Predicting randomized clinical trial results with real-world evidence: A case study in the comparative safety of tofacitinib, adalimumab and etanercept in patients with rheumatoid arthritis

**Version:** 2.1

**Authors:**

Runsheng Wang 1, Patrick Ryan 1,2, Hamed Abedtash 3, Eldar Allakhverdiiev 4, Deepa Balraj 5, Juan Banda 5, Maura Beaton 1, Paul Biondich 6, Clair Blacketer 2, Richard Boyce 25, Alison Callahan 5, Ray Chen 1, Young-Geun Choi 7, Ivan John Clement 8, Frank DeFalco 2, Sara Dempster 9, Stephen Deppen 10, 11, Jon Duke 12, Scott Duvall 11, Dmytro Dymshyts 4, Thomas Falconer 1, Pavel Grafkin 4, Shaun Grannis 6, Jill Hardin 2, Ross Hayden 6, George Hripcsak 1, Tommy Huynh 14, Yeesuk Kim 14, 15, Christopher Knoll 2, Martin Lavallee 16, Evan Minty 5,17, Akihiko Nishimura 15, Paul Petraro 19, Melanie Philofsky 4, Aaron Potvien 12, Christian Reich 8, Jenna Reps 2, Peter Rijnbeek 20, Patrick Ryan 1,2, Paola Saroufim 21, Lisa Schilling 19, Trey Schneider 12, Martijn Schuemie 2,15, Anthony Sena 2,20, Nigam Shah 5, Andrey Soares 19, David Sontag 22, Marc Suchard 15, Joel Swerdel 2, Mui Van Zandt 8, Rohit Vashisht 13, James Weaver 2, Chunhua Weng 1, Andrew Williams 23, Ross Williams 20, Jin Zhou 24, George Hripcsak 1, Kristin Kostka 8

1 Columbia University, 2 Janssen Research and Development, 3 Eli Lilly, 4 Odysseus Data Services, 5 Stanford University, 6 Regenstrief Institute, 7 Fred Huchinson Cancer Research Center, 8 IQVIA, 9 S. Dempster Consulting, 10 Vanderbilt University Medical Center, 11 US Veterans Affairs, 12 Georgia Tech Research Institute, 13 UCSF, 14 Hanyang University, 15 UCLA, 16 Virginia Commonwealth University, 17 University of Calgary, 18 Boehringer Ingelheim, 19 University of Colorado, Denver Anschutz Medical Campus, 20 Erasmus MC, 21 Case Western Reserve University, 22 MIT, 23 Tufts Medical Center, 24 University of Arizona, 25 University of Pittsburgh

**Date:** 2 May 2018

**Update:** June 2, 2020

**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.

**Disclosures:**

RW has received consulting fees from Novartis and Eli Lilly and Company (less than $10,000 each); PBR, CB, FD, JH, CK, JR, PR, MS, AS, and JW are employees of Janssen Research and Development. KK, MVZ, CR are employees are IQVIA. HA is an employee of Eli Lilly and Company. DD, PG and MP are employees of Odysseus Data Services. MS has received grants from the National Science Foundation, National Institutes of Health and IQVIA and consults for Janssen Research & Development and Private Health Management. GH has received grants from the US National Institutes of Health (R01 LM006910) and Janssen Research & Development. All authors will file ICMJE forms and provide updated disclosures at the time of manuscript submission.

# Table of contents

[1 Table of contents 2](#_Toc42154198)

[2 List of abbreviations 3](#_Toc42154199)

[3 Abstract 4](#_Toc42154200)

[4 Amendments and Updates 4](#_Toc42154201)

[5 Milestones 4](#_Toc42154202)

[6 Rationale and Background 4](#_Toc42154203)

[7 Study Objectives and Hypotheses 5](#_Toc42154204)

[8 Research methods 6](#_Toc42154205)

[8.1 Overview 6](#_Toc42154206)

[8.2 Data Source(s) 6](#_Toc42154207)

[8.3 Study population 7](#_Toc42154208)

