****

1. Research Question (include all primary and secondary objectives)

The main research question of this study is: *What are the population-level characteristics of the databases within the OHDSI federated network?*

The specific aims of this study are as follows:

- To generate and compare data counts and summary statistics of the databases within the OHDSI network (that the data owners can provide in compliance with IRB, GDPR, HIPAA) to support future network research and inform study feasibility.
- To describe the databases across the OHDSI network and inform data owners about the quality of their data by learning what a "typical" OMOP CDM standardized database looks like. This will be done by characterizing the heterogeneity, granularity, timeliness, and domain coverage of the participating databases.

2. Background (briefly describe relevant information to justify the research)

The Observational Health Data Sciences and Informatics (OHDSI) federated network is a collaborative effort aimed at leveraging healthcare data from multiple institutions for large-scale federated observational research. In its current state there are over 500 data sources from over 49 countries mapped to the OMOP Common Data Model, the standard that enables such ambitious evidence generation. One major challenge of federated network studies is the assessment of network data quality, study feasibility and data fitness-for-use across these data sources in such a way that does not strain the time and resources of data holders while still supporting rigorous evidence generation that engenders trust and buy-in from the larger research community.

To facilitate collaborative research efforts and ensure the quality and integrity of the data across the OHDSI network, it is imperative to understand the characteristics and variability of the databases within the network. This study aims to collect summary statistics from participating sites to describe the databases and learn about the network as a whole. The output of the study will inform and enhance the research capabilities of the OHDSI community by enabling rapid data quality and fitness-for-use assessments.

3. Methods

a. Describe the study design and the database(s) that will be utilized, including the specific sources from which you will collect data or samples. Include in your description whether you will use clinical data (e.g., EPIC) or a research resource (provide IRB protocol number) Provide your inclusion/exclusion criteria and describe your method of case/patient/sample identification.

This study will be an international observational, retrospective cohort study based on secondary EHR and healthcare claims data which have been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model

(CDM). The OMOP CDM is HIPAA compliant, created to support network research studies.

The study will utilize the aggressively deidentified projection of the [Health System] OMOP CDM provided by the [IRB#] resource protocol. The CDM provided includes patient data from approximately 2.1 million patients seen at [Health System] since July 2016. All data available in the OMOP CDM version 5.4 will be included in this study.

This study will not involve consenting/recruitment/remuneration/in-person activities as it will only perform a secondary analysis of existing clinical data. No chart review will be performed at this site. Upon IRB approval we plan to execute pre-defined queries and summary statistics using the DbDiagnostics R package, available on GitHub (https://github.com/ohdsi/DbDiagnostics), against local databases to generate summary statistics. The study team will also analyze aggregate-level descriptive information about the data and develop network-based characterizations.

Summary statistics, query outputs, and all other results from this analysis will be stored on a [Secure Computing Environment] shared folder with access granted only to the PI and study team members listed on the corresponding IRB protocol (IRB##). The study team will use SFTP to make aggregated results from the DbDiagnostics analysis temporarily available for sharing with the study coordinating center. The summary statistics generated will be kept by the OHDSI Coordinating Center for reuse to support network research.

Inclusion Criteria:

- All patients with available clinical data in the aggressively-deidentified OMOP database will be included in this study (IRB## resource protocol).

b. If your study involves data/biospecimens from participants enrolled under other research studies with a written consent or under a waiver of consent, please list the IRB application numbers for those studies. Please note: Certificate of Confidentiality (CoC) protections applied to the data in source studies funded by NIH or CDC will extend to this new study if the funding was active in 2016. If this situation applies, Section 36, question 6 in the application will need to be answered "Yes" and "Hopkins Faculty" should be selected in question 7. No other documents are required.

N/A

c. Clarify whether there was an ethical review process for initial collection/derivation of data/biospecimens. If applicable, provide the determination of the ethics committee or IRB study number. Indicate whether consent was obtained and whether the consent covered the use as proposed in this research.

The resource protocol from which this data is obtained underwent a separate IRB review. The protocol number for the approved resource is IRB IRB00445313.

d. If biological materials are involved please describe all the experimental procedures and analyses in which they will be used.

