RESEARCH PROTOCOL

PROPOSAL FOR PEDIATRIC CHARACTERIZATION ACROSS A NETWORK

Version 0.1

# List of Abbreviations

|  |  |
| --- | --- |
| CCAE | IBM MarketScan(R) Commercial Claims and Encounters |
| CDM | Common Data Model |
| CUIMC | Columbia University Irving Medical Center |
| DA | Disease Analyzer |
| EHR | Electronic Health Record |
| EHDEN | European Health Data & Evidence Network |
| JMDC | Japan Medical Data Center |
| MDCD | IBM MarketScan(R) Multi-State Medicaid Database |
| MDCR | IBM MarketScan(R) Medicare Supplemental Database |
| OHDSI | Observational Health Data Science and Informatics |
| OMOP | Observational Medical Outcomes Partnership |
| OPTUM\_DOD | Optum(R) de-identified Clinformatics©Data Mart Database Date of Death |
| OPTUM\_EHR | Optum(R) de-identified Electronic Health Record Dataset |
| CD | CohortDiagnostics |
| DD | Database Diagnostics |

# Responsible Parties

## Investigators

|  |  |
| --- | --- |
| **Investigator** | **Institution/Affiliation** |
| Jill Hardin | Epidemiology, Janssen Research and Development, LLC, Titusville, NJ, USA |
| Erica A Voss | Epidemiology, Janssen Research and Development, LLC, Titusville, NJ, USA |
| Clair Blacketer | Epidemiology, Janssen Research and Development, LLC, Titusville, NJ, USA |
| James P Gilbert | Epidemiology, Janssen Research and Development, LLC,Titusville, NJ, USA |
| Talita Duarte | Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain |
| Daniel Prieto -Alhambra | Pharmaco- and Device Epidemiology, CSM-NDORMS,  University of Oxford, Oxford, UK |

## Disclosures

This study is undertaken within Observational Health Data Sciences and Informatics (OHDSI), an open collaboration.

Data sources associated with the EHDEN network received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA.

JH, EAV, CB, and JPG are employees Janssen Research & Development, LLC and are Johnson & Johnson shareholders.

# Amendments and Updates

Table 3.1 lists any protocol amendments made over time.

Table 3.1: Protocol amendments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number | Date | Section of study protocol | Amendment or update | Reason |
|  |  |  |  |  |

# Milestones

Table 4.1 lists the study milestones.

Table 4.1: Study milestones

|  |  |
| --- | --- |
| Milestone | Planned / actual date |
| Start of analysis |  |
| End data collection |  |
| End of analysis |  |
| Results presentation |  |

# Rationale and Background

Use of data diagnostics for evaluation of the databases in the OHDSI network would be useful to identify those that would be suitable for use in pediatric and adult studies.

Characterization of conditions using the datasources available through the OHDSI network using the CohortDiagnostics (CD) tool would provide feasibility estimates for data partners to conduct studies. This data network could then be used for many types of studies ranging from descriptive to predictive across therapeutic areas. The output from CD are age distributions and clinical characteristics that would help identify datasources where pediatric studies could be conducted and would advance research, knowledge, and health in pediatrics.

Several pediatric data networks exist for a range of conditions (Appendix Table1) however, none enumerate subjects by individual data contributor. Some of these pediatric data networks have been converted to the common data model (CDM) format while others have not. Benefits of the proposed study is that because all the data is converted to the CDM.

Lastly background rates for the phenotypes included in Appendix Table 1 will be generated and evaluated. Knowing the background rates among children and adults is important because it will give visibility into how often these events occur in the in the age groups and databases included in this study.

# Study Objectives

The objectives of this study are to characterize several conditions and enumerate the number of pediatric subjects available for study using the observational databases available through the OHDSI network and using the CohortDiagnostics (CD) (https://github.com/OHDSI/CohortDiagnostics) and the cohortIncidence tool (<https://github.com/OHDSI/CohortIncidence>) for the phenotypes among the pediatric and adult populations.

Our aim is purely descriptive. The hope with this work is the evidence we provide can be used as a resources when evaluating fit for purpose datasources for pediatric studies.

# Research Methods

## Data Sources

We intend to study data with different data source provenance (e.g., electronic medical records, insurance claims) as well data representing different populations (privately insured employees or patients with limited income) and data with different origins (US and non-US). This study will be run on datasets that have been converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [1] [2].