[8.3.1 Inclusion/Exclusion criteria for real world analysis (RWD) 7](#_Toc42154209)

[8.3.2 Inclusion/Exclusion criteria to replicate clinical trial 7](#_Toc42154210)

[8.3.3 Executable Cohort Definitions 8](#_Toc42154211)

[8.4 Outcomes 8](#_Toc42154212)

[8.4.1 Primary outcome 8](#_Toc42154213)

[8.4.2 Secondary outcomes 8](#_Toc42154214)

[8.4.3 Negative control outcomes 8](#_Toc42154215)

[8.5 Covariates 9](#_Toc42154216)

[8.5.1 Propensity score covariates 9](#_Toc42154217)

[9 Data Analysis Plan 9](#_Toc42154218)

[9.1 Calculation of time-at risk 9](#_Toc42154219)

[9.2 Model Specification 10](#_Toc42154220)

[9.2.1 Pooling effect estimates across databases 10](#_Toc42154221)

[9.3 Analyses to perform 10](#_Toc42154222)

[9.4 Output 10](#_Toc42154223)

[9.5 Evidence Evaluation 11](#_Toc42154224)

[10 Study Diagnostics 12](#_Toc42154225)

[10.1 Sample Size and Study Power 12](#_Toc42154226)

[10.2 Cohort Comparability 12](#_Toc42154227)

[10.3 Systematic Error Assessment 12](#_Toc42154228)

[11 Strengths and Limitations of the Research Methods 12](#_Toc42154229)

[12 Protection of Human Subjects 12](#_Toc42154230)

[13 Management and Reporting of Adverse Events and Adverse Reactions 12](#_Toc42154231)

[14 Plans for Disseminating and Communicating Study Results 12](#_Toc42154232)

[15 Appendix 1 12](#_Toc42154233)

[15.1 Concepts excluded from covariates 12](#_Toc42154234)

[15.2 Negative control outcomes 12](#_Toc42154235)

[16 References 13](#_Toc42154236)

# List of abbreviations

ATC Anatomic Therapeutic Chemical

CYCLOPS Cyclic coordinate descent for logistic, Poisson and survival analysis

SNOMED Systematized Nomenclature of Medicine

OHDSI Observational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

T Target cohort

C Comparator cohort

O Outcome cohort

PS Propensity Scores

LASSO Least absolute shrinkage and selection operator

MDRR Maximum detectable relative risk

CI Confidence Interval

DOD Date of Death

MACE Major Adverse Cardiovascular Event

NMSC Non-melanoma skin cancer

ADA Adalimumab

ETN Etanercept

TOF Tofacitinib

RA Rheumatoid arthritis

# Abstract

This study aims to assess the cardiovascular safety of tofacitinib, in comparison to adalimumab and etanercept, in patients with rheumatoid arthritis, using real world evidence. In addition, we will replicate the design and population inclusion criteria of an ongoing phase 3b/4 randomized clinical trial (NCT02092467), with the aim of predicting the RCT results using real-world evidence. In this study, we will analyze data from observational databases across the OHDSI network using the OHDSI Population Level Estimation (PLE) package framework to perform this comparative safety study.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| 0.1 | 2 May 2018 | P. Ryan | First draft |
| 1.1 | 15 March 2019 | R. Wang | Second draft |
| 2.1 | June 2, 2020 | R. Wang | Third draft |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Milestones

|  |  |
| --- | --- |
| Milestone | Planned / Estimated Date |
| Start of analysis | 2 May 2018 |
| End of initial analysis | 3 May 2018 -> June 2020 |
| End of final analysis | 30 May 2019 -> August 2020 |
| Submission of manuscript | 31 July 2019 -> October 2020 |
|  |  |

# Rationale and Background

Rheumatoid arthritis (RA) is a chronic inflammatory condition that primarily affecting peripheral joints, causing joint damage and loss of function. Treatment options for patients with RA have significantly expanded in the past two decades, including conventional synthetic disease-modifying antirheumatic drugs (csDMARDs), biologic DMARDs (bDMARDs) and targeted synthetic DMARDs. When patients have inadequate response to csDMARDs, a second agent, either bDMARDs or tsDMARDs, is added, based on the expected efficacy and safety profile of the drug.