N/A

e. Specify the targeted number of individuals from whom you plan to include data/samples in this secondary use. Please be sure to specify the initial/largest cohort of eligible cases from which you will identify the final sample. Where applicable, please include an estimate of the time period that will be covered (e.g., will you include data within a certain range of dates?)

This study will include all 2.1 million patients and their records contained in the aggressively-deidentified OMOP database which covers 01/2016 through 06/2024.

f. Explain how your data are being extracted (manual chart review, bulk query). If you are planning to collect data from text documents (e.g., pathology/radiology reports) specify exactly how this will be accomplished. Are you planning to download text documents themselves? Storing copies of original documents from EPIC requires consultation with the CCDA and identification of an honest broker.

N/A. No chart review or data extraction will be done locally at [Health System].

g. Explain how your data are being recorded (paper, laptop, etc.).

N/A.

h. Explain how the data are being moved to the final storage location.

CDM data, as well as intermediary tables generated by the study analysis will be stored in a [Compute Platform] network managed by [Health System] IT in accordance with the parent protocol. Summary statistics, query outputs, and all other results generated from the database will move directly into the [Secure Computing Environment] study folder accessible only to study team members.

i. Provide the name and location of the server where the data will be stored.

The specific OMOP CDM projection that will be used in this study is housed on the server: [server.name.com] within the [Compute Platform] network managed by [Health System] IT.

j. Provide the name of the study team member responsible for data management and security.

[Study Team Member]

k. Provide any plans for de-identification of the dataset. Identifiers (MRN, Name) should be stored in a separate file with the data file using unique IDs.

We are requesting access to the projection of the OMOP CDM which has been determined by the CCDA to meet criteria for a limited dataset and has had sensitive data removed as described by the parent protocol (IRB##).

l. Explain how access to the data will be controlled and whether the access is logged.

Benjamin Martin and Haeun Lee as data managers will fully monitor and control access to database. We will grant access to the study folders via SAFE.

m. List the computer programs being used to store and to analyze the data.

R, Python, SQL, [Secure Computing Environment], and the HADES methods library maintained by OHDSI.

n. If you are using data from several sources explain what variables will be used to merge files.

N/A

o. Will the data set include any sensitive information (e.g., HIV status, psychiatric diagnosis)?

No. The LDS CDM will not include any sensitive information. Removal of sensitive information from the LDS will be coordinated with the CCDA as outlined in the parent protocol

p. Will the data set include any genomic data?

No

q. Will the data be used in collaborative efforts with other institutions? Describe the nature of the collaboration.

If data will be shared as part of this collaboration, please describe

i. The data that will be shared (e.g. individual-level data; summary-level data; aggregate data)
ii. The purpose of the data sharing

iii. How the sharing will be accomplished
iv. The security measures in place for the transfer of data to collaborators

No, the data provisioned for this study will not leave the [Health System] covered entity. Only publishable, aggregated results will be shared between institutions at any time.

r. Provide an estimate of how long it will take you to complete the study, including the time for data analysis.

2 years

s. Please describe any plans for data dissemination, including but not limited to, any plans to deposit data in external or internal repositories or share individual (person)-level data for publication purposes. Please consult the IRB's guidance on Sharing of Individual Level Research Data and describe how your dissemination plan aligns with these guidelines. Please be specific about the data that will be disseminated (including any tools needed to interpret it) and any methods that will be used to prepare the data for dissemination. If there are no plans nor requirements for individual-level sharing and data dissemination will consist of aggregate or summary-level data in publication, please state so. (If you have submitted a formal Data Management and Sharing Plan for funding purposes, please reference the plan here. These plans should be uploaded in Section 9 Item 4 of your eIRB application.)

This study has no plans nor requirements for sharing individual-level data. Any data shared/disseminated will only consist of aggregate or summary-level data and results. Specific summary statistics that will be generated and disseminated in this analysis are included in Section 4, Study Statistics (below). The summary statistics generated will be kept by the OHDSI Coordinating Center for reuse to support network research.

