|  |
| --- |
|  |
| **Observational Health Data Sciences and Informatics (OHDSI) Research Network\*** |
|  |
|  |
| **Evaluating Reliability of the Recently Published Observational Research Findings on the Association of Rivaroxaban vs Apixaban with Major Ischemic**  **or Hemorrhagic Events in Patients with Atrial Fibrillation** |
|  |
| 21 May, 2022  **Protocol** |

\* The Observational Health Data Sciences and Informatics (OHDSI) Research Network is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics, via open-source research and analytic tools. OHDSI has established an international network of researchers and observational health databases committed to generating reliable evidence with a central coordinating center housed at Columbia University.

# TABLE OF CONTENTS

1. TABLE OF CONTENTS 2

2. LIST OF ABBREVIATIONS 3

3. RESPONSIBLE PARTIES 3

3.1. Investigator(s) and Authors 3

3.2. Sponsor 3

4. PROTOCOL ABSTRACT 3

5. RATIONALE AND BACKGROUND 4

5.1. A framework describing desirable attributes of reliable evidence 4

5.2. A target study for replication: glucagon-like peptide 1 receptor agonists and chronic lower respiratory disease 6

6. STUDY Objectives 7

7. Research METHODs 7

7.1. Data sources 7

7.2. Study design and setting 9

7.2.1. Proposed study designs to examine evidence reliability 9

7.2.2. Target population and comparators 11

7.2.3. Time-at-risk 12

7.2.4. Outcome(s) of interest 12

7.2.5. Negative control outcome experiments 12

7.3. Additional variables of interest (demographic characteristics, confounders) 12

8. DATA ANALYSIS PLAN AND MODEL SPECIFICATION 12

8.1. Descriptive statistics 12

8.2. Propensity score adjustment 13

8.3. Evidence evaluation and synthesis 13

8.4. Outcome models and corresponding effect estimates 13

8.5. Calibration of effect estimates using the empirical null distribution 14

9. SAMPLE SIZE AND STUDY POWER 14

10. STRENGTHS AND LIMITATIONS OF THE RESEARCH METHODS 14

11. PROTECTION OF HUMAN SUBJECTS 15

12. MANAGEMENT AND REPORTING OF ADVERSE EVENTS AND ADVERSE REACTIONS 15

13. PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS 15

14. References 16

15. ApPEndix 1: list of candidate negative control outcomes 19

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| ATC | Anatomical Therapeutic Chemical Classification System |
| CC | Comparative cohort |
| CDM | Common data model |
| CEM | Common Evidence Model |
| DOACs | Direct oral anticoagulants |
| EHR | Electronic health record |
| HR | Hazard ratio |
| IPCW | Inverse probability of censoring weighting |
| IRB | Institutional Review Board |
| IRR | Incidence rate ratio |
| OHDSI | Observational Health Data Sciences and Informatics |
| RD | risk difference |
| sIPTW | Stabilized inverse-probability of treatment weighting |

# RESPONSIBLE PARTIES

## Investigator(s) and Authors

## Sponsor

Observational Health Data Sciences and Informatics (OHDSI) Center at the Roux Institute\*

# PROTOCOL ABSTRACT

A high-quality observational comparative cohort study conducted using a nation-wide sample of U.S. administrative claims was recently published in JAMA demonstrating a significantly increased risk of major ischemic or hemorrhagic events associated with treatment with rivaroxaban compared with apixaban.

We propose to independently replicate the original study according to its description in the recent publication and supplemental materials, using a similar database as was used in the original study. In addition to replicating the study, we plan to evaluate the robustness of the study findings by conducting sensitivity analyses, assessing: 1) changes to definitions of exposure and outcome phenotypes, 2) calibration of effect estimates using empirical null distributions, and 3) observable study diagnostics that inform the validity of a given analysis. Finally, we plan to explore the generalizability of the findings by executing the analysis on several study databases that vary with respect to the populations they include (e.g. U.S. and non-U.S.) and their mechanisms of data capture (e.g. administrative claims, electronic health records data).

We propose to conduct this study within the Observational Health Data Sciences and Informatics (OHDSI) research network. The OHDSI research network is an international collaboration of researchers seeking to advance observational data research in healthcare. OHDSI is an open network, inviting healthcare institutions across the globe with patient-level data to join the network by converting data to the OMOP CDM and participating in network research studies. Each site voluntarily opts-in to each respective network study. In each study, data remains at the site behind a firewall. No patient-level data pooling occurs across network sites; only aggregate results are shared.