These data sources include but are not limited to:

* IBM MarketScan® Commercial Claims and Encounters (CCAE)
* IBM MarketScan® Medicare Supplemental Beneficiaries (MDCR)
* IBM MarketScan® Multi-state Medicaid (MDCD)
* IQVIA® Disease Analyzer (DA) Germany (IQVIA\_GERMANY\_DA)
* IQVIA® Disease Analyzer (DA) France (IQVIA\_FRANCE\_DA)
* IQVIA® Australia Longitudinal Patient Data (IQVIA\_AUSTRALIA\_LPD)
* Japan Medical Data Center (JMDC)
* Optum® de-identified Electronic Health Record Dataset (OPTUM\_EHR)
* Optum® de-identified Clinformatics® Data Mart Database – Date of Death (OPTUM\_DOD)
* Various European Health Data & Evidence Network (EHDEN) Data Partners
* Various Observation Health Data Sciences and Informatics (OHDSI) Data Partners
* Other data sources from prospective data partners

## Study Populations

We will define all cohorts in the Atlas environment (<https://atlas.ohdsi.org/>). Detailed definitions can be found in the Appendix 2.

Phenotype algorithms have been built and evaluated for these conditions in existing observational databases converted to the CDM using version 5.1.3.

ATLAS is a web based open-source application and provides a unified interface to patient level data and analytics. ATLAS will be used to generate cohorts and their json representations will be used as the analytical inputs to the study.

### Target Cohort Definitions

For generation of background rates six target cohorts will be used and are listed in Appendix Table 2. The cohorts will be indexed on observation period start and end dates with the start date of the cohort representing the index date of the analysis. Patients with less than one year of observation prior to the index date will be excluded in order to better identify patients who have previously been diagnosed with the outcome of interest.

### Outcome Cohort Definitions

The background rates will use the incident phenotypes listed in Appendix Table 3 as the outcomes. Clean windows will include those listed in Appendix Table 3.

## Population Subgroups

The CD tool will be used to characterize the conditions of interest. A custom feature will be included to assess the relevant pediatric age groups using the custom covariate builder function <https://ohdsi.github.io/FeatureExtraction/articles/CreatingCustomCovariateBuilders.html>.

We will consider the following age subgroups independently and jointly.

Age groups

Pediatric Age Groups - FDA:

* <2
* 3-5
* 6-11
* 12-17

General:

* 0–5
* 6–17
* 18–34
* 35–54
* 55–64
* 65–74
* 75–84
* Greater Than 85 (>=)

Male/female

## Times at risk

For background rates several time-at-risk windows will be used to examine the impacts on the sample size and provide rate stability.

We will include four TARs, relative to the visit date which is the index date for this objective. Days 1 – 30 days (Four Weeks)

* Days 1 – 30 days
* Days 1 – 90 days
* Days 1 – 180 days
* Days 1 – 365 days

## 7. 5 Analysis

Incidence will be calculated as both a rate and a proportion, and calculated for all four potential times-at-risk noted above. Incidence proportion is defined as the number of patients in the target cohort who enter the outcome cohort during the time at risk divided by the total number of patients in the target cohort. Incidence rate is defined as the number of outcome events that occur within the time-at-risk among the target cohort episodes divided by the total time at risk for all patients in the target cohort.

## 7.6 Study outputs

We will generate descriptive statistics using the OHDSI tools Cohort Diagnostics and CohortIncidence.

The following will be performed for every subgroup and TAR combination:

* Absolute number of outcomes
* Absolute number of outcomes prior to TAR start date
* Absolute number of patients at risk
* Absolute number of patients with outcomes prior to TAR start date
* Absolute number of patients with outcomes within TAR interval
* Incidence proportion of outcomes within TAR interval
* Incidence rate of outcomes within TAR interval

In the above outcome counts, multiple data elements (e.g., multiple diagnosis codes) within a clean window count as one outcome. Patients will contribute time-at-risk from the index date (date of first visit) until the earliest of 365 days after the index, their observation period end date, or the start date of the outcome event. Persons with prior outcome events will not begin to contribute time-at-risk until the clean window requirement is satisfied.

