Patient-Level Prediction:   
*Prediction of hemorrhagic transformation after cerebral infarction among patient**s: an OHDSI exercise*

Prepared on: 2018-10-12

Created by: Qiong Wang, Yuhui Zou, Jenna M. Reps, Peter R. Rijnbeek, Chan You, Erica A. Voss, Andrew Williams, Ross Williams, Mui Van Zandt, Ray Chen, Gowtham Rao, Kristin Feeney Kostka, Thomas Falconer, Jami N. Jackson Mulgrave, Suranga N. Kasthurirathne, Henry Morgan Stewart, Margarita Fernandez-Chas, Rohit Vashisht, Stephen R. Pfohl, Christian Reich, Patrick B. Ryan

# Table of Contents

[**Table of Contents**](#_631f97f8uaj6) **2**

[**List of Abbreviations**](#_akx0uqwty2j3) **4**

[**Responsible Parties**](#_vzcye1dwmtjo) **4**

[**Executive Summary**](#_vjqo2695i1jf) **4**

[**Rational & Background**](#_uh8dwzgrpnjb) **4**

[**Objective**](#_wa9kcd85pndk) **5**

[**Methods**](#_x1ihhv5vcp1v) **5**

[Study Design](#_9wm4t7yid0cr) 5

[Data Source(s)](#_84vxx841ck33) 6

[Internal Validation](#_ues3ov96sq3p) 6

[External Validation](#_977werel8zqh) 7

[Study Populations](#_t8jq63be1uv8) 14

[Target Cohort(s) [T]](#_6vj4f49qupg3) 14

[Outcome Cohorts(s) [O]](#_lriey44v7ls4) 15

[Time at Risk](#_huh41npjzanc) 15

[Additional Population Settings](#_dwgmkw7xys1u) 16

[Statistical Analysis Method(s)](#_p5pxpof2w4bb) 16

[Algorithms](#_kno246a3d006) 16

[Model Evaluation](#_plnaopk7a6pe) 17

[Quality Control](#_skxbtz6y5m4s) 19

[Tools](#_r16toapxyv2m) 19

[**Diagnostics**](#_jxdwbx1tn13) **19**

[**Data Analysis Plan**](#_tywr6hngzlhe) **19**

[Algorithm Settings](#_5sqe5qslcxgi) 19

[Covariate Settings](#_ad23wk7zjn7z) 20

[Model Development & Evaluation](#_ashyxof0g1yk) 22

[Analysis Execution Settings](#_xf12gbwu8khe) 23

[**Strengths & Limitations**](#_ijb15xejdw68) **23**

[**Protection of Human Subjects**](#_kb75a48fim6) **23**

[**Plans for Disseminating & Communicating Study Results**](#_shqishkc678h) **23**

[**Tables & Figures**](#_swafqhoualm0) **24**

[Incidence Rate of Target & Outcome](#_lvom7a7pspdu) 24

[Characterization](#_zl0sakkjyn0) 24

[**Appendices**](#_yxfcdzi9ocuv) **25**

[Study Generation Version Information](#_4ygba2bl0op5) 25

[Code List](#_mg1ybrtllsg8) 26

[Complete Analysis List](#_3cst5b7ldllq) 29

[**References**](#_u2bxgoxlzfgt) **30**

# List of Abbreviations

|  |  |
| --- | --- |
| **Abbreviation** | **Phrase** |
| AUC | Area Under the Receiver Operating Characteristic Curve |
| CDM | Common Data Model |
| O | Outcome Cohort |
| OHDSI | Observational Health Data Sciences & Informatics |
| OMOP | Observational Medical Outcomes Partnership |
| T | Target Cohort |
| TAR | Time at Risk |

# Responsible Parties

Members of the OHDSI Community.

Lead Author: Qiong Wang, MSc1,2,3

Supporting investigators: Jenna M Reps, PhD3,7, Kristin Feeney Kostka, MPH3,10, Patrick B Ryan, PhD3,5,7, Yuhui Zou, MD4, Peter R Rijnbeek, PhD3,9, RuiJen Chen, MD3,5,6, Gowtham Rao, MD, PhD3,7, Seng Chan You, MD, MS3,8, Henry Morgan Stewart, PhD3,10, , Erica A Voss, MPH,3,7,9 Andrew E Williams, PhD3,11, Ross D Williams, MSc3,9,Mui Van Zandt, BS3,10, Thomas Falconer, MS,3,5, Suranga N Kasthurirathne, PhD3,12,13, Margarita Fernandez-Chas, PhD3,10, Rohit Vashisht, PhD3,14, Stephen Pfohl, BEng3,14, Nigam Shah, MBBS, PhD3,14, Qing Jiang, PhD1, Christian Reich, MD, PhD3,10\*, Yi Zhou, PhD15\*

