REF	9114	0.03%
Homozygous REF to Heterozygous ALT	146393	0.48%
Homozygous REF to Homozygous ALT	6293	0.02%

Table 4. Types of call changes given by the concordance measure between the two imputed files (23andMe and Complete Genomics of the same individual).

Homozygous Reference Region Imputation Accuracy Test; Chromosome 1 Subregion

A second assessment was carried specifically to assess how well Beagle 5.4 could impute homozygous reference regions which were discarded in the preprocessing of Complete Genomics files.

Table 5 presents the results of the assessment of Beagle 5.4's ability to impute homozygous reference regions. Of the 3291 total positions shared between the imputed reference region of chromosome 1 spanning positions 13380 to 998047 and the golden reference file, 3261 positions were correctly imputed as homozygous reference, resulting in an overall difference of 0.91%.

CHR1 Subregion Alleles	Measure
Total Positions	3291
Homozygous Reference	3261
Heterozygous Alternate	29
Homozygous Alternate	1
Changed Percentage	0.91%

Table 5. Homozygous reference region imputation accuracy results. Shown are the total positions shared between an imputed region in chromosome 1 and its golden reference.

Population Structure – Chronic Disease Prevalence

The exploration of chronic disease prevalence in the Harvard PGP dataset compared to the NHANES 2017-2018 dataset showed overall that most chronic diseases are underrepresented in the Harvard PGP study population, with the exceptions of asthma and cancer. Comparisons of rates for each condition can be viewed in Figure 22.

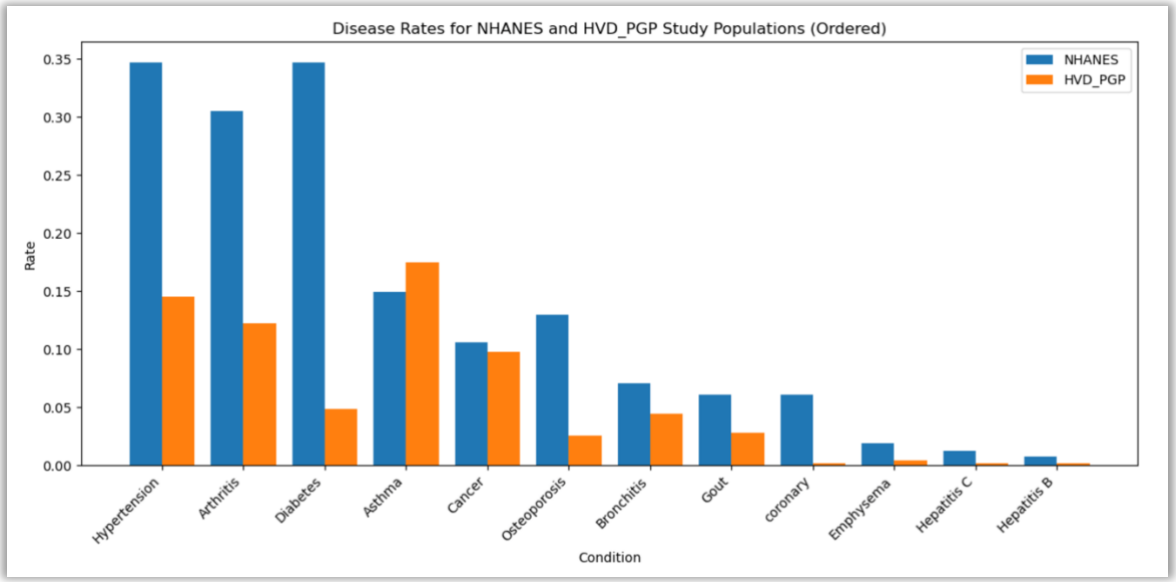


Figure 22. Bar plot displaying the prevalence of 12 chronic diseases in both the Harvard PGP study population (n = 900) and NHANES 2017-2018 survey population (n varies, ~6000-10,000 per condition).

Population Structure – Clinically Significant Variants

The Harvard PGP study population was not found to possess any P/LP SNP variants relating to the genes specified in the disease categories listed in the Figure 18. Genes lists corresponding to disease categories acquired from Y.-C. C. Hou et al. 2020 reference paper [8]. All P/LP variants that were identified are visualized in Figure 23. In the study population, ~83% of individuals were found to be heterozygous for at least one P/LP variant, 12% had at least on homozygous finding, and 17% had no P/LP variant findings. Of the two most prevalent variants, the rs11594656 variant of the IL2RA gene and rs721048 of the EHBP1 were both classified as pathogenic by ClinVar. The rates of occurrence of these two variants were notably high, therefore their origin in the data was explored further. For the most common IL2RA variant, 7.47% of the genotypes were imputed. File origins were traced to 65.9% of 23andMe files showing the variant and 47.5% of Complete Genomics files. For the variant related to the EHBP1 gene, 13.3% of the genotypes were imputed, and a balanced 33.9% of 23andMe files and 33.7% of Complete Genomics files possessed the variant.