## 7. 6 Tools to conduct analyses

The CohortDiagnostics (CD) module will be used to characterize the cohorts listed in Appendix Table 2. The GitHub repo for the CohortDiagnostics tool is located here: <https://github.com/OHDSI/CohortDiagnostics>

A publicly available shiny application will be deployed using the OHDSI ShinyDeploy server located here: <https://github.com/OHDSI/ShinyDeploy>.

Background rates will be calculated via the R package <https://github.com/OHDSI/CohortIncidenceModule>.

## 7.7 Logistics of Network Analyses

Executing a Federated StudySites will run the study analysis package locally on their data coded according to OMOP CDM. Only aggregate results will be shared with the study coordinator. Result files will be automatically staged into a ZIP file that can be transmitted using the OhdsiSharing R Library (http://ohdsi.github.io/OhdsiSharing/) or through a site’s preferred SFTP client using a site-specific key provisioned by the OHDSI Study Coordinator. Local data stewards are encouraged to review study parameters to ensure minCellCount function follows local governance. At a minimum, it is encouraged to keep this value to >5 to avoid any potential issues with re-identification of patients. Note: covariates are constructed using controlled ontologies from the OMOP standard vocabularies though some labels may be replaced with publication-friendly labels due to space restrictions of the submitting journal.

# Strengths and Limitations

## Strengths

* The fully specified study protocol is being published before analysis begins.
* Dissemination of the results will not depend on estimated effects, avoiding publication bias.
* All analytic methods have previously been verified on real data.
* All software is freely available as open source.
* Use of a common data model allows extension of the experiment to future databases and allows replication of these results on licensable databases that were used in this experiment, while still maintaining patient privacy on patient-level data.
* Use of multiple databases allows estimating consistency to add credibility and supports generalizability.
* Because this is a federated, distributed network study, the data holders are in control of if they want to participate both before and after results are generated.

## Limitations

* Misclassification of study variables is unavoidable in secondary use of health data, so it is possible to misclassify treatments, covariates, and outcomes; we do not expect differential misclassification, so bias will most likely be towards the null.
* The electronic health record databases may be missing care episodes for patients due to care outside the respective health systems; bias will most likely be towards the null.

# Protection of Human Subjects

This work does not involve human subjects research. The project does, however, use de-identified human data collected during routine healthcare provision. Confidentiality of subject records will be maintained always. All study reports will contain aggregate data only and will not identify individual subjects or physicians. At no time during the study will the sponsor receive subject identifying information.

All data partners executing this study within their data sources will have received their own approvals in accordance with their institutional governance prior to execution. This study will be executed across a federated and distributed data network, where analysis code is sent to participating data partners and only aggregate summary statistics are returned, with no sharing of patient-level data between organizations.

# Management and Reporting of Adverse Events and Adverse Reactions

This study uses coded data that already exist in electronic databases. In these types of databases, it is not possible to link (i.e., identify a potential causal association between) a particular product and medical event for any specific individual. Thus, the minimum criteria for reporting an adverse event (i.e., identifiable patient, identifiable reporter, a suspect product and event) are not available and adverse events are not reportable as individual adverse event reports. The study results will be assessed for medically important findings.

# Plans for Disseminating and Communicating Study Results

Open science aims to make scientific research, including its data process and software, and its dissemination, through publication and presentation, accessible to all levels of an inquiring society, amateur or professional [8]. Open science delivers reproducible, transparent, and reliable evidence. All aspects of study (except private patient data) will be open, and we will actively encourage other interested researchers, clinicians and patients to participate.

## Transparent and re-usable research tools

The protocol and all source code will be available below prior to execution across the federated and distributed network.

Protocol: <<ENCEPP LINK>>

Protocol: <<OHDSI GITHUB PROTOCOL>>

Code: <https://github.com/ohdsi-studies/PediatricCharacterization>

## Continuous sharing of results

We will store and openly communicate all of these results as they become available using a user-friendly web-based app.

## Scientific meetings and publications

We will deliver multiple presentations at scientific venues and will also prepare multiple scientific publications for clinical, informatics, and statistical journals.

## General public

We believe in sharing our findings that will guide clinical care with the general public. This study will use social-media (Twitter-OHDSI) to facilitate this.

# References

1. Observational medical outcomes partnership common data model (OMOP CDM). <https://github.com/OHDSI/CommonDataModel>
2. Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. Studies in Health Technology and Informatics 2015;216:574–8.