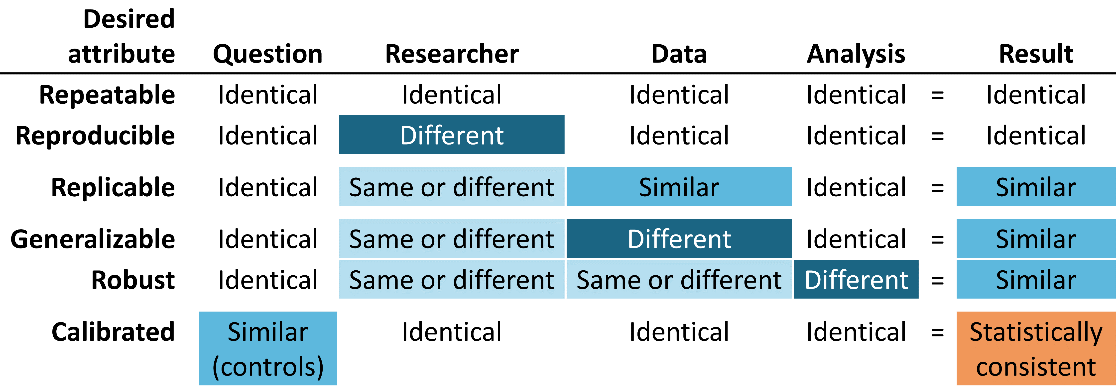
# RATIONALE AND BACKGROUND

Observational research using real-world data provides an opportunity to observe how medications are used in a real-world setting and to assess the risks and benefits associated with medications in actual practice. Ongoing development of study designs and analytic methods have equipped observational researchers with a powerful toolkit to reduce and/or mitigate sources of bias. For example, developments in propensity score analysis have enabled control for large, highly-dimensional sets of confounders and developments in negative control outcome experiments allow for the assessment of residual systematic bias, which can be used to calibrate effect estimates and confidence intervals. When proper study designs and analytic methods are used to address the right study questions, observational research using real-world data can provide valuable and reliable evidence on the effectiveness and safety of medical products, as they are used in real-world clinical practice.(1)

Despite substantial improvements in methods and data resources used in observational causal inference research, a large number of published observational studies still conclude with the statement “Randomized trials are warranted to confirm our findings,” even in cases where the recommended randomized trial is highly unlikely to be conducted in a timely manner (e.g. due to challenges with feasibility and/or cost). Implicit in this trend is the assumption that observational research cannot provide reliable evidence absent confirmation with experimental/randomized research. Given that many of-interest health research questions simply cannot be realistically assessed using experimental methods, it is critical that systems of evidence evaluation are developed to apply the findings of observational research studies more meaningfully. Despite commonplace concerns about the reliability of evidence generated by observational research, established practices to improve the reliability of empirical evidence (e.g. reproduction and replication) are severely underutilized, an issue which has received widespread attention over the last decade.(2-8)

## A framework describing desirable attributes of reliable evidence

Figure 1 below presents a general framework for evaluating six desirable attributes of reliable evidence: repeatable, reproducible, replicable, generalizable, robust, and calibrated. A brief description of the framework is provided below; however more detailed discussion of the framework can be found in [Section 14.1 of the Book of OHDSI (Attributes of Reliable Evidence)](https://ohdsi.github.io/TheBookOfOhdsi/EvidenceQuality.html).(9)



**Figure 1**. Desired attributes of reliable evidence (source: Book of OHDSI)

Evidence is considered repeatable when the same researcher applying the same analytic strategy to address the same study question with the same data would get the identical result. Generating repeatable evidence requires a clear process and absent of subjective judgement or interpretation. Evidence is considered reproducible when a different researcher can perfectly reproduce that analysis and, using the same data, is able to generate the identical result. In the case of research relying on retrospective data analysis, reproducible evidence requires the complete specification of the study protocols in human-readable and computer-executable form. Ensuring that findings from a given study are repeatable and reproducible is one of the most important principals for providing reliable evidence of causal relationships.

Evidence replicability, which relates to the Bradford-Hill consistency criterion(10), implies that the identical analysis applied to a similar but different data resource / population, generates a similar finding. In other words, has the association been observed in multiple contexts (e.g. across multiple different insurance claims databases)? Generalizable evidence represents findings that can be reproduced across multiple different sources of data (e.g. administrative claims databases, large electronic health records [EHR] databases, single-facility EHR systems) which vary with respect to the populations they include and their mechanisms of data capture.

Expanding on repeatability, reproducibility, replicability, and generalizability, evidence is further strengthened when it can be shown to be robust to variation in analytic choices. Sensitivity analyses commonly included in observational research studies typically seek to explore the robustness of findings to variation in potentially important analytic choices (e.g. definition of exposures and outcomes). However, observational evidence should be considered even more robust if it can be consistently observed using broad range of study designs which have orthogonal analytic assumptions and complementary strengths and limitations. For example, a comprehensive evaluation of robustness for a comparative cohort study may also include an assessment of whether similar study findings can be consistently generated using alternate study self-controlled designs.

Lastly, reliable evidence must be calibrated, meaning the findings must be inspected for residual systematic error. Statistical artifacts should be able to be empirically demonstrated to have well-defined properties, such as a 95% confidence interval having 95% coverage probability (which can be discerned from negative control outcome experiments). Negative controls have been shown to be a powerful tool for identifying and mitigating systematic error in observational studies.(11-14)

## Candidate study for replication: direct oral anticoagulants and bleeding and hemorrhagic events

Direct oral anticoagulants (DOACs) are the preferred anticoagulants for patients with atrial fibrillation. The beneficial and adverse effects of DOACs, however, are closely associated with their plasma concentrations. The difference in dosing schedules of rivaroxaban and apixaban can lead to different plasma concentration and hence different safety and efficacy profile. In their recent publication, Ray et al compared major ischemic and hemorrhagic outcomes in Medicare beneficiaries with atrial fibrillation who initiated anticoagulation treatment with rivaroxaban or apixaban.(15) The primary outcome was a composite of major ischemic and hemorrhagic events, an integrated measure of benefits and harms of anticoagulation for patients with atrial fibrillation. The results suggested significantly increased risk of the primary outcome for rivaroxaban compared with apixaban (RD, 2.7 (95%CI, 1.9-3.5) per 1000 person-years; HR, 1.18 (95% CI, 1.12-1.24)).

