# RESEARCH PROTOCOL:
# How Often? All-by-all Drug-Condition Incidence Rate: Protocol for an OHDSI Network Study

**Authors:**
George Hripcsak, MD, Columbia University
Patrick Ryan, PhD, Janssen Research and Development

**Date:** September 5, 2023

**Acknowledgment:** The analysis is based in part on work from the Observational Health Sciences and Informatics collaborative. OHDSI (http://ohdsi.org) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.
The authors declare the following disclosures: Dr. Ryan is an employee of Janssen Research & Development.

## 1. List of Abbreviations

| | |
|---|---|
| ADE | Adverse Drug Event |
| HES | Hospital Episode Statistic |
| CDM | Common Data Model |
| EHDEN | European Health Data and Evidence Network |
| OMOP | Observational Medical Outcomes Partnership |
| OHDSI | Observational Health Data Science and Informatics |
| RxNorm | US-specific terminology that contains all medications available on the US market |
| SNOMED | Systematized Nomenclature of Medicine |

## 2. Responsible Parties

| **Investigator/Author** | **Institution/Affiliation** |
|---|---|
| George Hripcsak | Department of Biomedical Informatics, Columbia University, NY, USA |
| Patrick Ryan | 1) Janssen Research and Development, Titusville, NJ, USA 2) Department of Biomedical Informatics, Columbia University, NY, USA |
| Elise Ruan | Department of Biomedical Informatics, Columbia University, NY, USA |

## 3. Abstract

In this study, we use large-scale observational data to estimate the incidence rate of clinical conditions in two sets of cohorts of patients, (1) cohorts defined in the OHDSI phenotype library and (2) cohorts defined as the post-exposure period following initiation of all marketed drugs.

## 4. Rationale and Background

Incidence rates serve many purposes in clinical medicine. The background rate of disease may be of interest to assign resources for detection and treatment, background rates may be used as comparisons for disease surveillance such as in vaccine adverse event detection, and post-treatment incidence rates may be used to assess the potential impact of possible adverse events due to the treatment. Given valid cohort definitions, incidence rates can be calculated efficiently and without the complex analysis that would be required for causal inference.

We propose two sources of cohorts. One is to enlist the OHDSI community to specify a set of clinical problems of interest, expressed as a set of cohort definitions of target populations relevant to those

problems and the cohort definitions of the outcomes whose incidence is to be estimated within those targets. Each OHDSI workgroup or interest group will define and validate the cohort definitions and place them in the OHDSI phenotype library. Incidence rates will be estimated across the network for those outcomes within those target populations.

The second source addresses drug safety. The vast majority of potential drug side effects have never been actually measured for any drug. Generally, only suspected side effects and side effects that are sufficiently common are detected and measured. The rest are assumed to be negative. The incidence of effects after initiation of a drug does not imply causality or its lack, but it does indicate the impact of potential side effects; if the rate is sufficiently low, then the effect is of less concern. The SNOMED CT disorder concept hierarchy offers the opportunity approximate side effect definitions based on billing codes and problem lists. While these are not validated phenotype definitions, they represent an incremental step towards understanding side effects. Given the limited data on side effects, making this simple calculation of absolute risk over a large scale can help clinicians and patients start the conversation of risk with an empirical number as the upper bound of risk. The drugs are defined as drug ingredients on the market specified in RxNorm and RxNorm Extension.

## 5. Objective

We aim to identify the incidence of clinical conditions and other events within defined populations for set periods of time. This analysis will be executed across multiple databases and both summarized numerically and graphically from an aggregate table as well as detailed for each individual database. These results will then be shared via a searchable, public web site.

## 6. Methods

The study is an observational cohort study based on routinely-collected health care data which is mapped to the OMOP CDM.

### 6.1 Data Sources

The analyses will be performed across a network of observational healthcare databases in the Observational Health Data Sciences and Informatics (OHDSI) international initiative. All databases will have been transformed into the Observational Medical Outcomes Partnership (OMOP) Common Data Model, version 5. This common data model converts data across multiple countries, languages, and vocabularies to a set of standard vocabularies in a common database schema. Databases are managed locally by the different sites and data owners, study code is shared open-source across sites, code is executed locally, and the results are returned to the coordinating center for the study.