Tofacitinib, an oral Janus kinase inhibitor, was approved for use in patients with moderate-to-severe rheumatoid arthritis by the FDA in 2012. As part of the approval, FDA required a postmarketing study to examine the long-term effects of tofacitinib on cardiovascular events, malignancies, and serious infections; the Phase 3b/4 randomized clinical trial uses adalimumab and etanercept as active comparators and is scheduled to complete in 2019.

Regulatory approvals with post-marketing commitments for obtaining additional evidence are commonplace, as it is understood that the complete safety and efficacy product of a medicine cannot be fully understood during a product’s initial development. Additional prospective studies, such as randomized clinical trials and clinical registries, are time-consuming and resource-intensive, and may still not provide sufficient evidence to resolve all uncertainties about the effects of a medical product in clinical practice. Real-world evidence, including population-level effect estimation from observational databases about the safety and comparative effectiveness of medical products, provides a complementary perspective toward – and potentially offers the possibility to supplant in select circumstances – randomized clinical trials. It would be the shared interest of regulators, product manufacturers, health systems, clinicians and patients if the desired evidence to understand a product profile and its appropriate use in medical practice could be obtained more quickly and with less resources, so long as the evidence was of sufficient quality to make appropriate decisions.

Regulatory authorities seek the highest quality available evidence to inform their regulatory decision-making, including the initial approval of products, ongoing pharmacovigilance, and extension to supplemental indications. While observational evidence has been used in post-marketing safety evaluations, there is not currently consensus across the research community as to whether the reliability of real-world evidence has been sufficiently demonstrated to constitute “substantial evidence” when evaluating the effectiveness of new drugs or determining the need for further prospective studies. Retrospective analyses have shown mixed results about the ability of observational studies to replicate the findings from clinical trials. A systematic review comparing results from observational studies with randomized trials found the expected difference in effect estimates to be small. Some discordant findings, such as the association between hormone therapies and cardiovascular risk have been explained away through re-analysis, while other replication failures have been left unresolved. To our knowledge, there has been little work to prospectively evaluate the performance of retrospective observational database analyses in their ability to predict randomized trial results.

In this study, using the tofacitinib Phase 4 clinical trial as a case study, we will conduct a retrospective observational database analysis to replicate the clinical trial design and study population. We will report the results in anticipation of the trial completion, such that an objective assessment of observational study-RCT concordance can be made. We will provide an a priori framework for the study-RCT comparison, to avoid the risk of posthoc rationalization clouding the interpretation of the assessment. Further, we will examine the impact of trial’s inclusion criteria on the generalizability of the results across the real-world population currently using these treatments in practice. In this regard, we aim to not only provide specifically actionable information to clinicians about the comparative safety of tofacitinib, adalimumab and etanercept prior to the RCT read-out, but also provide a demonstration project to more generally support further policy about when retrospective observational databases can appropriately be used to inform regulatory decision-making.

# Study Objectives and Hypotheses

The objective of this study is to assess the cardiovascular safety of tofacitinib, in comparison to adalimumab and etanercept, in patients with RA, in both real world cohorts, and in cohorts that replicate trial participants in clinical trial NCT02092467

The primary hypotheses include:

1. There is no difference in the incidence of acute myocardial infarction between tofacitinib and adalimumab in patients with RA
2. There is no difference in the incidence of acute myocardial infarction between tofacitinib and etanercept in patients with RA

Secondary hypotheses include:

1. There is no difference in the incidence of ischemic stroke between tofacitinib and adalimumab in patients with RA
2. There is no difference in the incidence of ischemic stroke between tofacitinib and etanercept in patients with RA

# Research methods

## Overview

This is a retrospective, comparative cohort study based on OHDSI, a multi-database network, to examine the incidence of acute myocardial infarctions and ischemic stroke in patients with rheumatoid arthritis who were newly started on tofacitinib, adalimumab or etanercept.