The attached PDF files include the full study manuscript and its supplemental materials, which represents the study protocol that our study designs, analyses and evaluation will be based around. A brief summary of their study approach is included below.

The study was conducted using the US Medicare data and included Medicare beneficiaries with atrial fibrillation initiating standard dose or reduced dose of rivaroxaban or apixaban between January 1, 2013 and November 30, 2018 were identified. The author conducted a retrospective new user cohort study and compared new users of rivaroxaban (standard or reduced dose) with new users of apixaban (standard or reduced dose) to assess 1) the risk of hemorrhagic or ischemic events, 2) nonfatal extracranial bleeding and 3) total mortality. Patients had to have a diagnosis of atrial fibrillation/flutter within 90 days prior to initiation of rivaroxaban or apixaban. To ensure regular contact with medical care, patients had to have continuous enrollment during the preceding year with at least one outpatient visit and one filled

prescription (other than the study anticoagulant).

The study controlled for 208 covariates associated with anticoagulation choice and outcomes using stabilized inverse probability treatment weighting (sIPTW). The propensity score distributions in the rivaroxaban and apixaban groups had good overlap, with all standardized

differences no more than 0.01. After re-weighting, the authors estimated adjusted hazard ratios (HRs) and adjusted risk difference (RD) using Cox proportional hazards and Poisson regression, respectively. The author accounted for the weighting induce dependencies using with modified sandwich variance estimation for the proportional Cox models and variances for Poisson models were calculated via generalized estimating equations. HRs and RDs were calculated for patients according to dose. The authors performed a series of sensitivity analysis to assess the effects of alternate definitions, analytic choices, the maximum allowed gap in anticoagulant days of supply or follow up time. The also assessed the potential influence of unmeasured confounding by calculating the E-value.

Generally speaking, this study represents a high-quality analysis relative to the majority of retrospective observational causal inference research published using secondary data. The analysis seeks to address threats to validity and evaluate robustness with respect to several analytical assumptions, within the context of a comparative cohort design. However, additional work is needed to further examine the study’s reliability, specifically: reproducibility, replicability, generalizability, robustness, and calibration.

# STUDY Objectives

**Objective 1.** Attempt to independently replicate the Ray et al. study as closely as possible using the IQVIA OpenClaims database, according to its description in the recent publication and supplemental materials. (15)

**Objective 2**. Evaluate the robustness of the Ray et al. study findings by conducting sensitivity analyses in the IQVIA OpenClaims database, assessing:

* Sensitivity to changes in definitions of exposure and outcome phenotypes,
* Sensitivity to empirical calibration(11-14) of effect estimates using negative and positive control outcomes, and
* Using observational study diagnostics to assess the robustness of their findings and risk of systematic bias before and after calibration.

**Objective 3**: Evaluate the generalizability of findings produced by the original Ray et al. protocol across different study databases, including U.S. and non-U.S. populations.

# Research METHODs

## Overview

The general approach for this study is to sequentially build a body of evidence which seeks to examine the reliability of the original Ray et al. finding. As the first building-block of this reliability assessment, we plan to address Objective 1 by replicating the original analysis as closely as possible using OHDSI tools, within data formatted to the OHDSI CDM, based on direct inspection of the original study’s code and published manuscript. Assuming the study findings can be replicated in a similar database, we will proceed to Objective 2, where we evaluate whether the study findings are robust to plausible variations in study design choices, including alternate definitions of target/comparator cohorts, outcome phenotypes, and study design. To do so, we plan to conduct sensitivity analyses that vary one design choice/parameter at a time (as opposed to exploring the combination of several parameters). Specifically, we will also assess the robustness of the findings by looking into other age groups (below 65 years) and different time at risk windows (bellow and above 6 months). Furthermore, we will assess robustness by assessing whether study findings are consistent after empirical calibration of effect estimates using negative and positive control outcomes. We will also use standard OHDSI study diagnostics to assess whether the data and design are fit-for-purpose and can be expected to generate inferences with minimal systematic error. Finally, to address Objective 3, we will assess the generalizability of findings by deploying the original Ray et al. protocol across different databases in the OHDSI Research Network (including U.S. and non-U.S. populations). Doing so allows us to evaluate whether the findings are consistent when applied to data which include different study populations and exhibit different mechanisms of data capture.

Uncertainty regarding the reliability of findings generated by observational studies limits the impact of observational research findings on real-world clinical practice. Collectively, we believe the body of evidence generated by these analyses will provide a comprehensive and holistic assessment of evidence reliability, which could either strengthen confidence in the original study finding or identify important gaps in the evidence which can be addressed by further research using alternate study designs and data resources. We believe that either result will improve the field’s understanding of this evidence and, if appropriate, enable changes in patient-care/treatment decision-making.

