| | |
|---|---|
| **Title** | Health Equity Research Assessment (HERA) |
| **Protocol version identifier** | 6.0 |
| **Date of last version of protocol** | July 5, 2023 |
| **Authors** | Noémie Elhadad<br>Tony Sun<br>Harry Reyes Nieva<br>Anthony Sena<br>Patrick Ryan |

## RATIONALE AND BACKGROUND

While health equity research has identified diagnosis and treatment disparities across gender and race, most published research has focused on one condition/intersectional demographic at a time, with cohorts usually including patients from only one dataset. The goal of the OHDSI Health Equity Research Assessment (HERA) is to enable systematic, large-scale characterizations of health disparities and differences simultaneously across intersectional groupings [e.g. race, gender, age], while also including data from numerous OHDSI databases.

HERA will serve as a hypothesis generation platform for health disparities researchers to investigate scenarios when significant differences are identified. By providing downloadable summary statistics that display findings across institutions, HERA will enable the systematic identification and quantification of differences in disease and treatment patterns across various sub-groups.

In this document, we describe the cohorts created and some initial analyses performed. Given a single cohort, prevalence rates are computed for each demographic subgroup of interest (race alone, gender alone, race+gender) for treatments and for conditions occurrences. These computations are carried out for patients prior to entry into cohort, at index time, and following entry into cohort. After computation, these summary data can be downloaded from each site using the HERA Dashboard, a centralized OHDSI website that enables exploratory data analysis.

## R PACKAGE AND SITE FOR EXPLORATORY DATA ANALYSIS

The R study package can be found on Github:

- https://github.com/ohdsi-studies/HERACharacterization
- Setup and installation instructions are listed in the STUDY-PACKAGE-SETUP.md file.
- Package running instructions can be found in the STUDY-EXECUTION.md file.
- Contact Anthony Sena (asena5@its.jnj.com) for technical issues and getting the OHDSI sharing key for uploading results
- Contact Tony Sun (tys2108@cumc.columbia.edu) for questions about data visualization and further analytical results

Exploratory data analysis of all results can be found at:
https://data.ohdsi.org/HERACharacterization/

Additional visualizations and interpretations of results can be found at:
https://even.dbmi.columbia.edu/HERA/

## STUDY MATERIALS AND DESIGN

This section identifies the data sources, phenotype definitions, subgroups, and analysis methods used to estimate differences in disease diagnosis. A set of experiments is delineated to estimate the sensitivity of differences in diagnosis to these factors. The overarching research question is: **how do factors such as data source and various intersectional demographics [e.g. race, gender, age] influence the estimated differences in the prevalence of disease and treatments between subgroups?**

Using target and comparator subgroups, differences in prevalence can be measured. Any difference observed between the target and comparator subgroups could be due to statistical

noise, biological difference, or health disparity arising from social determinants of health or provider bias. Availability of these quantities across different databases or across multiple cohorts might help disentangle some of these biases. But ultimately, the purpose of the study is to generate hypotheses for health equity researchers to investigate in more depth.

## A. Study Materials

### A.1. Data sources

We study data sources with different provenance as well data representing different populations. For example, populations studied include privately insured, employed patients in the IBM MarketScan Commercial Claims and Encounters (CCAE), patients with limited income in the IBM MarketScan Multi-state Medicaid (MDCD), and patients from international databases. Some data sources include race data, which we've listed below:

- Columbia University Irving Medical Center (CUIMC) EHR
- IBM MarketScan Multi-state Medicaid (MDCD)
- Optum® de-identified Clinformatics Data Mart Database (Clinformatics)
- Oregon Health and Science University's Research Data Warehouse (OHSU)
- Pennsylvania State Electronic Health Records System (PSHEHR)
- Stanford Medicine Research Data Repository (Stanford)
- Tufts Medical Center Electronic Health Record (TMC)
- University of Massachusetts Chan Medical School Health Record (UMass Chan)
- University of Tennessee Health Science Center Enterprise Data Warehouse (UTHSC)

Other data sources did not include race, but were queried for gender differences (listed below):

- IBM MarketScan Commercial Claims and Encounters (CCAE)
- IBM MarketScan Medicare Supplemental Beneficiaries (MDCR)
- IQVIA Disease Analyzer (DA) Germany
- IQVIA Longitudinal Patient Database (LPD) France
- IQVIA Longitudinal Patient Database (LPD) Australia EMR
- IQVIA Longitudinal Patient Database (LPD) Italy
- IQVIA Longitudinal Patient Database (LPD) Belgium
- IQVIA Ambulatory EMR (AmbEMR)
- Japan Medical Data Center (JMDC)
- IQVIA Adjudicated Health Plan Claims (PharMetrics Plus)

### A.2. Outcome phenotypes and cohort definitions

To analyze differences in disease prevalence, we create and analyze differences for all patients in databases with observations in the same year (2017), split by age deciles.