1Biomedical Engineering School, Sun-Yat Sen University, Guangzhou, China; 2The Third Affiliated Hospital of Guangzhou Medical University, Guangzhou, China; 3Observational Health Data Sciences and Informatics, New York, New York, USA; 4Department of Neurosurgery, General Hospital of Southern Theatre Command of PLA, Guangzhou, China; 5Department of Biomedical Informatics, Columbia University, New York, New York, USA; 6Department of Medicine, Weill Cornell Medical College, New York, New York, USA; 7Janssen Research and Development, Raritan, New Jersey, USA; 8Department of Biomedical informatics, Ajou University School of Medicine, Suwon, Korea; 9Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; 10IQVIA, Durham, North Carolina, USA; 11Tufts Medical Center, Institute for Clinical Research and Health Policy Studies, Boston, Massachusetts, USA; 12Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, Indiana, USA; 13Department of Epidemiology, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, Indiana, USA; 14Stanford Center for Biomedical Informatics Research, Stanford, California, USA; 15Department of Biomedical Engineering, Zhongshan School of Medicine, Sun Yat - Sen University, Guangzhou, China

# Executive Summary

The objective of this study is to develop and validate patient-level prediction models for patients in 2 target cohorts of patients with ischemic stroke treated with (i) all (ii) tread with antithrombotic to predict 1 outcome(s) of Cerebral Hemorrhage for 2 time at risk window(s):

1. Time at Risk Settings #1

Risk Window Start: 1, Add Exposure Days to Start: FALSE,

Risk Window End: 30, Add Exposure Days to End: FALSE,

1. Time at Risk Settings #2

Risk Window Start: 1, Add Exposure

Days to Start: FALSE, Risk Window End: 90, Add Exposure Days to End: FALSE).

The prediction will be implemented using 1 algorithms (Lasso Logistic Regression).

# Rational & Background

Hemorrhagic transformation (HT), which refers to a spectrum of ischemia-related brain hemorrhage, is a frequent spontaneous complication of ischemic stroke, especially after thrombolytic therapy (Sussman & Connolly, 2013). Therefore, HT limits the use of tissue plasminogen activator (tPA) treatment, the only method of clinical management of acute ischemic stroke. To search for new treatments as well as intervention measures for HT, it is important to understand its underlying mechanism and identify its predictors (Zhang, Yang, Sun, & Xing, 2014). Limited information about predictors is available, but some suggested atrial fibrillation, the nature of cerebral embolism and infarction, hyperglycemia, cholesterol levels could be useful.

Ischemic strokes occur when the arteries to the brain become narrowed or blocked, causing severely reduced blood flow (ischemia). Anticoagulants are often prescribed to thin the blood and reduce blood clots. Hemorrhagic Transformation occurs when a blood vessel in your brain leaks or ruptures and can occur after patients start. However it is unknown which patients exposed to an anticoagulant and had a prior ischemic stroke go on to have a hemorrhagic transformation within 7 days or 30 days.

*CITATION:*

*CITATION:* [*https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113*](https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113)

# Objective

The objective is to develop and validate patient-level prediction models for the following prediction problems:

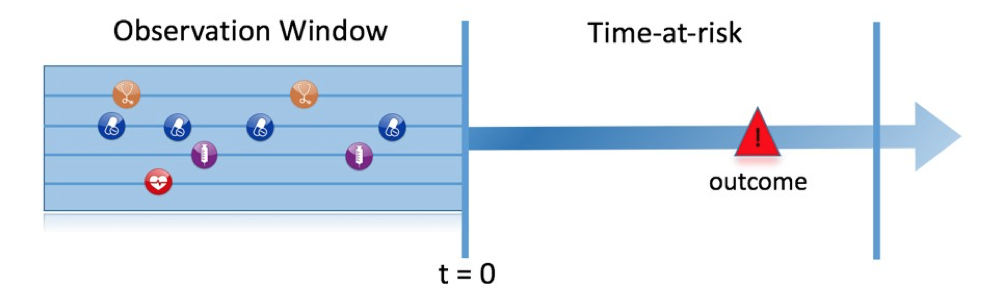
|  |  |  |
| --- | --- | --- |
| **Target Cohorts** | **Outcome Cohorts** | **Time at Risk** |
| [T1] Ischemic stroke + age >45-same day cerebral hemorrhagic  [T2] Ischemic stroke + antithrombotic + age >45-same day cerebral hemorrhagic | [O] Cerebral Hemorrhagic, exclude hemorrhagic caused by trauma | [Time at Risk Settings #1] Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 30, Add Exposure Days to End: FALSE,  [Time at Risk Settings #2] Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 90, Add Exposure Days to End: FALSE |

# Methods

## Study Design

This study will follow a retrospective, observational, patient-level prediction design. We define 'retrospective' to mean the study will be conducted using data already collected prior to the start of the study. We define 'observational' to mean there is no intervention or treatment assignment imposed by the study. We define 'patient-level prediction' as a modeling process wherein an outcome is predicted within a time at risk relative to the target cohort start and/or end date. Prediction is performed using a set of covariates derived using data prior to the start of the target cohort.

Figure 1, illustrates the prediction problem we will address. Among a population at risk, we aim to predict which patients at a defined moment in time (t = 0) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.



***Figure 1: The prediction problem***

*Citation: Jenna Reps, Martijn J. Schuemie, Marc A. Suchard, Patrick B. Ryan and Peter R. Rijnbeek (2018). PatientLevelPrediction: Package for patient level prediction using data in the OMOP Common DataModel. R package version 3.0.0.*

We follow the PROGRESS best practice recommendations for model development and the TRIPOD guidance for transparent reporting of the model results.