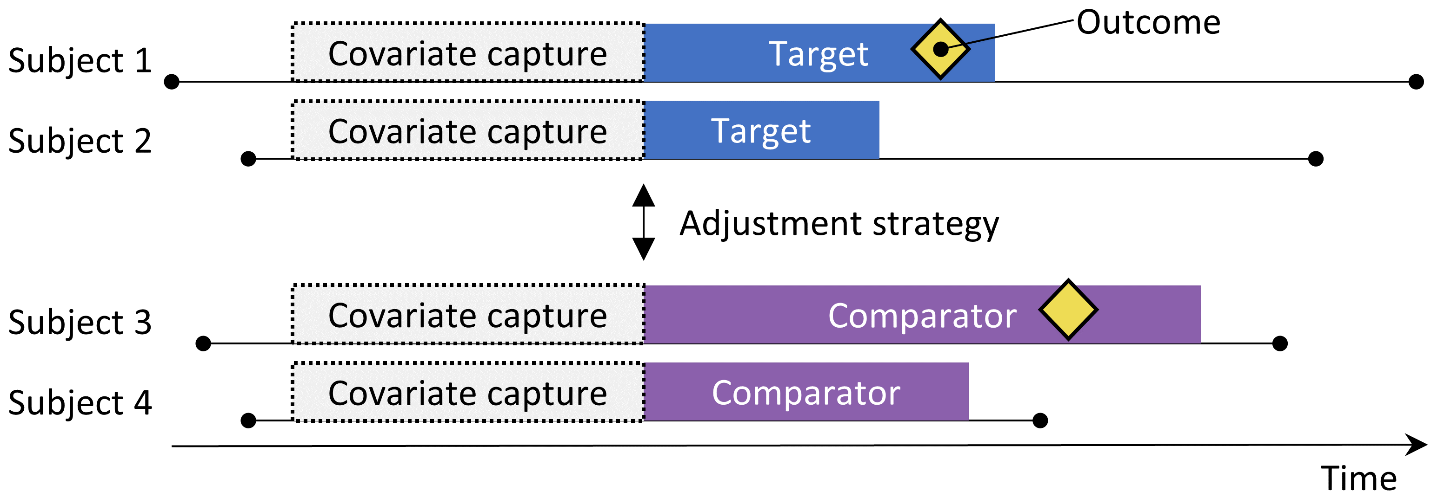
## Study design and setting

### Proposed study designs to examine evidence reliability

There is only one study design will be used as part of this evaluation: a new user cohort comparison (CC). We will not attempt to examine other study designs as we believe neither self-controlled cohort, and a self-controlled case series are appropriate for the current question.

New-user comparative-cohort (CC) design

In the CC design (Figure 2, below), subjects observed to initiate the target treatment are compared to those initiating the comparator treatment. To adjust for differences between the two treatment groups, several adjustment strategies are available, which include stratification, matching, or weighting by the propensity score, or by adding baseline characteristics to the outcome model. To delineate clear temporality between exposure, outcome and baseline confounders, the characteristics included in the propensity model or outcome model only include those captured prior to treatment initiation. The CC design can be employed using several different outcome models, allowing for the estimation of various effect estimates including logistic regression to estimate the risk ratio, Cox proportional hazards models (time-to-event analysis) to estimate the hazard ratio, and Poisson regression to estimate the incidence rate ratio. Additional description of the CC design and related analytic approaches is provided in [Section 12.1 of the Book of OHDSI](https://ohdsi.github.io/TheBookOfOhdsi/PopulationLevelEstimation.html#CohortMethod).



**Figure 2.** Schematic illustrating the new-user cohort comparison (CC) design

We will use the new user CC design in order to as closely as possible replicate the analysis described in the recent publication by Ray et al in a similar data source.

In addition to evaluating replicability, and generalizability, we also plan to evaluate the robustness of the findings to alternate choices in analytic choices, including 1) alternate definitions of exposure and/or outcome phenotypes, sensitivity to empirical calibration of effect estimates using negative and positive control outcomes. As a further perspective on the robustness of study findings, we plan to employ observational study diagnostics in order to inform our understanding of whether findings are systematically biased. Additional descriptions of proposed diagnostic procedures and empirical calibration methods are provided in Section 7.4 and Section 8.5 of this protocol, respectively.Furthermore, we plan to explore whether the study question addressed by Ray et al. generates findings that are robust to alternate study designs choices, which are described below.

### Target population and comparators

In this study, we will seek to assess patients initiating standard or reduce dose of rivaroxaban for atrial fibrillation/flutter. In analyses employing the CC design, we explicitly plan to replicate the comparator group from Ray et al which includes initiating standard or reduce dose of apixaban for atrial fibrillation. In order to examine the robustness of the findings from the CC design analysis, we will explore the association in other age groups (below 65 years and all ages) and different time at risk windows (bellow and above 6 months). Beyond the original definition of target populations in the Ray et al., we will conduct sensitivity analyses that vary the target and comparator populations of interest. Medication information is available at the level of the RxNorm ingredient.

### Time-at-risk

Time exposed will be considered the time-on-treatment, in accordance the original Ray et al. protocol, which will be defined using the days supply data that accompany prescribing events. If an otherwise feasible data set is available that does not contain reliable information on days supply, on-treatment time may be inferred using product labels to inform dosing schedules. tables will be created using products labels to infer dosing schedules. Beyond the original definition of time-at-risk described in the Ray et al. protocol, we will explore sensitivity analyses that vary the time-at-risk definition for all three study designs.