The 2017 cohorts require three years of prior observation, and are split by age:

- "Persons observed in 2017 with 3yr prior observation (Age 0-9)"
- "Persons observed in 2017 with 3yr prior observation (Age 10-19)"
- "Persons observed in 2017 with 3yr prior observation (Age 20-29)"
- "Persons observed in 2017 with 3yr prior observation (Age 30-39)"
- "Persons observed in 2017 with 3yr prior observation (Age 40-49)"

- “Persons observed in 2017 with 3yr prior observation (Age 50-59)”
- “Persons observed in 2017 with 3yr prior observation (Age 60-69)”
- “Persons observed in 2017 with 3yr prior observation (Age 70-79)”

We additionally create 124 disease-specific cohorts defined which were split into 87 chronic and 37 acute conditions. These disease-specific cohorts were created to enable health equity researchers to assess differences in specific phenotype prevalences across datasources.

The 87 chronic condition phenotypes are: Abdominal pain, Allergic rhinitis, Alzheimer's disease, Ankylosing spondylitis, Anosmia or hyposmia or dysgeusia, Asthma, Atopic dermatitis and eczema, Attention deficit hyperactive disorder, Autism, B-cell lymphoma, Behçet's syndrome, Bipolar disorder, Celiac disease, Chronic kidney disease, Chronic lymphoid leukemia, Cirrhosis, Constipation, COPD, Crohn's disease, Dementia, Depression, Dysphagia, Dyspnea, Epilepsy, Erythema multiforme, Facial palsy, Gastritis Gastroesophageal reflux, Giant cell arteritis, Glaucoma, Granulomatosis, Guillain-Barré, Hepatitis C, Hidradenitis suppurativa, HIV, Hyperlipidemia, Hypertension, Hypoglycemia, Hypothyroidism, Idiopathic thrombocytopenic purpura, Inflammatory bowel disease, Insomnia, Jaundice, Lower back pain, Malignant neoplasm of kidney, Malignant neoplasm of liver, Malignant neoplasm of respiratory tract, Malignant tumor of bladder, Malignant tumor of colon, Malignant tumor of esophagus, Malignant tumor of stomach, Migraine incident, Multi-system inflammatory syndrome, Multiple myeloma, Multiple Sclerosis, Myelodysplastic syndrome, Myocarditis, Narcolepsy, Nausea, Neck pain, Non-Hodgkins lymphoma, Obesity, Osteoarthritis, Osteoporosis, Parkinson's, Peripheral vascular disease, Posttraumatic stress disorder, Psoriasis, Psoriatic arthritis, Psychosis, Pulmonary hypertension, Rheumatoid arthritis, Schizophrenia, Sclerosis, Sjogren's, Sleep apnea, Systemic lupus erythematosus, Takayasu's disease, Thromboangiitis obliterans, Thrombocytopenia, Thrombotic microangiopathy, Thyroiditis, Transverse myelitis, Type 1 diabetes, Type 2 diabetes, Type B hepatitis, Ulcerative colitis, Vasculitis associated with ANCA

The 37 acute condition phenotypes are: Acute disseminated encephalomyelitis, Acute kidney injury, Acute myeloid lymphoma, Acute myocardial infarction, Acute pancreatitis, Acute respiratory failure,Acute tubular necrosis, Anaphylaxis, Angioedema, Aseptic meningitis, Atrial fibrillation, Cardiac arrest, Cardiac arrhythmia, Cardiogenic shock, Chilblains, Deep vein thrombosis, Flu, Fracture of bone of hip region, Gastrointestinal bleeding, Gout, Heart failure, Hemorrhagic stroke, Hepatic failure, Ischemic stroke, Kidney stone, Optic neuritis, Otitis, Pneumonia, Pulmonary embolism, Respiratory failure, Sepsis, Stress cardiomyopathy, Stroke, Toxic shock syndrome, Transient ischemic attack, Tuberculosis

Phenotypes are based on existing OHDSI phenotypes. To account for the wide range of data sources, the clinical context for some phenotypes is broadened to account for differences in data models. All phenotype definitions are publicly available in the OHDSI phenotype library (https://github.com/OHDSI/PhenotypeLibrary), and directly implemented in the R package.

The mapping of which specific phenotype is used in the study is available at online, at: https://github.com/ohdsi-studies/HERACharacterization/blob/master/inst/settings/CohortsToCreateFeature.csv

### A.3. Additional considerations about the cohort definitions

Another aspect of the population relates to the temporal span of a database. Over time, medical guidelines change, new medication becomes available, and events such as the COVID-19 pandemic can significantly impact overall healthcare utilization. With that in mind, we also keep track of the time periods of each database. The 2017 cohorts are all based on the same date span, and as such are aiming to control for potential temporal drifts.

To minimize differences in clinical settings, geography, coding, and data collection processes across databases, population subgroups are created within each database and compared within database only.

The population available for creating subgroups can be altered by requiring hospitalization or an inpatient visit within the cohort definition. This is sometimes used in previous work on electronic health records to ensure that a patient in data is acutely sick. For some of the acute phenotypes, we have required hospitalization or an inpatient visit, but for most chronic phenotypes we have no visit requirement.