*Citation: Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. doi: 10.1371/journal.pmed.1001381. Epub 2013 Feb 5. Review. PubMed PMID: 23393430; PubMed Central PMCID: PMC3564751.*

*Citation: Collins, G., et al. (2017.02.01). 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.' from https://www.equator-network.org/reporting-guidelines/tripod-statement/*

## Data Source(s)

All analysis was performed on databases that were converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

*Citations: OMOP Common Data Model: 'OMOP Common Data Model (CDM).' from https://github.com/OHDSI/CommonDataModel.*

### Internal Validation

|  |  |
| --- | --- |
| **Optum’s Clinformatics® Extended Data Mart – Socio-Economic Status (OPTUM SES)** | |
| Database Description | Optum Clinformatics® Extended DataMart is an adjudicated administrative health claims database for members with private health insurance, who are fully insured in commercial plans or in administrative services only (ASOs), Legacy Medicare Choice Lives (prior to January 2006), and Medicare Advantage (Medicare Advantage Prescription Drug coverage starting January 2006). The population is primarily representative of US commercial claims patients (0-65 years old) with some Medicare (65+ years old) however ages are capped at 90 years. It includes data captured from administrative claims processed from inpatient and outpatient medical services and prescriptions as dispensed, as well as results for outpatient lab tests processed by large national lab vendors who participate in data exchange with Optum. Optum SES provides socio-economic status for members with both medical and pharmacy coverage and location information for patients it at the US Census Division.    The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted thirds-party laboratory service provider) for a non-random sample of the population (coded with LOINC codes). |

### External Validation

|  |  |
| --- | --- |
| **Truven Health MarketScan® Commercial Claims and Encounters Database (CCAE)** | |
| Database Description | Truven Health MarketScan® Commercial Claims and Encounters Database (CCAE) represent 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. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.    The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted thirds-party laboratory service provider) for a non-random sample of the population (coded with LOINC codes). |

|  |  |
| --- | --- |
| **Truven Health MarketScan® Medicare Supplemental and Coordination of Benefits Database (MDCR)** | |
| Database Description | Truven Health MarketScan® Medicare Supplemental and Coordination of Benefits Database (MDCR) represents health services of retirees 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.    The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted thirds-party laboratory service provider) for a non-random sample of the population (coded with LOINC codes). |

|  |  |
| --- | --- |
| **Truven Health MarketScan® Multi-State Medicaid Database (MDCD)** | |
| Database Description | Truven Health MarketScan® Multi-State Medicaid Database (MDCD) 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. Members maintain their same identifier even if they leave the system for a brief period however the dataset lacks lab result data.    The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes (NDC), inpatient and outpatient medical claims which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data does not contain laboratory results. |

|  |  |
| --- | --- |
| **Stanford University Translational Research Integrated Database Environment (STRIDE)** | |
| Database Description | Electronic health record data derived from all patients treated as outpatients and inpatients at Stanford Hospital and Clinics from 1995 to 2013, including structured clinical data and unstructured clinical notes. |

|  |  |
| --- | --- |
| **Regenstrief Institute / Indiana Network for Patient Care (INPC)** | |
| Database Description | Population-based, longitudinal, and structured coded and text data captured from hospitals, physician practices, public health departments, laboratories, radiology centers, pharmacies, pharmacy benefit managers, and payers in the Indiana Network. |

|  |  |
| --- | --- |
| **IQVIA Disease Analyzer (DA) Germany** | |
| Database Description | The IQVIA (Formerly QuintilesIMS) Disease Analyzer (DA) Germany database consists of data collected from physician practices and medical centers for all ages. Mostly primary care physician data however some data from specialty practices (where practices are electronically connected to each other) and some lab data is included. Key attributes include demographics, prescriptions as prescribed at brand level, diagnosis, lab measurements, actions (e.g. referrals, sick notes). |

## 

|  |  |
| --- | --- |
| **IQVIA PharMetrics Plus** | |
| Database Description | Patient-Centric, Closed Claims Database of fully adjudicated pharmacy, hospital and medical claims at the anonymized patient level sourced from commercial payers |

## 

|  |  |
| --- | --- |
| **IQVIA Hospital Charge Detail Master (CDM)** | |
| Database Description | Anonymized patient level data are sourced from hospital charge detail masters (CDM) and collected from resource management software within short-term, acute-care and non-federal hospitals. |

## 

|  |  |
| --- | --- |
| **IQVIA Open Claims** | |
| Database Description | 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 Populations