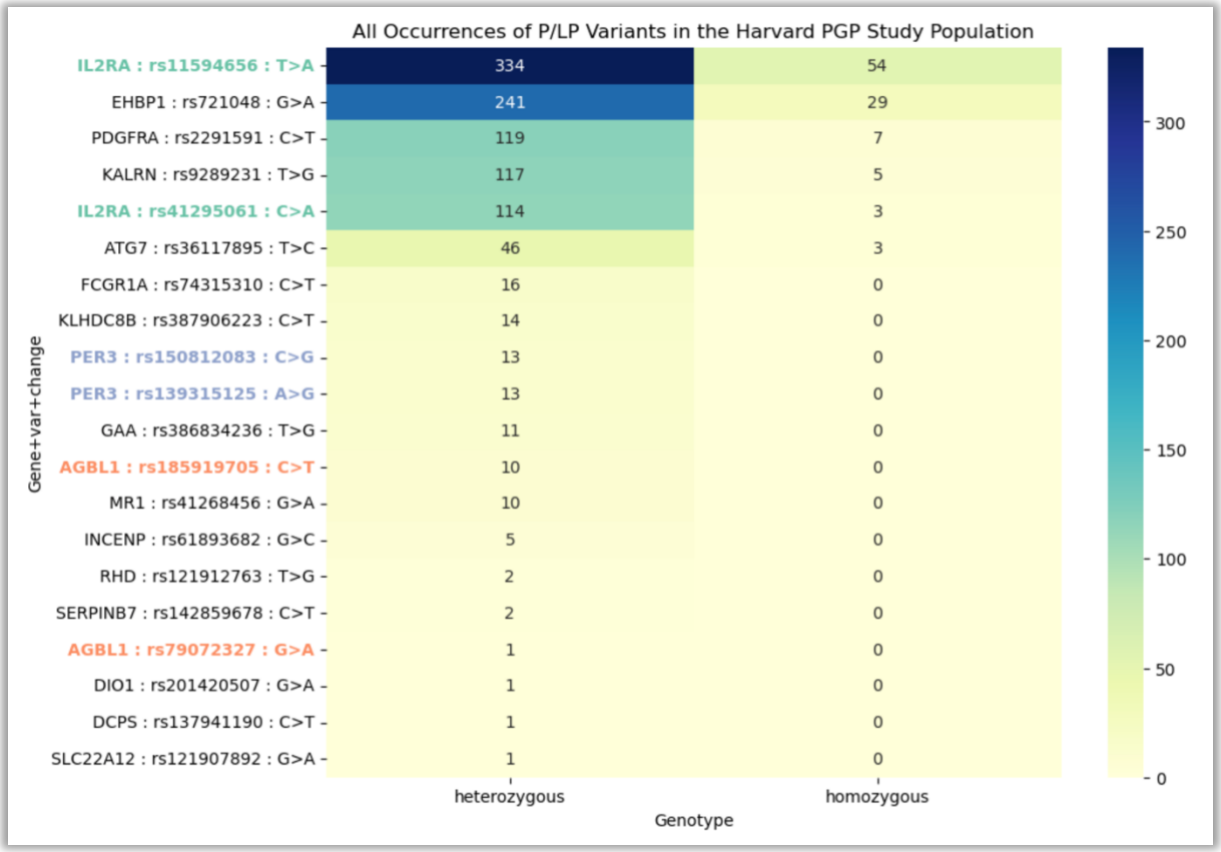


Figure 23. Heatmap displaying all occurrences of P/LP variants found in the Harvard PGP study population, by genotype. Gene-variant pairs are colored if there is more than one occurrence of the gene.

As a proof of concept for of the ability to relate genotypes to phenotypes of the predefined disease categories with the selected tools, the heatmap in Figure 24 is an example that illustrates total heterozygous and homozygous P/LP SNP variants for genes of the cancer category present in the Harvard PGP study population, had they existed, irrespective of medical record phenotypes. The bar plot in Figure 25 is an example that shows to what degree each affected gene also had related phenotypes in the medical records. True medical record cancer occurrences were used to generate these example figures, while the variants used were of benign or of unknown clinical significance.

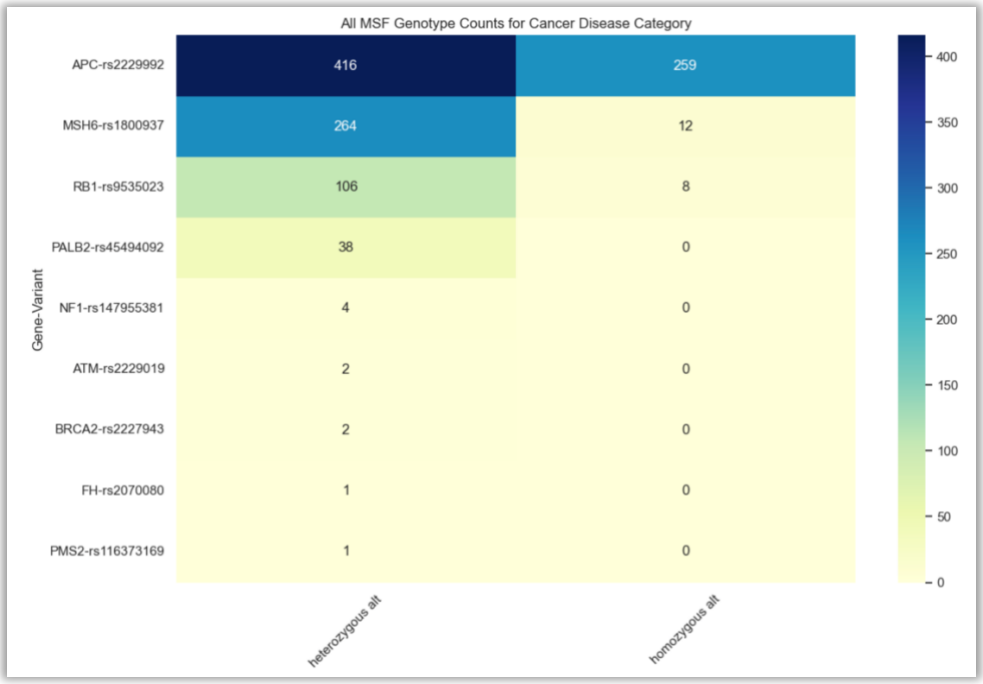


Figure 24. Heatmap example relating P/LP variants occurrences and associated genes to the cancer disease category. Heatmap was generated as proof of concept.

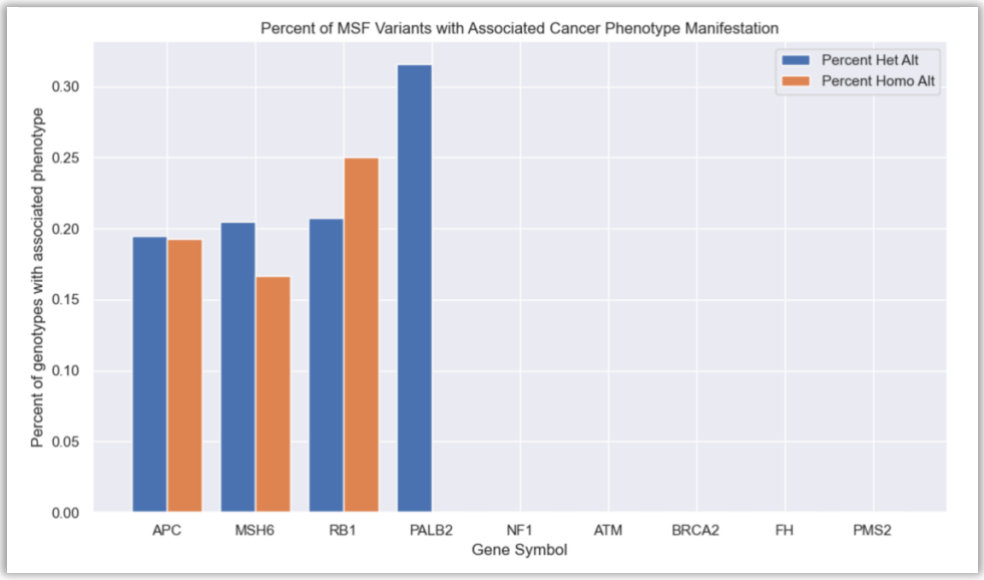


Figure 25. Bar plot showing frequency of genes affected by P/LP variants by genotype with corresponding phenotype occurrences in the OMOP CDM. Figure was generated as proof of concept.

PRS Analysis

The GERD cases (n=94) and control (n=188) cohort information were extracted from the OMOP CDM and used to run the PRS analysis. After pulling corresponding VCF files for each cohort individual, the total number of controls dropped by 25 due to an error where some individuals without valid genome files were included in the OMOP CDM. The resulting control cohort was n=163, making for a case to control ratio of 1:1.73. The cases cohort consisted of 53 white males and 41 white females aged 20 and older. The control cohort was made up of 69 white males and 30 white females ages 20 and older. The gender discrepancy was controlled for as a covariate in the PRS analysis.