# Appendix

## Appendix 1

**Appendix Table 1. Overview of existing pediatric data networks**

| **Name** | **Geographical Area/Catchment** | **Years of Data** | **Research Focus** |
| --- | --- | --- | --- |
| The Children’s Data Network ([CDN](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9053132/)) | children born in California – CA birth and death records; incarceration records; social service and education data | 2014-present | Evidence based policy research |
| [PEDSnet](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4078288/pdf/amiajnl-2014-002743.pdf) | eight children’s hospitals,  two existing patient-centered disease-specific pediatric networks addressing inflammatory bowel disease  and complex congenital heart disease, a newly  formed pediatric obesity network, and two national  data partners; includes observational data on over 5 million children with at least 1 clinical encounter and at least 1 coded diagnosis during or after 2009. Specific partners include:  Children’s  Hospital of Philadelphia (CHOP), Cincinnati  Children’s Hospital Medical Center (CCHMC),  Children’s Hospital Colorado, Nemours Children’s  Health System, Nationwide Children’s Hospital,  St. Louis Children’s Hospital, Seattle Children’s  Hospital, and Boston Children’s Hospital. With  regional catchment areas extending across 22 states; ImproveCareNow (ICN) Inflammatory  Bowel Disease Network, the National Pediatric  Cardiology Quality Improvement Collaborative  (NPCQIC) and  Healthy Weight  obesity prevention network. | 2009-present | Variety of outcomes and [data quality](https://academic.oup.com/jamia/article/24/6/1072/3238563) |
| The Shriners Hospitals for Children Learning Health System [SHOnet](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9284925/pdf/LRH2-6-e10305.pdf) | deidentified data  extracted from the electronic health record (EHR) from the Shriners Hospitals for  Children speacialty pediatric health care system. Two million patient  encounters between 2011 | 2011-present | Cerebral palsy (new network so just getting started) |
| The Newborn Screening Translational Research Network ([NBSTRN](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8293037/)) | Data on infants diagnosed at birth. Provides tools and resources for researchers to conduct research relevant to NBS programs for rare diseases for which data has been siloed locally. Infrastructure includes tools for the analytical and clinical validation of screening tests; the collection, analysis, sharing, and reporting of longitudinal laboratory and clinical data on newborn-screened individuals; developed the Longitudinal Pediatric Data Resource (LPDR) to capture, store, analyze, visualize, and share genomic and phenotypic data over the lifespan of NBS identified newborns |  | Muscular Dystrophy; Sickle Cell Disease; primary immunodeficiency |
| Pediatric Research in Office Settings ([PROS](https://www.aap.org/en/research/pediatric-research-in-office-settings-pros/about-pediatric-research-in-office-settings-pros/)) | Practice-based research network of the American Academy of Pediatrics (AAP)  Established in 1986  > 400 unique practices have participated in at least 1 study since 2010 |  | Obesity, flu vaccine, hpv vaccination; child poverty; adolescent substance abuse |
| Comparative Effectiveness Research through Collaborative Electronic Reporting ([CER2](https://www.aap.org/en/research/pediatric-research-in-office-settings-pros/comparative-effectiveness-research-through-collaborative-electronic-reporting/)) | 222 Practicing Sites  27 States  2119 Pediatricians and nurse practitioners  >2 Million Covered Lives |  | Pediatric Hypertension  ADHD diagnosis and prescribing  Patterns of prescribing for psychotropic mediactions  Prescribing patterns and safety profile of antipsychotics risperidone and aripiprazole  Benchmarking antibiotic use to inform outpatient stewardship |
| The Pediatric Research Consortium [PeRC](https://www.research.chop.edu/pediatric-research-consortium/publications) | the primary care practice-based research network at Children’s Hospital of Philadelphia; launched in 2002 |  | Adolescent health; antibiotic stewardship; asthma; ADHD; Autism spectrum disorder; behavioral health; cerebral palsy; community acquired pneumonia; concussion; developmental delay |
| Pediatric Practice Research Group ([PPRG](https://www.luriechildrens.org/en/research/research-areas/health-services-policy-research/pprg/)) | A partnership of the Department of Pediatrics at Lurie Children’s and over 50 pediatric practices. PPRG operates within the Mary Ann & J. Milburn Smith Child Health Outcomes, Research and Evaluation Center of the Manne Research Institute. |  | Atopic dermatitis – survey and chart review |
| Pediatric PittNet: University of Pittsburgh CTSI PBRN ([Pediatric PittNet](https://pedspittnet.pitt.edu/)) | Pediatric PittNet is a practice-based research network that supports collaboration between pediatric primary care practices and University of Pittsburgh researchers to improve the health of children. |  | Pediatric plaque psoriasis; several other outcomes ranging across therapeutic areas |
| Oklahoma Child Health Network ([OCHRN](https://octsi.ouhsc.edu/ochrn)) | collaborative effort of the Oklahoma Chapter of the American Academy of Pediatrics, Department of Pediatrics, and the OU Medical Center – Oklahoma only; Contributing to studies are pediatricians, primary care practitioners and specialists - specialty clinics are Gastroenterology, Pulmonology, Cardiology, and Neurology |  | vaccine decision making, parent-provider partnerships, referral communication and immunizations. |
| Continuity Research Network ([CORNET](https://www.academicpeds.org/groups-networks/research-networks/cornet/)) | 125 pediatric residency programs with nearly 6,500 pediatric residents caring for over 1 million patients across the United States and Canada. |  | To study and improve health, healthcare, and inequities in health and healthcare delivery among children and families, particularly those disproportionately affected by health inequities.  To study and improve resident education in pediatric continuity practices.  To engage residents in pediatric primary care research. |
| Cardiac Specific Pediatric | | | |
| Pediatric Heart Network ([PHN](https://www.pediatricheartnetwork.org/)) | research across a group of hospitals in the United States and Canada to accelerate the discovery of new treatments for congenital and pediatric-acquired heart disease |  | goal of the PHN is to improve outcomes and quality of life for individuals with congenital and pediatric-acquired heart disease |
| National Pediatric Cardiology-Quality Improvement Collaborative (NOTE: this is part of PEDSNET) |  |  |  |
| The Fontan Outcomes Network ([FON](https://www.fontanoutcomesnetwork.org/care-centers-overview)) | US based - 12 care centers that have worked to establish a strong foundation for our network. FON welcomed an additional 19 care centers to the network in summer 2022. | Launched in 2021 |  |
| The Australian and New Zealand Fontan Registry https://www.fontanregistry.com/ | Australia and New Zealand only - population-based registry collecting health information on all patients who have undergone the Fontan procedure (used to treat single-ventricle congenital heart disease) living in Australia and New Zealand. |  |  |