### Target Cohort(s) [T]

|  |  |
| --- | --- |
| **Cohort Name** | **Description** |
| T1 -PLP-T(Cerebral infarction+Warfarin) | Qiong-PLP-T(Cerebral infarction+Warfarin)  Initial Event Cohort  People having any of the following:   * a condition occurrence of Qiong-Concept(Cerebral infarction)1   with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.  For people matching the Primary Events, include:  Having all of the following criteria:   * with the following event criteria:   + with age >= 45   Limit cohort of initial events to: all events per person.  Inclusion Rules  Inclusion Criteria #1: Warfarin  Having all of the following criteria:   * at least 1 occurrences of a drug exposure of Qiong-Drug(Warfarin)2   starting between 7 days Before and 30 days After event index date  Limit qualifying cohort to: all events per person.  End Date Strategy  No end date strategy selected. By default, the cohort end date will be the end of the observation period that contains the index event.  Cohort Collapse Strategy:  Collapse cohort by era with a gap size of 0 days. |
| T2-PLP-T(Cerebral infarction+ Dabigatran) | Qiong-PLP-T(cerebral infarction+Dabigatran)  Initial Event Cohort  People having any of the following:   * a condition occurrence of Any Condition   with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.  For people matching the Primary Events, include:  Having all of the following criteria:   * with the following event criteria:   + with age >= 45   Limit cohort of initial events to: all events per person.  Inclusion Rules  Inclusion Criteria #1: Drug-Dabigatran  Having all of the following criteria:   * at least 1 occurrences of a drug exposure of Qiong-Drug(Dabigatran)2   starting between 7 days Before and 30 days After event index date  Limit qualifying cohort to: all events per person.  End Date Strategy  Date Offset Exit Criteria  This cohort defintion end date will be the index event's start date plus 30 days  Cohort Collapse Strategy:  Collapse cohort by era with a gap size of 0 days. |
| T3-PLP-T(Cerebral infarction+ aspirin) | Qiong-PLP-T(cerebral infarction+aspirin)  Initial Event Cohort  People having any of the following:  a condition occurrence of Any Condition  with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.  For people matching the Primary Events, include:  Having all of the following criteria:  with the following event criteria:  with age >= 45  Limit cohort of initial events to: all events per person.  Inclusion Rules  Inclusion Criteria #1: Drug(aspirin)  Having any of the following criteria:  at least 1 occurrences of a drug exposure of Qiong-Drug(Aspirin)6  starting between 7 days Before and 30 days After event index date  Limit qualifying cohort to: all events per person.  End Date Strategy  Date Offset Exit Criteria  This cohort defintion end date will be the index event's start date plus 30 days  Cohort Collapse Strategy:  Collapse cohort by era with a gap size of 0 days. |
| T4-PLP-T(Cerebral infarction+ Clopidogrel) | Qiong-PLP-T(Cerebral infarction+Clopidogrel)  Initial Event Cohort  People having any of the following:   * a condition occurrence of Qiong-Concept(Cerebral infarction)5   with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.  For people matching the Primary Events, include:  Having all of the following criteria:   * with the following event criteria:   + with age >= 45   Limit cohort of initial events to: all events per person.  Inclusion Rules  Inclusion Criteria #1: Clopidogrel  Having all of the following criteria:   * at least 1 occurrences of a drug exposure of Qiong-Drug(Clopidogrel)12   starting between 7 days Before and 30 days After event index date  Limit qualifying cohort to: all events per person.  End Date Strategy  Date Offset Exit Criteria  This cohort defintion end date will be the index event's start date plus 30 days  Cohort Collapse Strategy:  Collapse cohort by era with a gap size of 0 days. |
| T5-PLP-T(Cerebral infarction+ Riberoxaban) | Qiong-PLP-T(Cerebral infarction+Riberoxaban)  Initial Event Cohort  People having any of the following:   * a condition occurrence of Qiong-Concept(Cerebral infarction)5   with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.  For people matching the Primary Events, include:  Having all of the following criteria:   * with the following event criteria:   + with age >= 45   Limit cohort of initial events to: all events per person.  Inclusion Rules  Inclusion Criteria #1: Riberoxaban  Having all of the following criteria:   * at least 1 occurrences of a drug exposure of Qiong-Drug(Rivaroxaban)15   starting between all days Before and all days After event index date  Limit qualifying cohort to: all events per person.  End Date Strategy  No end date strategy selected. By default, the cohort end date will be the end of the observation period that contains the index event.  Cohort Collapse Strategy:  Collapse cohort by era with a gap size of 0 days. |

The figure below is a visual description of the cohort. [The picture Needs revise]

**INDEX**

Exposure to an Anticoagulant

Diagnosed with Ischemic stroke

7 days

30 days

No Hemorrhagic Stroke

### Outcome Cohorts(s) [O]

|  |  |
| --- | --- |
| **Cohort Name** | **Description** |
| O1 -PLP-O(cerebral hemorrhagic) | Initial Event Cohort  People having any of the following:  a condition occurrence of Qiong-Conditon(Cerebral Hemorrhage)1  with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: all events per person.  Limit qualifying cohort to: earliest event per person.  End Date Strategy  No end date strategy selected. By default, the cohort end date will be the end of the observation period that contains the index event.  Cohort Collapse Strategy:  Collapse cohort by era with a gap size of 0 days.  Reference 1 = Appendices, Code List, Concept Set #1 |

### Time at Risk

|  |
| --- |
| **Time at Risk** |
| ***[Time at Risk Settings #1]***  Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 30, Add Exposure Days to End: FALSE |
| ***[Time at Risk Settings #2]***  Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 7, Add Exposure Days to End: FALSE |

### 

### Additional Population Settings

***Population Settings #1***

|  |  |
| --- | --- |
| **Item** | **Settings** |
| minTimeAtRisk | 364 |
| requireTimeAtRisk | TRUE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 1 |
| washoutPeriod | 365 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 99999 |
| Binary | TRUE |
| removeSubjectsWithPriorOutcome | FALSE |
| riskWindowEnd | 365 |
| firstExposureOnly | TRUE |