A p-value threshold of 0.0318 was selected by PRSice2 for the inclusion of SNPs in the PRS model. The resulting model had an R^2 value of ~0.039. See Figure 26 and Figure 27 for visualizations of the thresholding results.

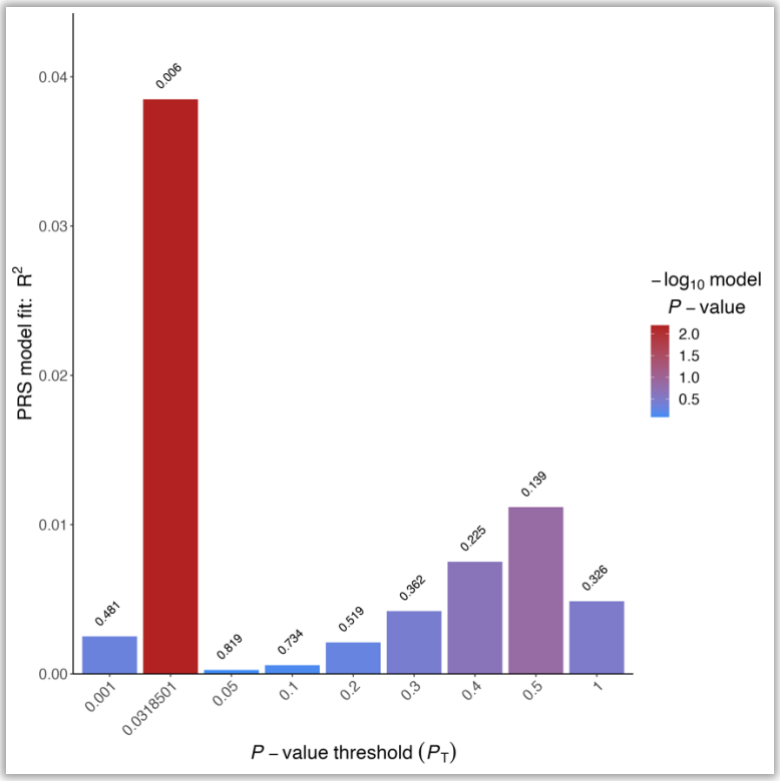


Figure 26. Bar plot from PRSice2 showing PRS models including variants from varying broad SNP P-value thresholds. The bar for the best model derived from the high-resolution run is shown in red. The PRS R^2 is the R^2 of the full model (PRS + covariates) minus the R^2 of the null model (covariates only).

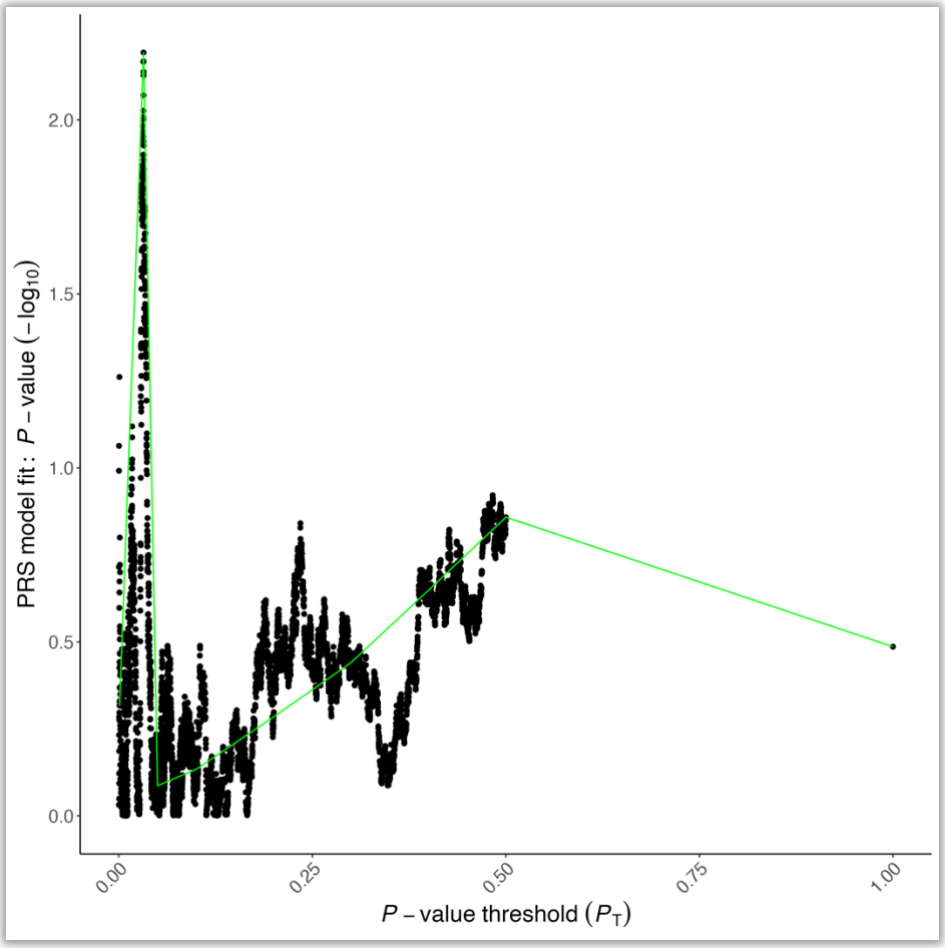


Figure 27. High Resolution PRSice2 plot of P-value thresholds for GERD PRS model. The green line connects the points at the broad P-value thresholds from Figure 26.

The first 6 principal components of the PRS target data were calculated and displayed in Figure 28, illustrating the relatively small amount of explained variance by any given component in the target data.