The data sources are described in section 8.2.

The study population and cohort definition are described in section 8.3.

The outcomes are described in section 8.4.

The covariates and method to handle the covariate are described in section 8.5.

The analysis plan is described in section 9

## Data Source(s)

The analyses will be performed across OHDSI network in multiple observational databases. This database has been transformed into the OMOP Common Data Model, version 5.0 or higher. The complete specification for OMOP Common Data Model is available at: <https://github.com/OHDSI/CommonDataModel>.

Data sources expected to participate to include:

|  |  |  |
| --- | --- | --- |
| **Data Source Name** | **Contributor** | **Description** |
| Optum® De-Identified Clinformatic Data Mart Database – Date of Death (DOD) | Janssen | Optum’s de-identified Clinformatic Data Mart Database – Date of Death is a medical records database. The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using Natural Language Processing (NLP). |
| IBM MarketScan® Commercial Database (CCAE) | Janssen | Data from individuals enrolled in United States employer-sponsored insurance health plans. The data includes adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. Additionally, it captures laboratory tests for a subset of the covered lives. The patients in this database are aged under 65. |
| IBM MarketScan® Multi-State Medicaid Database (MDCD) | Janssen | Adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. The dataset lacks lab result data. |
| IBM MarketScan® Medicare Supplemental Database (MDCR) | Janssen | Represents health services of retirees (aged 65 or older) in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. These data include adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy). Additionally, it captures laboratory tests for a subset of the covered lives. |
| IQVIA LRxDx OpenClaims | IQVIA | Pre-adjudicated claims at the anonymized patient level collected from office-based physicians and specialists via office management software and clearinghouse switch sources for the purpose of reimbursement. We do have adjudicated claims for a subset of the medical claims data. |

## Study population

### Inclusion/Exclusion criteria for real world analysis (RWD)

For the real world comparison, the inclusion criteria for the study population are:

* a clinical diagnosis of rheumatoid arthritis;
* age over 18 years on index date;
* newly started on tofacitinib (TARGET cohort) or adalimumab (COMPARATOR 1 cohort) or etanercept (COMPARATOR 2 cohort), after January 1, 2013
* have at least 365 days of continuous observation time prior to the index date.

Index date is defined as the first date of tofacitinib, adalimumab or etanercept being prescribed. Those who fulfill the definition of more than one cohort will be included in the cohort that has the earliest index date.

We will exclude patients with a diagnosis of psoriasis, psoriatic arthritis, ankylosing spondylitis, uveitis, inflammatory bowel diseases, sarcoidosis and systemic lupus erythematosus.

### Inclusion/Exclusion criteria to replicate clinical trial

For the comparison that resembling the clinical trial, we rephrase the trial inclusion and exclusion criteria into executable search criteria in observational databases, as indicated in the following table:

|  |  |
| --- | --- |
| CLINICAL Trial (NCT02092467) | Observational study |
| Moderate to severe rheumatoid arthritis | Clinical diagnosis of rheumatoid arthritis, with use of methotrexate prior to index date. |
| Adults >/= 50 years | Adults >/= 50 years |
| Taking methotrexate without adequate control of symptoms | At least one record of prior exposure to methotrexate |
| Have at least one cardiovascular risk factor (eg, current smoker, high blood pressure, high cholesterol levels, diabetes mellitus, history of heart attack, family history of coronary heart disease, extra-articular RA disease) | At least one condition record for hypertension, hyperlipidemia, diabetes mellitus, or myocardial infarction prior to index exposure or at least one record suggesting current smoker status |
| No Current or recent infection | None. As in clinical practice, it is extremely unlikely to start biologics when a patient has active infection. |
| No Clinically significant laboratory abnormalities | None. Vague. Also it is unlikely that to start biologics in a patient with clinically significant laboratory abnormalities. |
| No Pregnancy | Not necessary, as we only study patients older than 50 years. |

Patients are followed until they 1) meet the primary endpoint by having AMI or Stroke; 2) exit the data source (censored); 3)discontinuation of the drug (defined by either a different biologic was prescribed), or 60-90 days after the last dispense of study drugs (censored).