***Population Settings #2***

|  |  |
| --- | --- |
| **Item** | **Settings** |
| minTimeAtRisk | 29 |
| requireTimeAtRisk | TRUE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 1 |
| washoutPeriod | 365 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 99999 |
| Binary | TRUE |
| removeSubjectsWithPriorOutcome | FALSE |
| riskWindowEnd | 30 |
| firstExposureOnly | TRUE |

## Statistical Analysis Method(s)

### Algorithms

|  |  |
| --- | --- |
| **Algorithm** | **Description** |
| Lasso Logistic Regression | Lasso logistic regression belongs to the family of generalized linear models, where a linear combination of the variables is learned and finally a logistic function maps the linear combination to a value between 0 and 1. The lasso regularization adds a cost based on model complexity to the objective function when training the model. This cost is the sum of the absolute values of the linear combination of the coefficients. The model automatically performs feature selection by minimizing this cost. We use the Cyclic coordinate descent for logistic, Poisson and survival analysis (Cyclops) package to perform large-scale regularized logistic regression: https://github.com/OHDSI/Cyclops |

### Model Evaluation

The following evaluations will be performed on the model:

|  |  |
| --- | --- |
| **Evaluation** | **Description** |
| Box Plots | The prediction distribution boxplots are box plots for the predicted risks of the people in the test set with the outcome (class 1: blue) and without the outcome (class 0: red). |
| Calibration Plot | The calibration plot shows how close the predicted risk is to the observed risk. The diagonal dashed line thus indicates a perfectly calibrated model. The ten (or fewer) dots represent the mean predicted values for each quantile plotted against the observed fraction of people in that quantile who had the outcome (observed fraction). The straight black line is the linear regression using these 10 plotted quantile mean predicted vs observed fraction points. The two blue straight lines represented the 95% lower and upper confidence intervals of the slope of the fitted line. |
| Demographic Summary Plot | This plot shows for females and males the expected and observed risk in different age groups together with a confidence area. |
| Precision Recall Plot | The precision-recall curve is valuable for dataset with a high imbalance between the size of the positive and negative class. It shows the tradeoff between precision and recall for different threshold. High precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). A high area under the curve represents both high recall and high precision. |
| Prediction Distribution Plots | The preference distribution plots are the preference score distributions corresponding to i) people in the test set with the outcome (red) and ii) people in the test set without the outcome (blue). |
| ROC Plot | The ROC plot plots the sensitivity against 1-specificity on the test set. The plot shows how well the model is able to discriminate between the people with the outcome and those without. The dashed diagonal line is the performance of a model that randomly assigns predictions. The higher the area under the ROC plot the better the discrimination of the model. |
| Smooth Calibration Plot | Similar to the traditional calibration shown above the Smooth Calibration plot shows the relationship between predicted and observed risk. the major difference is that the smooth fit allows for a more fine grained examination of this. Whereas the traditional plot will be heavily influenced by the areas with the highest density of data the smooth plot will provide the same information for this region as well as a more accurate interpretation of areas with lower density. the plot also contains information on the distribution of the outcomes relative to predicted risk. However the increased information game comes at a computational cost. It is recommended to use the traditional plot for examination and then to produce the smooth plot for final versions. |
| Test-Train Similarity Plot | The test-train similarity is presented by plotting the mean covariate values in the train set against those in the test set for people with and without the outcome. |
| Variable Scatter Plot | The variable scatter plot shows the mean covariate value for the people with the outcome against the mean covariate value for the people without the outcome. The size and color of the dots correspond to the importance of the covariates in the trained model (size of beta) and its direction (sign of beta with green meaning positive and red meaning negative), respectively. |

## Quality Control

The PatientLevelPrediction package itself, as well as other OHDSI packages on which PatientLevelPrediction depends, use unit tests for validation.

*Citation: Jenna Reps, Martijn J. Schuemie, Marc A. Suchard, Patrick B. Ryan and Peter R. Rijnbeek (2018). PatientLevelPrediction: Package for patient level prediction using data in the OMOP Common DataModel. R package version 3.0.0.*

## Tools

This study will be designed using OHDSI tools and run with R.

*Citation: R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*

More information about the tools can be found in the Appendix 'Study Generation Version Information'.

# Diagnostics

Reviewing the incidence rates of the outcomes in the target population prior to performing the analysis will allow us to assess its feasibility. The full table can be found in the 'Table and Figures' section under 'Incidence Rate of Target & Outcome'.

Additionally, reviewing the characteristics of the cohorts provides insight into the cohorts being reviewed. The full table can be found below in the 'Table and Figures' section under 'Characterization'.

# Data Analysis Plan

## Algorithm Settings

***Model Settings Settings #1 - LassoLogisticRegressionSettings***

|  |  |
| --- | --- |
| **Covariates** | **Settings** |
| seed |  |
| variance | 0.01 |

## Covariate Settings

The covariates (constructed using records on or prior to the target cohort start date) are used within this prediction mode include the following. Each covariate needs to contain at least 0.001 subjects to be considered for the model.