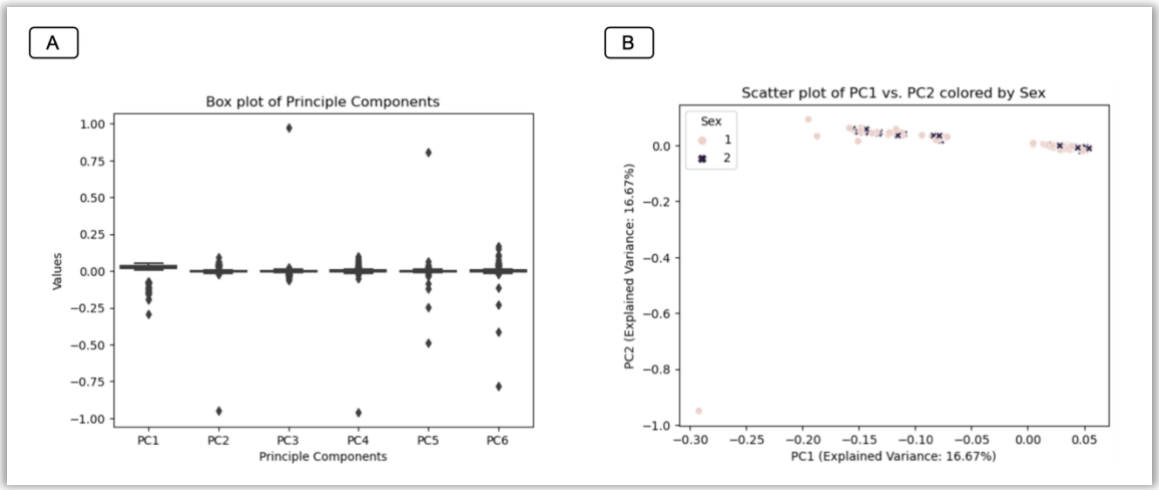
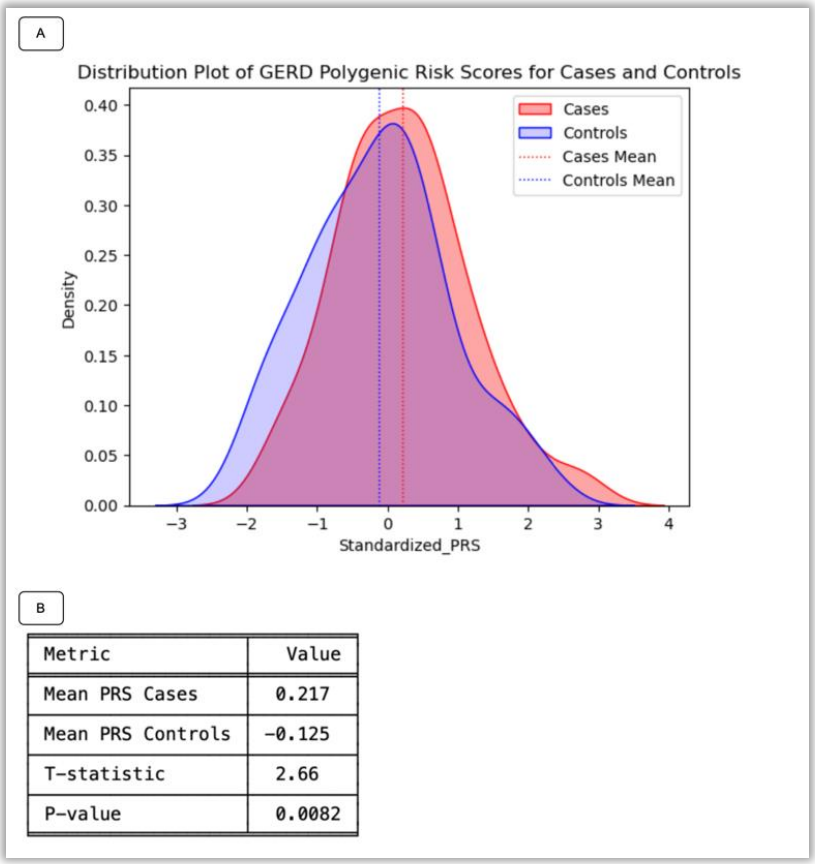


Figure 28. A. Box plot of the first 6 principal components generated from the PRS target data. B. Scatter plot visualizing the explained variance of the first 2 principal components, labeled by Sex.

The resulting PRS scores for each individual were analyzed and tested for association with the GERD target phenotype. The normal distribution-standardized scores for case and control affected statuses were plotted in Figure 29, and a standard t-test was carried out to test for significance. The mean standardized PRS for cases

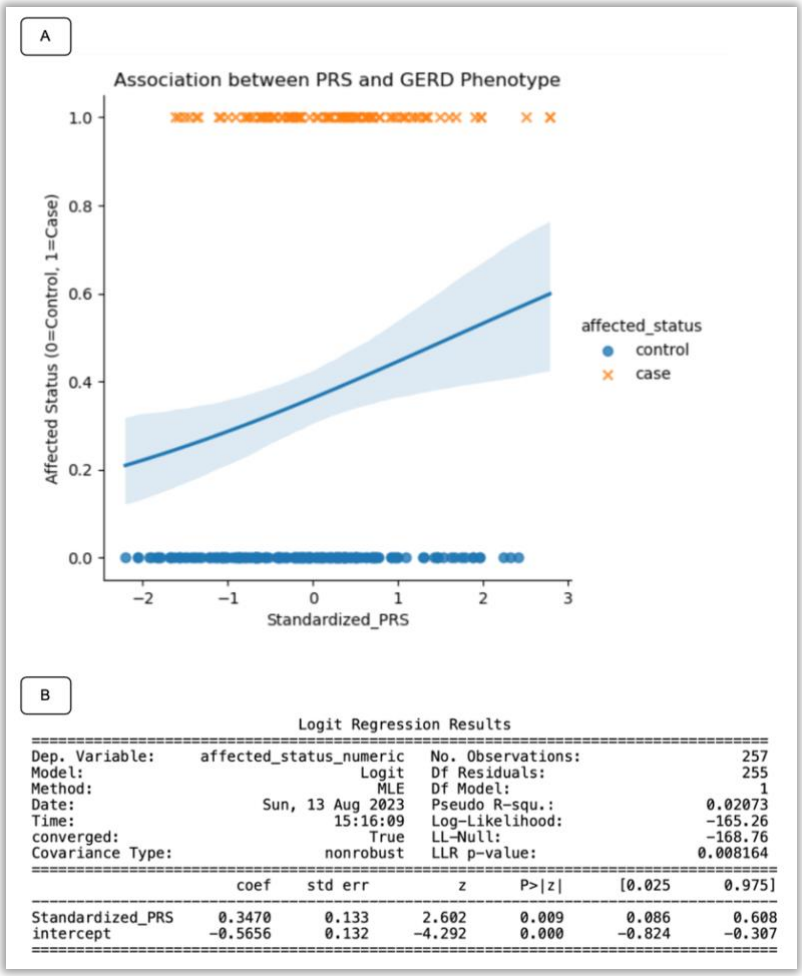
was 0.217 and -0.125 for controls, with a t-statistic of 2.66 and p-value of 0.0082. The null hypothesis of no association between PRS and GERD phenotype was rejected.



Metric	Value
Mean PRS Cases	0.217
Mean PRS Controls	−0.125
T-statistic	2.66
P-value	0.0082

Figure 29. PRS Association testing part 1. Plot A visualizes the normalized PRS scores per individual, colored by affected status. PRS means are indicated. Table B displays significance testing results with PRS means, T-statistic and associated P-value.