For the secondary analyses, the criteria marked with a star (\*) will be removed one at a time, starting with the last one, to explore the influence of these restrictions on the estimated effect.

Both cohorts were further restricted to those subjects whose index date was in the time period when both T and C were observed.

### Executable Cohort Definitions

8.3.3.1 Real World Study

* TARGET: New users of tofacitinib with rheumatoid arthritis

URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

* COMPARATOR 1: New users of adalimumab with rheumatoid arthritis

URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

* COMPARATOR 2: New users of etanercept with rheumatoid arthritis

URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

8.3.3.2 Study replicating RCT criteria

* TARGET: New users of tofacitinib with rheumatoid arthritis satisfying RCT criteria

URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

* COMPARATOR 1: New users of adalimumab with rheumatoid arthritis satisfying RCT criteria

URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

* COMPARATOR 2: New users of etanercept with rheumatoid arthritis satisfying RCT criteria

URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

## Outcomes

### Primary outcome

Myocardial infarction: LEGEND definition [ref]: URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

### Secondary outcomes

Ischemic stroke: LEGEND definition [ref]: URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

Cardiovascular death: URL: <https://ohdsi.org/web/atlas/#/cohortdefinition/XXX>

### Negative control outcomes

Negative controls are concepts known to not be associated with the target or comparator cohorts, such that we can assume the true relative risk between the two cohorts is 1. Negative controls are selected using a similar process to that outlined by Voss et al. [[5](#_ENREF_5)]. Person counts of all potential drug-condition pairs are reviewed in observational data; this person count data helps determine which pairs are even probable for use in calibration. Given the list of potential drug-condition pairs, the concepts in the pairs must meet the following requirements to be considered as negative controls: (1) that there is no Medline abstract where the MeSH terms suggest an association between the drug and the condition [[6](#_ENREF_6)], (2) that there is no mention of the drug-condition pair on a US Product Label in the “Adverse Drug Reactions” or “Postmarketing” section [[7](#_ENREF_7)], (3) there are no US spontaneous reports suggesting that the pair is in an adverse event relationship [[8](#_ENREF_8), [9](#_ENREF_9)], (4) that the OMOP Vocabulary does not suggest that the drug is indicated for the condition, (5) that the concepts are usable (i.e. not too broad, not suggestive of an adverse event relationship, not pregnancy related), and (6) the exact concept itself is utilized in patient level data (i.e. concepts that are not usually used within the data are usually indicative a broad concept that has a child that is more specific). The remaining concepts are “optimized”, meaning parent concepts remove children as defined by the OMOP Vocabulary (e.g. if both “Non-Hodgkin’s Lymphoma” and “B-Cell Lymphoma” we selected, child concept “B-Cell Lymphoma would be removed for its parent “Non-Hodgkin’s Lymphoma”). Once potential negative control candidates were selected, manual clinical review to exclude any pairs that may still be in a causal relationship or similar to the study outcome will be performed to select the top concepts by patient exposure. The final list can be found in appendix 15.2.

For each negative control outcome, a patient enters the negative control outcome cohort at the occurrence of a diagnose code identified by the concepts listed above, or any one of its descendant codes.

## Covariates

### Propensity score covariates

Propensity scores (PS) will be used as an analytic strategy to reduce potential confounding due to imbalance between the target and comparator cohorts in baseline covariates. The propensity score is the probability of a patient being classified in the target cohort vs. the comparator cohort, given a set of observed covariates.

The types of baseline covariates used to fit the propensity score model will be:

* Demographics
  + Gender
  + Age group (5-year bands)
  + Index year
  + Index month
* Conditions
  + In prior 30d
  + In prior 365d
* Condition aggregation
  + SNOMED
* Drugs
  + In prior 30d
  + In prior 365d
  + Overlapping index date
* Drug aggregation
  + Ingredient
  + ATC Class
* Risk scores
  + Charlson comorbidity index

Specific covariates to be excluded from the propensity score model are labelled concepts to exclude as detailed in Appendix 15.1.