***Covariate Settings #1***

|  |  |
| --- | --- |
| **Covariates** | **Settings** |
| VisitCountMediumTerm | FALSE |
| ObservationShortTerm | FALSE |
| shortTermStartDays | -30 |
| MeasurementRangeGroupShortTerm | FALSE |
| ConditionOccurrenceLongTerm | FALSE |
| DrugEraStartLongTerm | FALSE |
| VisitCountShortTerm | FALSE |
| Chads2Vasc | TRUE |
| ConditionGroupEraStartLongTerm | FALSE |
| ConditionEraShortTerm | FALSE |
| Dcsi | TRUE |
| DrugGroupEraLongTerm | TRUE |
| DrugGroupEraShortTerm | FALSE |
| ConditionEraStartLongTerm | FALSE |
| temporal | FALSE |
| DemographicsIndexMonth | TRUE |
| ConditionOccurrencePrimaryInpatientLongTerm | FALSE |
| ConditionEraAnyTimePrior | FALSE |
| addDescendantsToInclude | FALSE |
| ConditionGroupEraStartMediumTerm | FALSE |
| ProcedureOccurrenceLongTerm | TRUE |
| DrugExposureLongTerm | FALSE |
| DrugEraStartShortTerm | FALSE |
| DistinctIngredientCountMediumTerm | FALSE |
| DistinctMeasurementCountShortTerm | FALSE |
| MeasurementRangeGroupLongTerm | FALSE |
| ConditionGroupEraOverlapping | FALSE |
| MeasurementRangeGroupMediumTerm | FALSE |
| DrugGroupEraStartMediumTerm | FALSE |
| MeasurementAnyTimePrior | TRUE |
| MeasurementMediumTerm | FALSE |
| includedCovariateIds |  |
| ConditionOccurrenceAnyTimePrior | FALSE |
| DistinctConditionCountLongTerm | FALSE |
| MeasurementValueLongTerm | FALSE |
| DrugEraShortTerm | FALSE |
| DrugGroupEraAnyTimePrior | TRUE |
| DrugEraOverlapping | FALSE |
| ConditionOccurrencePrimaryInpatientAnyTimePrior | FALSE |
| ConditionEraMediumTerm | FALSE |
| ConditionEraOverlapping | FALSE |
| ConditionEraStartShortTerm | FALSE |
| ObservationAnyTimePrior | TRUE |
| VisitConceptCountShortTerm | FALSE |
| DemographicsEthnicity | TRUE |
| DistinctIngredientCountLongTerm | FALSE |
| ConditionOccurrencePrimaryInpatientShortTerm | FALSE |
| DemographicsAgeGroup | TRUE |
| DistinctProcedureCountShortTerm | FALSE |
| DistinctObservationCountMediumTerm | FALSE |
| includedCovariateConceptIds |  |
| DrugGroupEraStartShortTerm | FALSE |
| addDescendantsToExclude | FALSE |
| DrugEraLongTerm | FALSE |
| DistinctConditionCountShortTerm | FALSE |
| ConditionGroupEraShortTerm | FALSE |
| ConditionEraStartMediumTerm | FALSE |
| VisitCountLongTerm | FALSE |
| DemographicsRace | TRUE |
| ProcedureOccurrenceAnyTimePrior | TRUE |
| DistinctObservationCountLongTerm | FALSE |
| ProcedureOccurrenceMediumTerm | FALSE |
| CharlsonIndex | TRUE |
| DemographicsPriorObservationTime | FALSE |
| MeasurementShortTerm | FALSE |
| DistinctProcedureCountMediumTerm | FALSE |
| ConditionEraLongTerm | FALSE |
| DrugGroupEraStartLongTerm | FALSE |
| DemographicsGender | TRUE |
| DeviceExposureAnyTimePrior | TRUE |
| ObservationLongTerm | TRUE |
| DemographicsIndexYearMonth | FALSE |
| ConditionOccurrenceMediumTerm | FALSE |
| longTermStartDays | -365 |
| DemographicsAge | FALSE |
| DrugGroupEraOverlapping | FALSE |
| DistinctMeasurementCountLongTerm | FALSE |
| MeasurementRangeGroupAnyTimePrior | FALSE |
| DistinctConditionCountMediumTerm | FALSE |
| DrugGroupEraMediumTerm | FALSE |
| ProcedureOccurrenceShortTerm | FALSE |
| ObservationMediumTerm | FALSE |
| ConditionGroupEraAnyTimePrior | TRUE |
| Chads2 | TRUE |
| DrugExposureAnyTimePrior | FALSE |
| DeviceExposureLongTerm | TRUE |
| DemographicsTimeInCohort | FALSE |
| DistinctMeasurementCountMediumTerm | FALSE |
| MeasurementValueShortTerm | FALSE |
| DeviceExposureMediumTerm | FALSE |
| ConditionGroupEraStartShortTerm | FALSE |
| ConditionOccurrencePrimaryInpatientMediumTerm | FALSE |
| MeasurementLongTerm | TRUE |
| DemographicsIndexYear | FALSE |
| MeasurementValueMediumTerm | FALSE |
| DrugEraStartMediumTerm | FALSE |
| MeasurementValueAnyTimePrior | FALSE |
| DistinctObservationCountShortTerm | FALSE |
| DrugEraMediumTerm | FALSE |
| ConditionGroupEraLongTerm | TRUE |
| DrugExposureShortTerm | FALSE |
| DistinctIngredientCountShortTerm | FALSE |
| DeviceExposureShortTerm | FALSE |
| mediumTermStartDays | -180 |
| DemographicsPostObservationTime | FALSE |
| VisitConceptCountLongTerm | FALSE |
| VisitConceptCountMediumTerm | FALSE |
| excludedCovariateConceptIds |  |
| ConditionGroupEraMediumTerm | FALSE |
| DrugExposureMediumTerm | FALSE |
| DistinctProcedureCountLongTerm | FALSE |
| DrugEraAnyTimePrior | FALSE |
| endDays | 0 |
| ConditionOccurrenceShortTerm | FALSE |

## Model Development & Evaluation

To build and internally validate the models, we will partition the labelled data into a train set (75%) and a test set (25%).