Some analysis require that a previous outcome event or other condition diagnoses can be observed for an individual in a database. But requiring such a pre-entry observation period in the database reduces the sample size of available patients. All cohort phenotype definitions used in this study require patients to have been continuously observed at least 365 days or more prior to cohort entry. The 2017 cohorts require 3 years of continuous observation.

### A.4. Population subgroups

The following population subgroups are considered in this study. We note that the analysis is restricted to a binary definition of gender as reported in most observational health data and that does not represent nuances of gender. Furthermore, in most data sources, gender and sex is conflated and coded as “Male/Female.” We also note that our focus on race is limited to Black and White, as reported in data sources (most often as self-reports or as recorded by administrative records), and similarly there is no concept of multi-racial identity, as is the case in most data sources.

- Sex alone
  - Male
  - Female
- Race alone
  - Black
  - White
- Race and sex combined
  - Black Female
  - Black Male
  - White Female
  - White Male

### A.5. Availability of demographic information

Black and White labels were chosen as the primary race labels because they are the largest subgroups available in the American OMOP databases (e.g., CUIMC, MDCD, OptumSES). We chose two races and two sexes to allow for binary comparisons across classes. For databases where race labels are not available (e.g., CCAE, MDCR, European IQVIA datasets, Japanese dataset, Pharmetrics Plus]), the HERA dashboard makes available data summarizing gender differences.

## B. Study Methods

**The research question** for this study is asked for each **cohort**, each **database**, and for three different **time windows** (365 days to 1 day prior to entry into cohort, at day of entry into cohort, and 1day to 365 days after entry to cohort), namely: **what are the differences in condition and drug prevalence among cohort subgroups** based on sex, race, and their intersection?

For each cohort, we compute the following:

- Absolute number of Female/Male, Black/White (for datasets where race is available), Female/Black, Female/White, Male/Black, Male/White patients
- Absolute number of conditions and drugs in cohort
- Prevalence of conditions and drugs for each subgroup in the cohort

To compare prevalence of a condition or treatment across subgroups, we calculate the risk ratio between a target and comparator subgroup, for binary comparison.

Given a cohort (e.g., persons observed in 2017 with 3yr prior observation, age 50-59), we define a target subgroup (Female observed in 2017 with 3yr prior observation, age 50-59) and a comparator subgroup (Male observed in 2017 with 3yr prior observation, age 50-59). Let $n_1$ be the number of Female patients in the cohort, and $n_2$ be the number of Male patients in the cohort. We compute the prevalence percentages of all conditions and drugs for the subgroups. Let $x_1$ be the number of Female patients who have at least one condition occurrence of "abdominal pain" in the time window of interest, and $x_2$ be the corresponding number for Male patients. The prevalence percentage for Female patients is thus $p_1 = \frac{x_1}{n_1}$. We can then use these prevalence percentages to calculate risk ratios for all conditions and drugs for target and comparator subgroups.

**Confidence intervals.** We compare the risk ratios across an aggregate of symptoms or medications by calculating confidence intervals for the risk ratio. For risk ratios (RRs), the variance of the log(RR) is calculated as: $Z_{critical}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$. [1]

**False discovery rates and false coverage rate control.** While the above method is satisfactory for a small number of comparisons of prevalence rates, the total number of comparisons of prevalence risk rates is large when considering all cohorts, databases, time windows, conditions, drugs, and subgroups.

To remedy this, we compute confidence intervals that control for the false coverage rate (the number of confidence intervals that do not correctly cover the probability of the true

hypothesis). We use the analog of the Benjamini-Hochberg procedure for false discovery rate control, the Benjamini-Yekutieli algorithm ($BY_q$) for false coverage rate control of multiple confidence intervals [2], which is also described and motivated in [3] (Chapter 20). The $BY_q$ algorithm works by first computing the p-values for all *N* hypothesis tests, ordering them in ascending order then defining a threshold *q* above which to reject a fraction of hypotheses as false, and computing the confidence intervals for the accepted hypotheses using the quantile α of the standard normal inverse cumulative distribution function.

**C. Example**

Given the following instance of the study's research question:
For each database, compare differences in condition and drug prevalence among Male and Female patients, in the 2017 cohort ages 50-59, and for the time window 1 year prior to entry into cohort.

This would be translated as selecting the following:

sex = "Male",
"Female"
domain = "Drug"
"Condition"
time-window = "-365d to -1d"
outcome = "Persons observed in 2017 with 3yr prior observation (Age 50-59)",
database = ["CUIMC", "CCAE", … , " PharMetrics Plus "]

**BIBLIOGRAPHY**


1. Confidence Intervals for Risk Ratios and Odds Ratios, 2021. Wayne W. LaMorte. Date last modified online: 21 April 2021, https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module8-CategoricalData/PH717-Module8-CategoricalData5.html

2. Benjamini Y, Yekutieli D, Edwards D, et al. False Discovery Rate: Adjusted Multiple Confidence Intervals for Selected Parameters [with Comments, Rejoinder]. Journal of the American Statistical Association 2005;100:71–93.

3. Efron B, Hastie T. Computer Age Statistical Inference. 2016; Stanford University:493.