All covariates that occur in fewer than 0.1% of the persons between the target and comparator cohorts combined will be excluded prior to model fitting for computational efficiency.

# Data Analysis Plan

## Calculation of time-at risk

The time-at-risk will be defined in two ways: 1) as the ‘intent to treat’ period, defined as the time from the cohort start date to observation period end date, and 2) as a ‘on treatment’ period, defined as the time from cohort start date to cohort end date (representing the period of persistent exposure without switching or augmentation of therapy).

As previously described, the “on-treatment” period would be: 1) meet the primary endpoint by having MACE; 2) exit the data source (censored); 3) discontinuation of the drug (defined by a different biologic (code set 4) was prescribed), or 60-90 days after the last dispense of study drugs (censored).

## Model Specification

In this study, we compare the target cohort with the comparator cohort for the hazards of outcome during the time-at-risk by applying a Cox proportional hazards model.

The time-to-event of outcome among patients in the target and comparator cohorts is determined by calculating the number of days from the start of the time-at-risk window (the cohort start date), until the earliest event among 1) the first occurrence of the outcome, 2) the end of the time-at-risk window, and 3) the end of the observation period that spans the time-at-risk start.

Patients with the outcome, observed prior to target or comparator cohort entry are not excluded from consideration.

Propensity scores will be used as an analytic strategy to reduce potential confounding due to imbalance between the target and comparator cohorts in baseline covariates. The propensity score is the probability of a patient being classified in the target cohort vs. the comparator cohort, given a set of observed covariates. In this study, the propensity score is estimated for each patient, using the predicted probability from a regularized logistic regression model, fit with a Laplace prior (LASSO) and the regularization hyperparameter selected by optimizing the likelihood in a 10-fold cross validation using 10 replications per fold, a starting variance of 0.01 and a tolerance of 2e-7. Covariates to be used in the propensity score model are listed in section 8.5.

The target cohort and comparator cohorts will be stratified into five quantiles of the propensity score distribution. The final outcome model will apply a conditional Cox proportional hazard model, conditions on the propensity score strata.

### Pooling effect estimates across databases

Random-effects meta-analytic estimates were generated using the Hartung-Knapp-Sidik-Jonkman and Dersimonian Laird methods to pool evidence across the databases for all comparison-outcome-analyses where there was sufficient homogeneity (I2 < 40%).

## Analyses to perform

The following analyses will be performed:

* A = 4 comparisons: 2 main comparisons (tofacitinib\_RWD vs. adalimumab\_RWD; tofacitinib\_RWD vs. etanercept\_RWD; tofacitinib\_RCT vs. adalimumab\_RCT; tofacitinib\_RCT vs. etanercept\_RCT) using all inclusion criteria, and removing RCT criteria for each main comparison.
* B = 2 outcomes: 1) acute myocardial infarction, 2) ischemic stroke
* C = 2 time-at-risk definitions: 1) intent-to-treat, 2) on treatment
* D = 1 model: Cox regression using propensity score stratification
* E = 6 databases: Optum ClinFormatics Extended Datamart, Optum Pan-therapeutic EHR, Truven MarketScan CCAE, Truven MarketScan MDCD, Truven MarketScan MDCR, IQVIA Open Claims

The total number of analyses is therefore A/4 x B/2 x C/2 x D/1 x E/6 = 96 analyses.

## Output

Covariate balance will be summarized in tabular form by showing the mean value for all baseline covariates in the target and comparator cohort, with the associated standardized mean difference computed for each covariate.

Once the propensity score model is fit, we will plot the propensity score distribution of the target and comparator cohorts to evaluate the comparability of the two cohorts. The plot will be scaled to the preference score, normalizing for any imbalance in cohort size. The covariates selected within the propensity score model, with associated coefficients will also be reported.