The hyper-parameters for the models will be assessed using 3-fold cross validation on the train set and a final model will be trained using the full train set and optimal hyper-parameters.

The internal validity of the models will be assessed on the test set. We will use the area under the receiver operating characteristic curve (AUC) to evaluate the discriminative performance of the models and plot the predicted risk against the observed fraction to visualize the calibration. See 'Model Evaluation' section for more detailed information about additional model evaluation metrics.

## Analysis Execution Settings

There are 1 target cohorts evaluated for 1 outcomes over 1 models over 1 covariates settings and over 2 population settings. In total there are 2 analysis performed. For a full list refer to appendix 'Complete Analysis List'.

# Strengths & Limitations

* Some strengths to consider:
* Some limitations to consider:
  + It may not be possible to develop prediction models for rare outcomes.
  + Not all medical events are recorded into the observational datasets and some recordings can be incorrect. This could potentially lead to outcome misclassification.
  + The prediction models are only applicable to the population of patients represented by the data used to train the model and may not be generalizable to the wider population.

# 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.

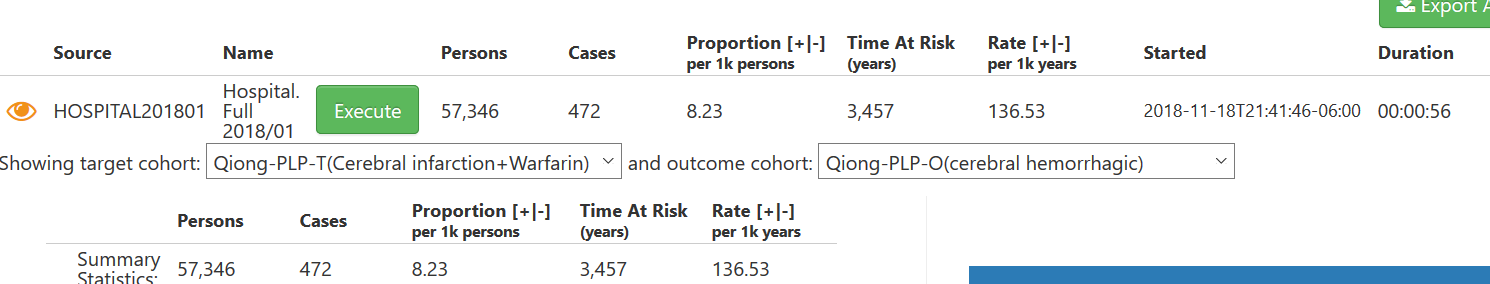
# Plans for Disseminating & Communicating Study Results

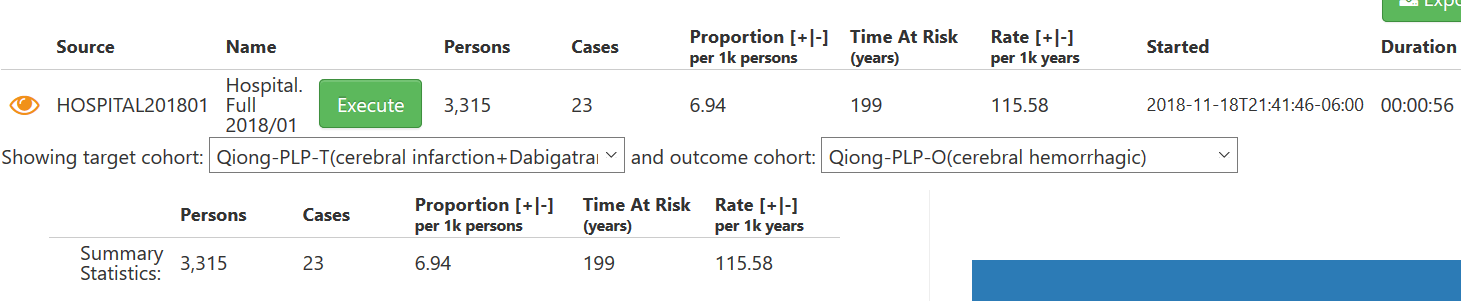
The goal will be to publish our findings. The target journal will be Stroke as JANE [*CITATION: http://jane.biosemantics.org/index.php*] suggests this to be one of the top five best fits based on the abstract. Secondary journal would be Neurobiology of Disease as this was also recommended by JANE.