Further quantification of the association between PRS and GERD phenotype was analyzed via logistic regression. A pseudo-R squared value of 0.0207 was obtained, indicating the model explained around 2.07% of the variation in the affected status label of individuals. Log likelihood and null log likelihoods were calculated to be -165.26 and -168.76 respectively, with an associated P-value of 0.00816 again indicating the improvement over the null model is unlikely to be due to chance. The PRS coefficient of 0.347 gives the estimated change in log-odds of having the GERD condition for a one-unit increase in the standardized PRS. Exponentiating the log odds coefficient to get the odds-ratio increase results in an increase in odds-ratio of 41.58% for a single-unit increase in standardized PRS score. The confidence intervals of 0.086 to 0.608 indicates the range in which PRS values for the population are most likely to lie. These results and an illustration of the logistic regression curve can be viewed in Figure 30.



Logit Regression Results

Dep. Variable:	affected_status_numeric	No. Observations:	257
Model:	Logit	Df Residuals:	255
Method:	MLE	Df Model:	1
Date:	Sun, 13 Aug 2023	Pseudo R-squ.:	0.02073
Time:	15:16:09	Log-Likelihood:	-165.26
converged:	True	LL-Null:	-168.76
Covariance Type:	nonrobust	LLR p-value:	0.008164

	coef	std err	z	P>|z|	[0.025	0.975]
Standardized_PRS	0.3470	0.133	2.602	0.009	0.086	0.608
intercept	-0.5656	0.132	-4.292	0.000	-0.824	-0.307

Figure 30. Logistic Regression results for GERD PRS association testing. Plot A illustrates the logistic regression curve and log-odds likelihood of affected status. Confidence interval is given in the shaded region. Table B presents the logistic regression statics for the regression model.

A second regression model was created with the objective of directly assessing PRS predictive capabilities. The regression model was trained and validated with 3-fold cross-validation, where a sum of squares regularization method (l2) and regression penalty value (c parameter) of 5 were selected as optimal hyper parameters. Predictions were made using the test set and yielded an overall accuracy of 64%. The area under the curve (AUC) of the ROC plot corresponding to the model performance was 0.74. For the controls (n=24), the model performed with precision, recall and f1-scores of 0.63, 0.96 and 0.77 respectively. For the cases (n=15), the model performed with scores of 0.67, 0.13 and 0.22. While the model was successful in correctly classify individuals from the control group, it had very poor performance in predicting the GERD phenotype group.

Taken together, these results indicate a statistically significant positive association between PRS and GERD phenotype in the Harvard PGP population, however, the PRS continuous predictor alone explains very little of the variation in the occurrence GERD and is not predictive of the class on its own.

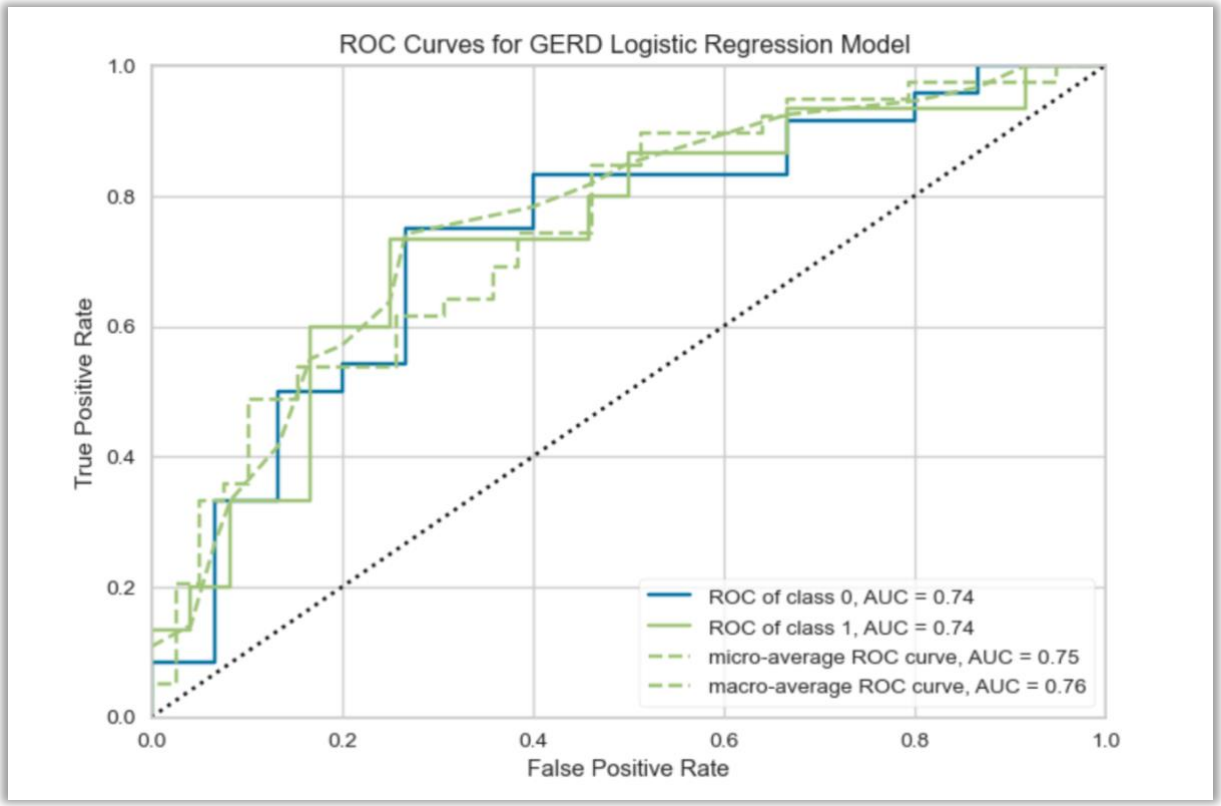
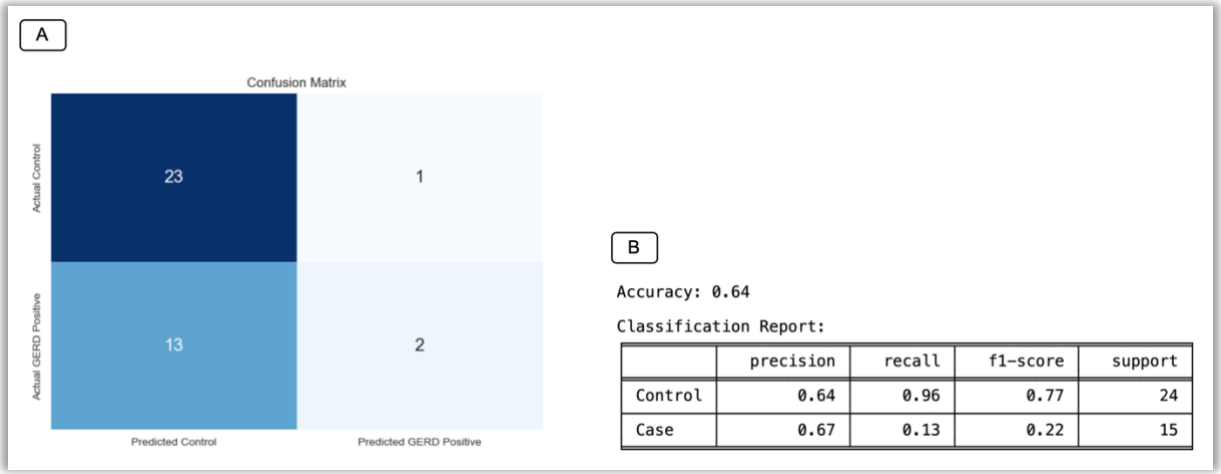


Figure 31. ROC AUC plot for the GERD logistic regression predictive model. Overall AUC for each class is given, along with overall micro-average and macro-average curves.



	precision	recall	f1-score	support
Control	0.64	0.96	0.77	24
Case	0.67	0.13	0.22	15

Figure 32. Summary of GERD logistic regression predictive model performance. Section A displays a confusion matrix illustrating the correct and incorrect class assignments, with B indicating the performance statistics by class.