### Outcome(s) of interest

We plan to replicate the primary and secondary outcomes as defined by Ray et al. Clinical input from experts in the relevant disease areas (cardiovascular disease) will be included during the process of identifying alternate outcome phenotypes. Alternate outcome phenotypes will be developed and evaluated using OHDSI’s CohortDiagnostic tool which provides a suite of analyses including syntax-/vocabulary-based concept-set diagnostics, cohort incidence rates over time, gender, and age, index event breakdowns, cohort overlap statistics, baseline and temporal characterizations, cohort comparison, and patient profiles. These tools allow for a data-driven, substantive inspection of phenotypes which, when interfaced with clinical knowledge, can ably identify problems in cohort definitions including the populations they do and do not capture and the timing of the events they identify. The CohortDiagnostic tool can be used to instantiate and evaluate cohorts in any dataset formatted to the OMOP CDM. This will enable us to understand outcome performance in all study databases.

### Negative control outcome experiments

Negative controls will be used to evaluate the potential impact of residual systematic error in the study designs, and to perform the empirical calibration of confidence intervals around effect estimates.(11-14) An outcome is considered a good candidate for a negative control if it is believed to have no causal relationship with the outcome or exposure (i.e. that a hypothetical, well-designed randomized trial would find no relationship between exposure and the negative control outcome). For the purposes of our evaluation of robustness, we plan to conduct an evaluation of a broad set of negative controls. Because we believe that confounding structures are unknowable, it is our view that using a large number of negative controls provides a more meaningful assessment of uncertainty than assessing a single negative control analysis, since it provides a distribution of potential systematic error rather than a single number. The set of negative control outcomes we propose using for this assessing is provided in Table xxx in Appendix 2. Additional description of empirical calibration methods is provided in Section 8.5 of this document.

### Positive control outcome experiments

To understand the behavior of a method when the true relative risk is smaller or greater than one requires the use of positive controls where the null is believed not to be true. In OHDSI we therefore use synthetic positive controls, (Schuemie, Hripcsak, et al. 2018) created by modifying a negative control through injection of additional, simulated occurrences of the outcome during the time at risk of the exposure. For example, assume that, during exposure to Rivaroxaban, n occurrences of our negative control outcome “ingrown nail” were observed. If we now add an additional n simulated occurrences during exposure, we have doubled the risk. Since this was a negative control, the relative risk compared to the counterfactual was one, but after injection, it becomes two.

To preserve confounding, we want the new outcomes to show similar associations with baseline subject-specific covariates as the original outcomes. To achieve this, for each outcome we train a model to predict the survival rate with respect to the outcome during exposure using covariates captured prior to exposure. These covariates include demographics, as well as all recorded diagnoses, drug exposures, measurements, and medical procedures. An L1-regularized Poisson regression (Suchard et al. 2013) using 10-fold cross-validation to select the regularization hyperparameter fits the prediction model. We then use the predicted rates to sample simulated outcomes during exposure to increase the true effect size to the desired magnitude. The resulting positive control thus contains both real and simulated outcomes.

Our process for creating positive controls synthesizes outcomes with a constant incidence rate ratio over time and between patients, using a model conditioned on the patient where this ratio is held constant, up to the point where the marginal effect is achieved. The true effect size is thus guaranteed to hold as the marginal incidence rate ratio in the treated. Under the assumption that our outcome model used during synthesis is correct, this also holds for the conditional effect size and the ATE. Since all outcomes are rare, odds ratios are all but identical to the relative risk. Additional description of positive control methods is provided in [Section 18.2.2 of the Book of OHDSI](https://ohdsi.github.io/TheBookOfOhdsi/MethodValidity.html#PositiveControls).

### Additional variables of interest (demographic characteristics, confounders)

For all designs, we will describe patient characteristics (prevalence) for each cohort and data source including but not limited to: gender, age, index month, conditions (SNOMED concepts and descendants) drugs (ATC classes and RxNorm ingredients), procedures, devices, measurements, the CHADS2Vasc score, and the Charlson Comorbidity Index.

## Data sources

This study will be conducted within the OHDSI research network. The OHDSI community has authored standards and tools to facilitate high-quality, reproducible research, including OMOP Standardized Vocabularies, the Common Data Model (CDM) and analytical methods packages. OHDSI network studies represent the culmination of a transparent, consistent and reproducible way to conduct research across a large number of geographically dispersed data. The OHDSI Research network is an international collaboration of researchers seeking to advance observational data research in healthcare. OHDSI is an open network, inviting healthcare institutions across the globe with patient-level data to join the network by converting data to the OMOP CDM and participating in network research studies. Each site voluntarily opts-in to each respective network study. In each study, data remains at the site behind a firewall. No patient-level data pooling occurs across network sites. Only aggregate results are shared.

The results of an observational study can be influenced by many factors that vary by the location of the data source such as patterns of treatment discontinuation/switching, genetic diversity, or environmental factors, overall health status: factors that may not have been possible to vary in the context of a clinical trial even if one exists for your same study question. A typical motivation to run an observational study in a network is therefore to increase diversity with respect to data sources, study populations, and patterns of clinical practice to inform how well the results generalize to the many types of people they will eventually be applied to. In other words, can the study findings be replicated across multiple sites, or do they differ and if they differ, can any insights be gleaned as to why?