## Appendix 2

**Target Cohorts:**

|  |  |
| --- | --- |
| **Cohort Name** | **JSON file # (included in package)** |
| persons at risk at start of year 2017-2022 with 365d prior observation | 14061 |
| persons at risk at start of year 2017 with 365d prior observation | 14063 |
| persons at risk at start of year 2018 with 365d prior observation | 14064 |
| persons at risk at start of year 2019 with 365d prior observation | 14065 |
| persons at risk at start of year 2020 with 365d prior observation | 14066 |
| persons at risk at start of year 2021 with 365d prior observation | 14067 |
| persons at risk at start of year 2022 with 365d prior observation | 14068 |

**Outcome Cohorts - Conditions and Clean Windows:**

|  |  |  |
| --- | --- | --- |
| **Cohort Name** | **JSON file # (included in package)** | **Clean Window** |
| Ulcerative Colitis | 10606 | 9999 |
| Autism | 3417 | 9999 |
| Crohn's Disease | 10616 | 9999 |
| Plaque Psoriasis | 10626 | 9999 |
| Major Depressive Disorder | 10628 | 9999 |
| Attention Deficit Disorder | 10640 | 9999 |
| Multiple Sclerosis | 10641 | 9999 |
| Chronic Lymphocytic Leukemia | 10642 | 9999 |
| Type 2 diabetes mellitus | 10647 | 9999 |
| Anaphylaxis non environmental | 10659 | 30 |
| Urinary tract infection | 12396 | 30 |
| Epilepsy | 12403 | 9999 |
| Migraine | 12468 | 30 |
| Burns | 13676 | 30 |
| Cystic Fibrosis | 13626 | 9999 |
| Down Syndrome | 13625 | 9999 |
| Motor Vehicle Accident | 13692 | 30 |
| Firearm Accident | 13631 | 30 |
| Suicide ideation attempt, including drug poisoning | 3429 | 30 |
| Asthma | 14041 | 9999 |
| Atopic Dermatitis | 12734 | 9999 |
| Otitis Media | 372 | 30 |