4. Discussion

Genomics Tools and Google Life Sciences API

At the time this thesis was produced, the Google Variant Transform Tool, part of the Cloud Life Sciences API suite, was identified as the best genomics research tool for the scope of this study. The convenience of the integrated cloud compute environment and support for other general purpose genomics research tools further motivated the tool's selection. Additionally, it was implemented without issue and the BigQuery variant schema was successfully utilized as intended. Unfortunately, on July 20th, 2023, Google announced that the Cloud Life Sciences API would be deprecated by July 2025. Its tools are no longer made available to

customers without the API previously enabled. It is stated that the tools under the Life Sciences API suite are now migrating to Google Batch service (see supplementary links), however, they will not be available in the form utilized by this project.

The criteria used to assess all tools in this study were based on the perspective of suitability for large scale genomics. For example, while Fairspace's RDF model is general purpose and may be scalable for the majority of research data needs, it was not considered scalable for variant representation as millions of 'hasSample' - 'hasDiagnosis' relations which would need to be tied to a single subject, resulting unnecessarily large and complex queries.

Imputation

Genotype imputation using a common reference panel was determined to be the most appropriate approach to makings the disparate 23andMe genotyping and Complete Genomics assembly variation files comparable. As Beagle 5.4 has been demonstrated to be more accurate than other imputation programs at imputing variants with MAF >= 5%, the basis of this study, it became the obvious choice. Regarding the imputation performance, the results of the first of two small-scale imputation assessments indicated that the concordance measure between an imputed 23andMe VCF file and imputed Complete Genomics VCF file from the same Harvard PGP profile 98.55% similar, with no abnormal rate of call type changes. The second small-scale analysis indicating Beagle 5.4's ability to impute homozygous reference regions, which were discarded from Complete Genomics files in the preprocessing step, showed a performance of 99.09% similarity in the imputed and golden-reference versions of the chromosome 1 region assessed. These results are in line with the accuracy to be expected from Beagle 5.4. A study using Beagle 5.4 and varying reference panel sizes to impute genotypes in cattle showed 94-95% concordance ranges when using a reference panel size of 1000 to impute MAF >=5% variants [23]. In another study analyzing the effects of different imputation quality control methods showed that using 2014 versions of the Beagle imputation program and the 1000 genomes reference in masked SNP groups of MAF >= 5% resulted in average concordance measures of 99% [24]. Overall, the imputation carried out in this project was not suspected to have introduced any major confounding changes, at least no greater than expected in modern genomics research, in the genomics data of the Harvard PGP study population.

Population Structure

The results of the population structure analysis concerning chronic disease prevalence were unexpected as 10 of the 12 diseases surveyed were underrepresented in the Harvard PGP study population in comparison to the NHANES 2017-2018 population. For most diseases, the underrepresentation exceeded 50%. However, asthma and cancer rates were on par with the rates seen in NHANES. The most likely explanation for this discrepancy in disease rates is that the medical records of the individuals from the Harvard PGP study population are not sufficiently up to date and / or are generally not comprehensive in their detail. In many cases the EHR source data for an individual only consisted of survey data. Additionally, not all survey types were not completed ubiquitously across Harvard PGP participants. This reality could be a confounding factor with regards to cohort generation, especially concerning the GERD PRS cohorts. The rate of occurrence of GERD in the study population was shown to be 14.9%, whereas epidemiological studies estimate GERD rates to occur between 18.1% and 27.8% in the U.S. population [25]. If the underrepresentation of GERD in the Harvard PGP study population is due to incomplete medical records, it is possible that individuals with unreported cases of GERD were selected for inclusion in the control cohort, potentially influencing the PRS results.

Regarding findings of P/LP variants in the Harvard PGP study population, 83% of individuals were found to be heterozygous for at least one P/LP variant, 12% had at least one homozygous occurrence and 17% had no clinically significant finding. The two most commonly occurring variants, rs11594656 of the IL2RA gene (n heterozygous = 334, n homozygous = 54) and rs721048 of the EHBP1 gene (n heterozygous = 241, n homozygous = 29), saw unexpectedly high rates given their ClinVar clinical significance classification of pathogenic. After exploring file sources and imputation rates of these variants, neither of these factors were suspected to be the source of the high rates of occurrence. After exploring the evidence for these variants in

ClinVar, it was found that both variants had only a single line of evidence indicating pathogenicity, with dates of publishing from 2007 and 2008 respectively [26][27]. A separate study not listed as evidence on the IL2RA variant indicated that only the homozygous TT manifestation of the variant confers the risk for the associated condition (T1D) [28].

In a separate study analyzing the frequency of pathogenic variant occurrences in asymptomatic individuals, it was noted that of the 150,000 pathogenic variants discovered at the time, many were originally discovered via small cohort studies of affected individuals. In addition, it was found that many pathogenic variants occur frequently in asymptomatic individuals, indicating that the pathogenic classification may be erroneous or have lower penetrance than previously thought [29]. This reality should also be considered when interpreting high rates of occurrence of pathogenic variants.

For the analysis of P and LP variants affecting genes from predetermined disease categories in the Harvard PGP study population and their potential connection to associated phenotypes, no P/LP variants were found. This initially unexpected result has several likely causes. The genes associated with each disease category in Figure 18 were taken from a reference paper performing a similar genotype-phenotype association analysis in a separate study population [8]. The variants this paper sought to analyze was derived from ClinVar, OMIM and HGMD and included variants of all types (insertions, deletions, transversions, etc.) related to 30,281 unique genes. Unfortunately, the variant inclusion criteria for each of the 3 sources were not specified, and the OMIM and HGMD datasets were not available without an agreement. In contrast, the variant dataset generated for this project only contained SNP variants from ClinVar relating to 4640 unique genes. This discrepancy on its own would greatly limit the amount of overlapping hits for variants between the two reference variant datasets. Additionally, no MAF filter was stated to have been applied to the variant calls in the study population, whereas a MAF > 5% was applied in this project.