OHDSI network data sources will be used as they are made available by data partners in the OHDSI research network. We will conduct feasibility assessments with volunteered data sources to determine if they have adequate sample and sufficient data to meaningfully inform the study (e.g. through calculation of the minimally-detectable risk ratio). All data sources will receive institutional review board approval or exemption for their participation before executing the study. Specifically, the IQVIA OpenClaims database will be used to evaluate replicability. Additional study resources will be used to evaluate the generalizability of the Ray et al. findings.

Table 1 (below) describes data sources that are currently available to be used in this study, pending assessment of feasibility. These sources encompass a variety of practice types, data capture, and populations. The embedded hyperlinks can be used to access more detailed descriptions of each data source. For each, we report a brief description and size of the population it represents as well as the calendar years and data that it captures.

**Table 1.** Already-available data sources that can be used to replicate/generalize findings from Ray et al.

| **Data source** | **Population** | **Patients** | **History** | **Data capture and short description** |
| --- | --- | --- | --- | --- |

## Fit-for-purpose evaluation

In order to assess the diagnostics that inform the robustness of estimates generated by a single analysis within a single data source, the following additional evaluations will be performed: power calculations estimating minimum detectable relative risk; and negative-control calibration plots to assess residual systematic bias. For the CC design, we will also calculate the following: preference score (a transformation of propensity score that adjusts for prevalence differences between populations) distributions to evaluate empirical equipoise and population generalizability; patient characteristics to evaluate cohort balance before and after propensity score adjustment, and Kaplan-Meier plots to examine proportionality assumptions in the analyses estimating the hazard ratio. In cases where a specific analysis produces study diagnostics that indicate potentially biased findings, we will note that when sharing the results of that analysis. Further description and justification for this approach is provided in Section 8.3 of this protocol. In Section 7.2.4 of this protocol, we also describe tools that are used to build and validate target/comparator/outcome phenotypes within study data sources.

# DATA ANALYSIS PLAN AND MODEL SPECIFICATION

## Descriptive statistics

In descriptive analyses, we will summarize patient characteristics at the point of entry into the study cohort, using proportions to describe dichotomous covariates and using the mean and standard deviation to describe continuous covariates (e.g. the Charlson Comorbidity Index). For the comparative cohort design, we will describe characteristics for each treatment group and will calculate the absolute standardized mean difference between the two.(16)

## Propensity score adjustment

To adjust for between-person confounding when using the CC design, propensity score models for each class pair and data source will be created using a data-driven process using regularized logistic regression. This process allows the data to inform which combinations of baseline patient characteristics, including demographics and previous conditions, drug exposures, procedures, and health-service-use behaviors are most predictive of treatment assignment. The propensity score model used in the comparative cohort design will include the following potential confounding variables: age, gender, the year and month of the exposure, presence of medical conditions, medications, medical procedures, and the number of distinct conditions, procedures, measurements (e.g., lab test results), and visit counts. For the initial objective, where we attempt to replicate the finding using the same approach used by Ray et al., we plan to assess baseline covariates within a 365 day look-back window.

The original study by Ray et al. used sIPTW methods to balance the study populations being compared. We may also choose to conduct sensitivity analyses exploring the use of propensity score matching and/or propensity-score stratification. Because sIPTW can produce extreme weights and can be difficult to deploy in multiple data settings without hands-on inspections of weights, we may choose to use matching or stratification when exploring the generalizability of findings across databases. It is important to note that these treatments appear to be highly comparable so we do not anticipate that the chosen method of confounding control will meaningfully impact findings.

## Evidence evaluation and synthesis

In Section 7.4 of this protocol, we describe the diagnostics that will be used to assess fit-for-purpose evaluations for each analysis within each database. Since the explicit intention of this study is to explore the reliability of an already-published analysis, we plan to present effect estimates from all analyses, regardless of whether they pass diagnostics. Doing so provides needed information about the scale/magnitude of various sources of bias that may inform the reliability of the original finding by Ray et al. In cases where failing diagnostics indicate potentially spurious effect estimates, we will specifically highlight that when presenting results. If appropriate, effect estimates generated by analyses that do satisfy diagnostic inspections will be aggregated across data sources using a random-effects model to produce meta-analytic estimates. Before combining into a meta-analytic estimate, we will inspect effect estimates for cross-database heterogeneity, as indicated by the I2 score, which quantifies the proportion of total variation across studies that is due to heterogeneity rather than chance.(17)

## Outcome models and corresponding effect estimates

The CC design can be employed using several different outcome models, allowing for the estimation of various effect estimates including: logistic regression to estimate risk ratio, Cox proportional hazards models (time-to-event analysis) to estimate the hazard ratio, and Poisson regression to estimate the incidence rate ratio.

## Calibration of effect estimates using the empirical null distribution

The distribution of effect estimates across all negative control outcome experiments will be used to fit an empirical null distribution, which collectively reflects both random and residual systematic error. The empirical null distribution is particularly well equipped to demonstrate systematic sources of selection bias and confounding bias and, to some degree, misclassification bias. The empirical null distribution will then be applied to the outcome of interest to calibrate the p-value and confidence interval. Empirical calibration serves as an important diagnostic tool to evaluate if the residual systematic error is sufficient to cast doubt on the accuracy of the unknown effect estimate.(11-14) The calibration effect plot will be generated and included alongside results in order to inform interpretation of the results. We will report the traditional and empirically calibrated effect estimates and confidence interval for each analysis.