The complete specification for OMOP Common Data Model version 5 is available at: https://github.com/OHDSI/CommonDataModel. We will run the study package on the CUIMC database for testing, and then on the OHDSI Network. CUIMC is defined as follows:

**Table 1.** Data sources formatted to the OMOP CDM participating in the study

| Abbre-viation | Data Source | Population Description | Sample Size | Data Type |
|---|---|---|---|---|
| CUIMC* | Columbia University Irving Medical Center | Patients of the Columbia University Irving Medical Center (New York City, USA) | 6 million | Inpatient and outpatient electronic health record data |

## 6.2 Target Cohorts

For the first cohort source, we will use the target cohorts defined in the OHDSI Phenotype Library, with the start date of the cohort representing the index date of the analysis. For the second cohort source, we will generate an exposure cohort for each drug ingredient in RxNorm and RxNorm Extension, defining the index date for the analysis as the first occurrence of a drug exposure in a patient's longitudinal record and having them exit the cohort at the time of discontinuation of the drug or a lapse in treatment

For both sources, we will exclude patients with less than one year of observation prior to the index date in order to better identify these patients who have previously been diagnosed with the outcome of interest.

## 6.3 Follow-up

Time-at-risk is defined as starting one day after the index date and ending 1) 30 days after the index date (regardless of cohort end date), 2) 365 days after the index date (regardless of cohort end date), 3) when the patient exited the cohort ("on treatment" for target cohorts based on drug exposure), or 4) the end of the patient's observation period in the database ("intent to treat" for target cohorts based on drug exposure).

## 6.4 Outcomes of Interest

For the first cohort source, we will use the outcome cohorts defined in the OHDSI Phenotype Library, with the start date of the cohort representing initiation of the outcome. For the second cohort source, we will generate outcome cohorts based on disorder concepts in the SNOMED CT hierarchy, creating a cohort for each concept and its descendants and excluding top-level terms such as a general disorder.

### 6.5 Stratifications

Each target cohort will be analyzed in full and stratified on factors based on the following pre-index characteristics, all stratum pending meeting minimum requirement of reportable cell counts (>2500):

- Sex (Male vs. Female)
- Age groups : 0-2, 3-12, 13-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+
- Index year:  2002-2022
- Source database

### 6.6 Analysis

All analyses will be performed using code developed for the OHDSI Methods library. The code for this study can be found at https://ohdsi.github.io/CohortIncidence/ and https://github.com/ohdsi-studies/HowOften.

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.

### 6.7 Logistics of Executing a Federated Analysis

Sites 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.

## 7. Sample Size and Study Power

The study package is designed to suppress any analyses which have less than 2500 unique persons. This cut point was informed by a power calculation to assess the computational cut point of when a cell count would be too small to merit additional subdivision within the target-stratum-feature combination. This means that each data owner will only generate results for target-stratum-feature pairs that meet this minimum threshold.

## 8. Strengths and Limitations

### 8.1 Strengths

We hope to generate the world's largest observational sets of analysis of secondary health data for the proposed incidence rates. We are running a multi-country, multi-center characterization study to

estimate the incidence rates. The use of a common data model and standard vocabularies ensure interoperability and portability of phenotypes utilized in this analysis. The use of a federated study model will ensure no movement of patient-level data from institutions participating in this analysis. This is critically important to ensure the protection of patient privacy in the secondary use of routinely collected patient data. Data custodians will remain in control of the analysis run on these data and will conduct their own site-based validation processes to evaluate case reports against public health reporting.

### 8.2 Limitations

Target and outcome phenotyping may be inaccurate as it is based on real world data, with the absence of such records taken to indicate the absence of a condition or event. Presence of records such as medication events indicate that an individual was prescribed or dispensed a particular drug, but this does not necessarily mean that an individual took the drug as originally prescribed or dispensed. The index date may also be misclassified depending on when the patient actually underwent an event versus when a record is recorded in the database.

We will communicate to viewers of the results via the searchable website that the findings are incidence rates based on observational data and do not reflect causality.

## 9. Protection of Human Subjects

Confidentiality of patient records will be maintained at all times. Data custodians will remain in full control of executing the analysis and packaging results. There will be no transmission of patient-level data at any time during these analyses. Only aggregate statistics will be captured. Study packages will contain minimum cell count parameters to obscure any cells which fall below allowable reportable limits. All study reports will contain aggregate data only and will not identify individual patients or physicians.

## 10. Plans for Disseminating and Communicating Study Results

The results will be used across multiple papers by the target cohorts, by stratification features, or by baseline characteristics, or outcomes. At least one paper will be written and submitted for publication to a peer-reviewed scientific journal. Study results will be shared via a searchable public web site after completion of the study. The results will also be presented at OHDSI in-person or virtual events.