A plot showing the propensity score distributions for both cohorts after stratification will be provided, with each quantile cut point shown as a vertical line. Covariate balance will be evaluated by plotting the standardized mean difference of each covariate before propensity score stratification against the standardized mean difference for each covariate after propensity score stratification.

An attrition diagram will be provided to detail the loss of patients from the original target cohort and comparator cohort to the subpopulations that remain after all design considerations have been applied.

The final outcome model, a conditional Cox proportional hazards model, will be summarized by providing the hazards ratio and associated 95% confidence interval. The number of persons, amount of time-at-risk, and number of outcomes in each cohort will also be reported.

## Evidence Evaluation

We have executed diagnostics to determine if the analysis can be appropriately conducted. The diagnostics include:

* Propensity score distribution
* Covariate balance before and after propensity score matching
* Estimation for negative and positive controls, to assess residual error
* Negative and positive control exposures and outcomes will be used to evaluate the potential impact of residual systematic error in the study design, and to facilitate empirical calibration of the p-value and confidence interval for the exposures and outcome of interest.

Negative control outcomes in the context of this study are outcomes that are not believed to be caused by neither tofacitinib nor adalimumab or etanercept, and where therefore the true hazard ratio is equal to 1. We will execute the same analysis used for the primary hypothesis to produce hazard ratio estimates for the negative controls. The distribution of effect estimates across all negative controls will be used to fit an empirical null distribution which models the observed residual systematic error. The empirical null distribution will then be applied to the target exposures and outcome of interest to calibrate the p-value [[11](#_ENREF_11)].

Positive control exposures and outcomes are pairs of exposures and outcomes where the hazard ratio is known to be of some magnitude greater than 1. We will synthesize positive controls by starting with the negative controls defined earlier, and adding additional, simulated outcomes during the time-at-risk until the desired true hazard ratio is achieved. The target hazard ratios are 1.5, 2 and 4. The negative and positive controls together will be used to estimate an empirical systematic error model, which will inform whether systematic error changes as a function of true effect size. The empirical systematic error model will then be applied to the target the target exposures and outcome of interest to calibrate the confidence interval [[10](#_ENREF_10)].

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. The calibration effect plot and calibration probability plots will be generated for review. We will report the traditional and empirically calibrated p-value and confidence interval for each negative control, as well as the hypothesis of interest.

# Study Diagnostics

## Sample Size and Study Power

## Cohort Comparability

## Systematic Error Assessment

# Strengths and Limitations of the Research Methods

Strength

* Cohort studies allow direct estimation of incidence rates following exposure of interest, and the new-user design can capture early events following treatment exposures while avoiding confounding from previous treatment effects. New use allows for a clear exposure index date.
* PS matching allow balancing on a large number of baseline potential confounders.
* Use of negative and positive control outcomes allow for evaluating the study design as a whole in terms of residual bias.

Limitations

* Even though many potential confounders will be included in this study, there may be residual bias due to unmeasured or misspecified confounders.

# Protection of Human Subjects

The use of the Optum Extended DoD database were reviewed by the New England Institutional Review Board (IRB) and was determined to be exempt from broad IRB approval, as this research project did not involve human subjects research.

# Management and Reporting of Adverse Events and Adverse Reactions

This study uses coded data that already exist in an electronic database. 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 are not reportable as individual adverse events reports. The study results will be assessed for medically important results.

# Plans for Disseminating and Communicating Study Results

The study results will be posted on the OHDSI website after completion of the study. At least one paper describing the study and its results will be written and submitted for publication to a peer-reviewed scientific journal.

# Appendix 1

## Concepts excluded from covariates

These concepts and their descendants are excluded when creating covariates: tofacitinib, adalimumab, etanercept drug concepts and their descendants.

|  |  |  |
| --- | --- | --- |
| Concept ID | Concept Name |  |

## Negative control outcomes

|  |  |
| --- | --- |
| Concept ID | outcomeName |
|  |  |

# References