GERD PRS

The PRS study undertaken using the GERD case and control cohorts showed that there was a positive association between PRS and the GERD phenotype ($\log(OR) = 0.347$, $P = 0.009$, pseudo-$R^2 = 0.0207$). The average standardized PRS for the cases cohort was 0.217 and -0.125 for the control cohort (T-statistic = 2.66). While these results are statistically significant, the amount of phenotypic variance explained by the continuous PRS variable is relatively low, with the pseudo-R^2 indicating an explained variance of only 2.07%. While this may even seem remarkably small, this result aligns with a large body of research showing PRS models often explain only a fraction (< 1%) of phenotypic variance [19].

Visual examination of the logistic regression curve in Figure 30 immediately indicates that the PRS scores generated by this model would have poor predictive ability, and this was confirmed after generating a new predictive regression model. The predictive logistic regression model showed moderately low precision for both groups (control = 0.64, case = 0.67), high recall for predicting controls and low recall for cases (control = 0.96, case 0.13) and a similar pattern in f1-score (control = 0.77, case = 0.22). In addition to the model having low predictive capabilities, it is not generalizable to a broader population as it is based solely on a base and study populations with European ancestry.

Lastly, the target sample size used in this study was suboptimal (n cases = 94, n controls = 188). The cases and control combined sample size only just exceeded the recommended cutoff of n = 100. Most PRS studies typically include target sample sizes in the magnitude of 1000s of individuals, and studies of this size are more likely to show highly significant results [19]. Specifically, it has been demonstrated that many studies publishing null results were likely caused by being underpowered rather than possessing bias or confounding factors. Additionally, it has been shown that the power of PRS association testing is greatest when using equal sized base and target data, whereas individual prediction is optimized by maximizing the size of the base data [30]. In the GERD PRS analysis undertaken, the target data came nowhere close to the size of the summary statistics used (base n = 361,194).

Improvements

There are various improvements that could be made to this study to better demonstrate the utility of the selected tools, lead to more telling variant analyses and enhance the potential clinical utility of the PRS model. The Harvard PGP dataset was both small and often contained EHRs sparsely populated with information. A larger study dataset would make the PRS model more informative and have provided a more solid basis for predictive model assessment. Additionally, inclusion of environmental risk factors and disease comorbidities (disease-disease networks) could give rise to models better able to predict absolute risk for disease, and do so in a demographic-stratified manner [31][32].

The genomics pipeline employed should have allowed for an intermediate variant dataset for use in the variant analyses that used a lower MAF threshold and included more than just SNPs. Alongside this improvement, a more comprehensive reference variant dataset that also includes mode of inheritance information for each variant would provide better insight into P/LP variant prevalence and their phenotype associations.

5. Conclusion

Overall, the OMOP CDM and Google Variant Transform tool used in tandem proved capable of population structure analyses and PRS studies despite some of the shortcomings of the datasets which the tools were applied to. The utilization of the Google Variant Transform tool within the context of the Cloud Life Sciences API, despite its subsequent deprecation, showcased the potential of cloud-based environments for variant analysis. Further, in light of the growing interest in leveraging EHR data for predictive modeling, this study contributes to the ongoing movement towards the integration of genetic and phenotypic data. The study's limitations serve as a reminder of the inherent complexities of real-world health data. As the field advances, the integration of genetic and phenotypic data shows promise for fostering a deeper understanding of complex disease relationships and improving predictive modeling for personalized healthcare.

6. Acknowledgments

I would like the thank The Hyve for the opportunity to complete my master’s thesis in such an innovative environment committed to improving open health science informatics with the objective of fostering better health decisions and better care. I am grateful for having Julia Kurps as my supervisor - always keen to hear my updates on the project and make sure I was moving in the right direction. Additionally, a great thanks to the other members of The Real-World Data Team (Sofia, Stefan, Azadeh and Jan) for being supportive and eager to help get a handle for OHDSI tools.

I would also like to acknowledge my examiner, Douwe Molenaar, for providing me with feedback and validation in the formation of my research goals.

A special thank you to Natalia Azcona Granada, for checking the sanity of my reasonings, providing me with unconditional support throughout my master’s project and always being there as my friend.

A loving thank you is sent to my friends and family for sending me their encouragement and support from the other side of the globe.

Lastly, I would like to extend my gratitude to the anonymous individuals who selflessly made their personal health information and genomics data publicly available. Without their contributions, open collaborative research on the intersection of human genomes and health outcomes would not be possible.

7. Data Availability

- **All 798 Harvard PGP participant study IDs can be found here:**
  - https://github.com/JamWithBread/Masters-Thesis-VU-Hyve/blob/main/Resources/Harvard-PGP-798-profiles.txt
- **ClinVar Variants Summary download can be found here:**
  - https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz
- **NHANES Chronic diseases dataset can be found here:**
  - **2017-2018** -https://wwwn.cdc.gov/Nchs/Nhanes/continuousnhanes/default.aspx?BeginYear=2017
- **Reference Assemblies Used:**
  - **GRCh37** - http://grch37.ensembl.org/Homo_sapiens/Info/Index
  - **GRCh36** - https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.12/
  - **Chain file** - https://github.com/JamWithBread/Masters-Thesis-VU-Hyve/blob/main/Resources/NCBI36_to_GRCh37.chain
- **1000 Genomes Reference Panel:**
  - **Bref3 files** - https://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/b37.bref3/

8. Code Availability

All code is made available via public GitHub repository at:

https://github.com/JamWithBread/Masters-Thesis-VU-Hyve

*List of software and versions also included

9. Supplementary links

Understanding the BigQuery variants schema	https://cloud.google.com/life-sciences/docs/how-tos/bigquery-variants-schema
Sharding per chromsosome and BigQuery Partitioning	https://github.com/googlegenomics/gcp-variant-transforms/blob/master/docs/query_cost.md
Cloud life sciences migrating to Batch	https://cloud.google.com/batch/docs/migrate-to-batch-from-cloud-life-sciences
OHDSI ATLAS setup guide	https://github.com/OHDSI/Atlas/wiki/Atlas-Setup-Guide
OHDSI web API setup guide	https://github.com/OHDSI/WebAPI
OHDSI Athena Vocabularies and Versions	https://athena.ohdsi.org/vocabulary/list
Complete Genomics variant ASM file structure	http://www.genomeinterpretation.org/asm-file-format.html

10. References

[1] S. J. Grannis *et al.*, "Evaluating the effect of data standardization and validation on patient matching accuracy," *J. Am. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 447–456, May 2019, doi: 10.1093/jamia/ocy191.

[2] "Data Standardization – OHDSI." 2020. [Online]. Available: https://www.ohdsi.org/data-standardization/

[3] "OMOP Common Data Model." OHDSI. [Online]. Available: https://ohdsi.github.io/CommonDataModel/

[4] Amelia J Averitt, Alexandra Orlova, Alexander Davydov, Oleg Zhuk, Michael N Cantor, and Gregory Klebanov, "Conversion of UK Biobank into the OMOP CDM: New Data for Inferences Between Episodic Care." 2021. [Online]. Available: https://www.ohdsi.org/2021-global-symposium-showcase-1/

[5] G. Hripcsak, M. J. Schuemie, D. Madigan, P. B. Ryan, and M. A. Suchard, "Drawing Reproducible Conclusions from Observational Clinical Data with OHDSI," *Yearb. Med. Inform.*, vol. 30, no. 01, pp. 283–289, Aug. 2021, doi: 10.1055/s-0041-1726481.