4. Study Statistics

We will use the DbDiagnostics R package that will run SQL queries on each site to produce descriptive statistics. The comprehensive list of summary statistics and counts that will be generated is as follows:

- Number of persons
- Number of persons by gender
- Number of persons by year of birth
- Number of persons by race
- Number of persons by ethnicity
- Number of persons with at least one day of observation in each month
- Number of persons with by observation period start month

- Number of persons with by number of observation periods
- Number of persons with by length of observation period, in 30d increments
- Number of persons with at least one visit occurrence, by visit_concept_id
- Number of distinct patients that overlap between specific domains – including death
- Number of persons with at least one concept_id, by measurement_concept_id

- Number of measurement occurrence records, by measurement_concept_id
- Number of measurement occurrence records, by measurement_source_concept_id
- Number of measurement records, by measurement_concept_id and value_as_concept_id
- Number of measurement records with no value (numeric, string, or concept)
- Number of persons with at least one concept_id, by condition_concept_id

- Number of condition occurrence records, by condition_concept_id
- Number of condition occurrence records, by condition_source_concept_id
- Number of persons with at least one concept_id, by procedure_concept_id
- Number of procedure occurrence records, by procedure_concept_id
- Number of procedure occurrence records, by procedure_source_concept_id
- Number of persons with at least one concept_id, by drug_concept_id
- Number of drug exposure records, by drug_concept_id
- Number of drug exposure records, by drug_source_concept_id
- Number of persons with at least one concept_id, by observation_concept_id
- Number of observation occurrence records, by observation_concept_id
- Number of observation occurrence records, by observation_source_concept_id
- Number of observation records, by observation_concept_id and value_as_concept_id
- Number of persons with at least one concept_id, by device_concept_id
- Number of device exposure records, by device_concept_id
- Number of device exposure records, by device_source_concept_id
- Distribution of numeric values, by measurement_concept_id and unit_concept_id

Metadata:

- Dataset name
- Name of the owner or licensee of the dataset
- Dataset DOI, if applicable
- Type of data in the database (EHR, administrative claims, clinical registry, etc.)
- OHDSI Standardized Vocabularies version
- Name of contact person responsible for network studies on the database
- Email address of contact person responsible for network studies on the database
- If the site has participated in an OHDSI network study before
- If there is someone at the site who can run an OHDSI study package
- How long in weeks it takes to get approval to run a study using the dataset
- How often the data are refreshed

5. Risks

a. Address the risk of loss of confidentiality.

The OMOP CDM and OHDSI methodologies make loss of confidentiality extremely unlikely. Despite efforts to secure the data, there is a risk of loss of confidentiality should an accidental breach occur with the provided CDM. Mitigation strategies are described by the parent protocol.

b. Discuss the steps you are taking to minimize this risk.

We will follow all Federal and state laws and [Health System] policies regarding patient privacy and HIPAA. All study team members have completed all clinical training modules and research modules regarding patient privacy and HIPAA. Investigators outside of the covered entity will complete HIPAA training as required by the Privacy Office. Please refer to parent protocol (IRB##) for a description of risk minimization.

c. Identify whether there are any additional risks and how you will minimize these risks.

N/A

d. Discuss your plan for reporting unanticipated problems or study deviations.

Unanticipated problems or study deviations will be reported to the IRB within 10 days of the PI becoming aware of such problems or deviations in accordance with the IRB policy on prompt reporting.

6. Requested Variables (Upload your data collection form in Section 20, Q 2 of the application. Do not use general terms, i.e. medical history. Be specific about what you plan to collect and indicate any coding scheme that will be used, e.g. yes/no.)

This CDM projection includes all elements from the Precision Medicine Analytics Platform (PMAP) which have been mapped to the OMOP CDM. The OMOP CDM contains an ontology of over nine million potential mappings. Due to the size of the mapped vocabulary, it is not practical to include a separate data specification sheet. An interactive tool for viewing the CDM vocabulary is available at https://athena.ohdsi.org/. The database specification for the OMOP CDM tables and columns for version 5.4 is outlined at https://ohdsi.github.io/CommonDataModel/cdm54.html.