# SAMPLE SIZE AND STUDY POWER

This is a retrospective cohort study all available observations will be included, and thus no a priori sample size or power calculations will be performed. For each distinct analysis, we will estimate the minimally detectable relative risk, which provides an estimate of the point at which the true effect size would be too small to detect using a 95% confidence interval.

# STRENGTHS AND LIMITATIONS OF THE RESEARCH METHODS

There are several important strengths, limitations, and assumptions that inform the use of the CC design. Perhaps most importantly, the CC design emulates a randomized clinical trial and, as such, is desirable with respect to the ease of interpretability and also its clear delineation of the temporality of confounder-exposure-outcome relationships. However, in order to clearly delineate temporality, the CC design requires relative accuracy ascertaining exposure initiation status and timing. Furthermore, the CC design is susceptible to unmeasured confounders (including time-variant and time invariant confounders), specifically but not limited to confounding by indication. Finally, the CC design can be biased if loss-to-follow-up is related to both exposure and disease status.

From its inception, the declared goal of the OHDSI community has been to establish an international collaborative by building on open-science values.(9) This work intending to interrogate the reliability of the Ray et al. study directly aligns with OHDSI’s open-science strategy and its commitment to enable more meaningful collaboration in the health research community. OHDSI has developed a broad toolkit to make such work possible, including the use of open-source software, public availability of all conference proceedings and materials, and transparent, open-access publication of generated medical evidence. These developments enable the OHDSI community to not just conduct high-impact observational health research, but to further our understanding of the reliability of existing health research. In keeping with this philosophy, the work conducted on this project and the generated results will also be shared in such a way (i.e. including open-source, executable code) that enables other researchers, both within and outside the OHDSI community, to re-run analyses and understand the reliability of our findings. It is our hope that this work represents an early step toward a more robust system of evaluating evidence quality.

# PROTECTION OF HUMAN SUBJECTS

Confidentiality of patient records will be maintained always. All study reports will contain aggregate data only and will not identify individual patients or physicians. At no time during the study will the sponsor receive patient identifying information except when it is required by regulations in case of reporting adverse events. We will seek approval from all data sources in accordance to our data use agreements as applicable.

# MANAGEMENT AND REPORTING OF ADVERSE EVENTS AND ADVERSE REACTIONS

This study uses coded data that already exist in electronic databases. In this type of database, it is not possible to link (i.e., identify a potential causal association between) a particular product and medical event for any 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 (AEs) are not reportable as individual AE reports. The study results will be assessed for medically important results.

# PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

In accordance with OHDSI’s open-science approach, all aspects of this research will be publicly shared. From an open-science perspective, one of the most important resources to share is the code for the actual execution of studies, such as studies from the OHDSI Research Network. In turn, these programs leverage the fully open source OHDSI stack, which can be inspected, reviewed and contributed to via GitHub. Furthermore, by providing executable code, we enable other researchers to reproduce the analyses we conduct as part of this evaluation of evidence reliability. Because of the privacy-sensitive nature of healthcare data, fully open, comprehensive patient-level datasets are typically not available. However, it is possible to leverage OMOP mapped datasets to publish important aggregated data and results sets to <http://data.ohdsi.org>.

# References

1. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS medicine. 2011;8(5):e1001026.

2. Challenges in irreproducible research (special issue). Nature. 2018;7707(May;557).

3. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circulation research. 2015;116(1):116-26.

4. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? Journal of the American Medical Informatics Association : JAMIA. 2018;25(8):963-8.

5. Harris JK, Johnson KJ, Carothers BJ, Combs TB, Luke DA, Wang X. Use of reproducible research practices in public health: A survey of public health analysts. PloS one. 2018;13(9):e0202447.

6. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. Jama. 2005;294(2):218-28.

7. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery. 2011;10(9):712.

8. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ digital medicine. 2019;2:2.

9. Observational Health Data Sciences and Informatics, Abedtash A, Blacketer B, Cloft C, Duke D, Falconer F, et al. The Book of OHDSI. 1st ed2020 April 16.

10. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58(5):295-300.

11. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of p-values using observational data. Statistics in medicine. 2016;35(22):3883-8.

12. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(11):2571-7.

13. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Statistics in medicine. 2014;33(2):209-18.

14. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philosophical transactions Series A, Mathematical, physical, and engineering sciences. 2018;376(2128).

15. Ray WA, Chung CP, Stein CM, Smalley W, Zimmerman E, Dupont WD, et al. Association of Rivaroxaban vs Apixaban With Major Ischemic or Hemorrhagic Events in Patients With Atrial Fibrillation. Jama. 2021;326(23):2395-404.

16. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in medicine. 2009;28(25):3083-107.

17. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ (Clinical research ed). 2003;327(7414):557-60.