[6] *Observational Health Data Sciences and Informatics. The Book of OHDSI.* 2021. [Online]. Available: https://ohdsi.github.io/TheBookOfOhdsi/

[7] N. Ahmadi, Y. Peng, M. Wolfien, M. Zoch, and M. Sedlmayr, "OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review," *Int. J. Mol. Sci.*, vol. 23, no. 19, p. 11834, Oct. 2022, doi: 10.3390/ijms231911834.

[8] Y.-C. C. Hou *et al.*, "Precision medicine integrating whole-genome sequencing, comprehensive metabolomics, and advanced imaging," *Proc. Natl. Acad. Sci.*, vol. 117, no. 6, pp. 3053–3062, Feb. 2020, doi: 10.1073/pnas.1909378117.

[9] S. M. Harrison and H. L. Rehm, "Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar," *Genome Med.*, vol. 11, no. 1, p. 72, Dec. 2019, doi: 10.1186/s13073-019-0688-9.

[10] T. Frayling, "Genome-wide association studies: the good, the bad and the ugly," *Clin. Med.*, vol. 14, no. 4, pp. 428–431, Aug. 2014, doi: 10.7861/clinmedicine.14-4-428.

[11] C. M. Lewis and E. Vassos, "Polygenic risk scores: from research tools to clinical instruments," *Genome Med.*, vol. 12, no. 1, p. 44, Dec. 2020, doi: 10.1186/s13073-020-00742-5.

[12] A. V. Khera *et al.*, "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations," *Nat. Genet.*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018, doi: 10.1038/s41588-018-0183-z.

[13] J. An *et al.*, "Gastroesophageal reflux GWAS identifies risk loci that also associate with subsequent severe esophageal diseases," *Nat. Commun.*, vol. 10, no. 1, p. 4219, Sep. 2019, doi: 10.1038/s41467-019-11968-2.

[14] C. Lu, B. Greshake Tzovaras, and J. Gough, "A survey of direct-to-consumer genotype data, and quality control tool (GenomePrep) for research," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 3747–3754, 2021, doi: 10.1016/j.csbj.2021.06.040.

[15] C. Charon, R. Allodji, V. Meyer, and J.-F. Deleuze, "Impact of pre- and post-variant filtration strategies on imputation," *Sci. Rep.*, vol. 11, no. 1, p. 6214, Mar. 2021, doi: 10.1038/s41598-021-85333-z.

[16] Brian Browning, "Beagle 5.4." University of Washington, Jul. 22, 2022. [Online]. Available: http://faculty.washington.edu/browning/beagle/beagle.html

[17] B. L. Browning, Y. Zhou, and S. R. Browning, "A One-Penny Imputed Genome from Next-Generation Reference Panels," *Am. J. Hum. Genet.*, vol. 103, no. 3, pp. 338–348, Sep. 2018, doi: 10.1016/j.ajhg.2018.07.015.

[18] A. D. Marino *et al.*, “A comparative analysis of current phasing and imputation software,” Genomics, preprint, Nov. 2021. doi: 10.1101/2021.11.04.467340.
[19] S. W. Choi, T. S.-H. Mak, and P. F. O’Reilly, “Tutorial: a guide to performing polygenic risk score analyses,” *Nat. Protoc.*, vol. 15, no. 9, pp. 2759–2772, Sep. 2020, doi: 10.1038/s41596-020-0353-1.
[20] J. A. Collister, X. Liu, and L. Clifton, “Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists,” *Front. Genet.*, vol. 13, p. 818574, Feb. 2022, doi: 10.3389/fgene.2022.818574.
[21] T. Watanabe, “Gastro-esophageal reflux disease symptoms are more common in general practice in Japan,” *World J. Gastroenterol.*, vol. 13, no. 31, p. 4219, 2007, doi: 10.3748/wjg.v13.i31.4219.
[22] S. W. Choi and P. F. O’Reilly, “PRSice-2: Polygenic Risk Score software for biobank-scale data,” *GigaScience*, vol. 8, no. 7, p. giz082, Jul. 2019, doi: 10.1093/gigascience/giz082.
[23] Y. Jiang, H. Song, H. Gao, Q. Zhang, and X. Ding, “Exploring the optimal strategy of imputation from SNP array to whole-genome sequencing data in farm animals,” *Front. Genet.*, vol. 13, p. 963654, Aug. 2022, doi: 10.3389/fgene.2022.963654.
[24] S. S. Verma *et al.*, “Imputation and quality control steps for combining multiple genome-wide datasets,” *Front. Genet.*, vol. 5, Dec. 2014, doi: 10.3389/fgene.2014.00370.
[25] H. B. El-Serag, S. Sweet, C. C. Winchester, and J. Dent, “Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review,” *Gut*, vol. 63, no. 6, pp. 871–880, Jun. 2014, doi: 10.1136/gutjnl-2012-304269.
[26] C. E. Lowe *et al.*, “Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes,” *Nat. Genet.*, vol. 39, no. 9, pp. 1074–1082, Sep. 2007, doi: 10.1038/ng2102.
[27] J. Gudmundsson *et al.*, “Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer,” *Nat. Genet.*, vol. 40, no. 3, pp. 281–283, Mar. 2008, doi: 10.1038/ng.89.
[28] H. M. Abdelrahman, L. M. Sherief, D. M. Abd Elrahman, A. Alghobashy, H. F. Elsaadani, and R. H. Mohamed, “The association of PTPN22 (rs2476601) and IL2RA (rs11594656) polymorphisms with T1D in Egyptian children,” *Hum. Immunol.*, vol. 77, no. 8, pp. 682–686, Aug. 2016, doi: 10.1016/j.humimm.2016.06.006.
[29] C. A. Cassa, M. Y. Tong, and D. M. Jordan, “Large Numbers of Genetic Variants Considered to be Pathogenic are Common in Asymptomatic Individuals,” *Hum. Mutat.*, vol. 34, no. 9, pp. 1216–1220, Sep. 2013, doi: 10.1002/humu.22375.
[30] F. Dudbridge, “Power and Predictive Accuracy of Polygenic Risk Scores,” *PLoS Genet.*, vol. 9, no. 3, p. e1003348, Mar. 2013, doi: 10.1371/journal.pgen.1003348.
[31] N. Chatterjee, J. Shi, and M. García-Closas, “Developing and evaluating polygenic risk prediction models for stratified disease prevention,” *Nat. Rev. Genet.*, vol. 17, no. 7, pp. 392–406, Jul. 2016, doi: 10.1038/nrg.2016.27.