# Appendix 1: Description of considered study designs

**Table S1. Overview of selected study designs including strengths and limitations of selected study design**

| **Study Design** |  | **Population** | **Example of Data Being Compared** | **Strengths** | **Weaknesses** | **Examples of Test Statistics or Regression Models** |
| --- | --- | --- | --- | --- | --- | --- |
| New user comparative cohort (CC) design |  | Exposed persons Persons exposed to an active comparator (often matched on potential confounders) | Incidence or incidence rates of those vaccinated vs. incidence or incidence rates of those unvaccinated | Standard design Easily implemented large observational data sources Propensity score methods control for potential confounders Accounts for seasonality; index date aligns risk period for target-exposed and comparator-exposed | Confounding by indication, other unmeasured confounders Susceptible to misclassification of exposure Unvaccinated control population may be limited Biased incidence rate ratio if loss to follow-up is affected by both exposure and disease status | Linear and logistic regression Cox regression Poisson regression (conditional logistic regression or Poisson regression if matched analysis) |
| Self-controlled case series (SCCS) design |  | Cases | Incidence rates of exposed time periods vs incidence rates of self-matched unexposed time periods | Self-controlled, adjusts for time-invariant confounders Multiple occurrences of independent events within an individual can be assessed | Time-varying confounding Reverse causality bias | Conditional Poisson regression |
| Self-controlled cohort (SCC) design |  | Exposed persons, but only cases informative | Incidence rates of exposed time periods vs. incidence rates of self-matched unexposed time periods | Self-controlled, adjusts for time-invariant confounders Less susceptible to misclassification of exposure | Time-varying confounding (less susceptible than SCCS because of less variation in periods under observation) Possible bias due to differences in baseline risk of outcome between time-periods. | Conditional Poisson regression |

\* Table contents partially informed by: Baker MA, Lieu TA, Li L, Hua W, Qiang Y, Kawai AT, et al. A vaccine study design selection framework for the postlicensure rapid immunization safety monitoring program. American journal of epidemiology. 2015;181(8):608-18.

# ApPEndix 2: list of candidate negative control outcomes

**Table S2.** Candidate negative control outcomes (N=51) to use in comparisons of antidiabetes treatments

|  | **Concept ID** | **Concept Code** | **Concept Name** |
| --- | --- | --- | --- |
| 1 | 4056591 | 212454006 | Traumatic rupture of lumbar intervertebral disc |
| 2 | 45757682 | 323351000119108 | Anomaly of jaw size |
| 3 | 4058388 | 19843006 | Hypertrophic scar |
| 4 | 439907 | 268002004 | Non-traumatic tendon rupture |
| 5 | 40481897 | 443483005 | Inadequate sleep hygiene |
| 6 | 374919 | 24700007 | Multiple sclerosis |
| 7 | 4090205 | 187251001 | Sequelae of tuberculosis |
| 8 | 438021 | 23123006 | Disorder due to and following fracture of upper limb |
| 9 | 4166231 | 47708004 | Genetic predisposition |
| 10 | 440933 | 77968000 | Effects of accidental exposure to external cause |
| 11 | 374013 | 230292008 | Secondary parkinsonism |
| 12 | 141667 | 27911000 | Laxity of ligament |
| 13 | 4216670 | 81302005 | Worried well |
| 14 | 77139 | 2630008 | Open wound of finger without complication |
| 15 | 132736 | 5758002 | Bacteremia |
| 16 | 374358 | 246815009 | Excess skin of eyelid |
| 17 | 4344500 | 239960007 | Impingement syndrome of shoulder region |
| 18 | 4128329 | 289903006 | Menopause present |
| 19 | 439438 | 162213003 | Crying, excessive |
| 20 | 4168681 | 274646000 | Irritability and anger |
| 21 | 4051630 | 23268009 | Malingering |
| 22 | 4310082 | 216551003 | Accidental poisoning by cannabis derivatives |
| 23 | 440858 | 32106001 | Cortex contusion without open intracranial wound |
| 24 | 439216 | 216545008 | Accidental poisoning by antidepressants |
| 25 | 440612 | 212899006 | Toxic effect of tobacco and nicotine |
| 26 | 373764 | 28394000 | Toxic encephalopathy |
| 27 | 432798 | 5285008 | Mechanical complication of internal orthopedic device, implant AND/OR graft |
| 28 | 374801 | 75441006 | Foreign body in ear |
| 29 | 440905 | 44400004 | Toxic effect of venom |
| 30 | 440053 | 71571008 | Infestation by insect |
| 31 | 441487 | 370977006 | Frostbite |
| 32 | 30234 | 209557005 | Neck sprain |
| 33 | 4019836 | 105412007 | Social exclusion |
| 34 | 381877 | 56713002 | Dysfunction of eustachian tube |
| 35 | 433577 | 122481008 | Hammer toe |
| 36 | 373478 | 41256004 | Presbyopia |
| 37 | 381295 | 39450006 | Senile cataract |
| 38 | 137016 | 66569006 | Retained dental root |
| 39 | 4213540 | 417347005 | Cervical somatic dysfunction |
| 40 | 4067069 | 201040000 | Callosity |
| 41 | 4209423 | 56294008 | Nicotine dependence |
| 42 | 80242 | 307945003 | Current tear of medial cartilage AND/OR meniscus of knee |
| 43 | 138384 | 111566002 | Acquired hypothyroidism |
| 44 | 4079750 | 239873007 | Osteoarthritis of knee |
| 45 | 261880 | 46621007 | Atelectasis |
| 46 | 378427 | 46152009 | Tear film insufficiency |
| 47 | 374375 | 18070006 | Impacted cerumen |
| 48 | 40479573 | 441527004 | Stimulant abuse |
| 49 | 4170260 | 419145002 | Nightmares |
| 50 | 46286594 | 987891000000105 | Problem related to lifestyle |
| 51 | 443274 | 275471001 | Psychostimulant dependence |