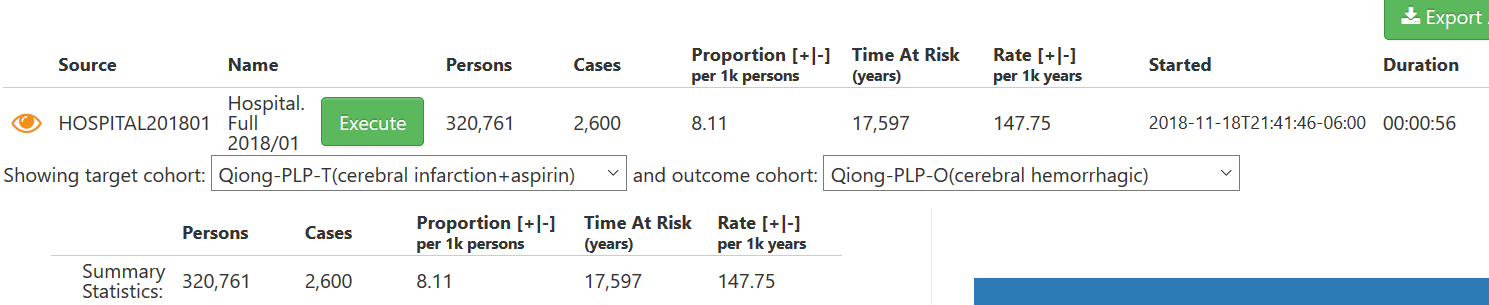
# Tables & Figures

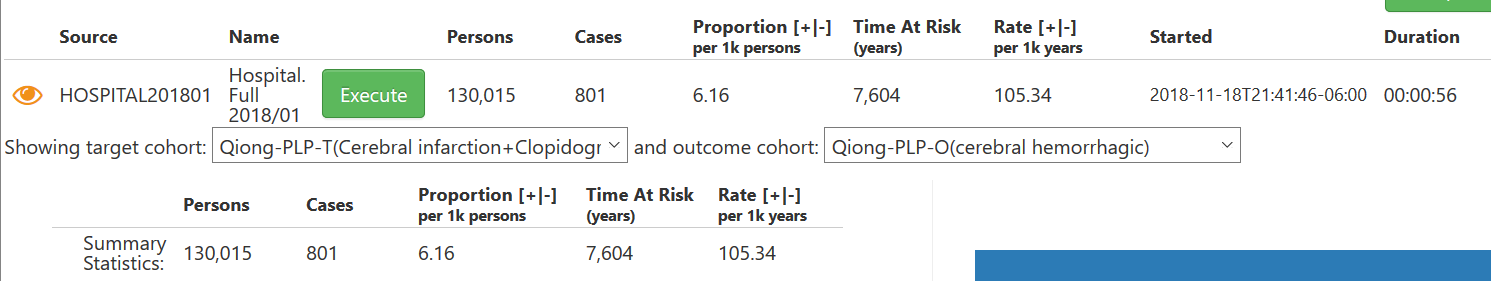
## Incidence Rate of Target & Outcome

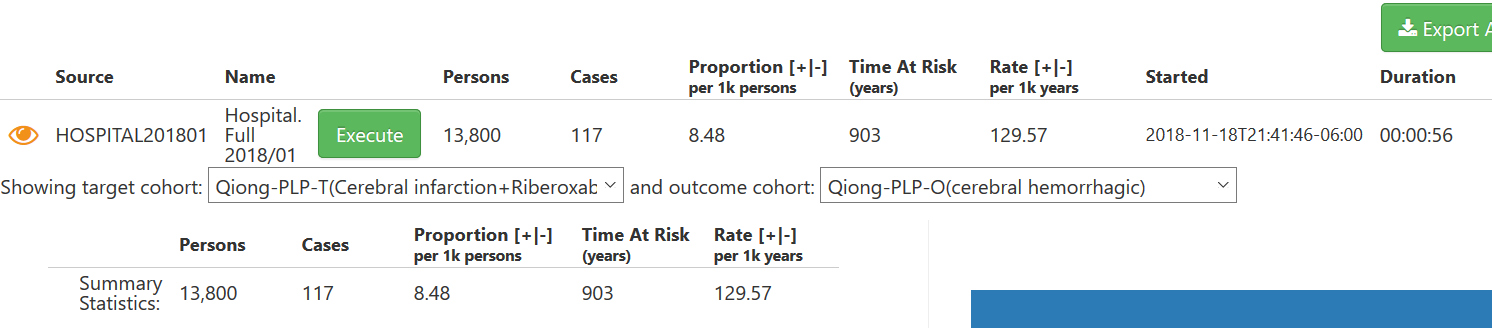
*<< add incidence here. >>*











## Characterization

To be completed.

*<< add characterization table here. >>*

*<< add results here. >>*

# Appendices

## Study Generation Version Information

Skeleton Version: PatientLevelPredictionStudy - v0.0.1

Identifier / Organization: Janssen Research and Development

## Code List

***Concept Set #1 - Qiong-Concept(Cerebral infarction)***

1. Qiong-Concept(Cerebral infarction)

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 4110189 | Cerebral infarct due to thrombosis of precerebral arteries | Condition | SNOMED | NO | YES | NO |
| 443454 | Cerebral infarction | Condition | SNOMED | NO | YES | NO |
| 4111714 | Cerebral infarction due to cerebral venous thrombosis, non-pyogenic | Condition | SNOMED | NO | YES | NO |
| 4108356 | Cerebral infarction due to embolism of cerebral arteries | Condition | SNOMED | NO | YES | NO |
| 45772786 | Cerebral infarction due to embolism of middle cerebral artery | Condition | SNOMED | NO | YES | NO |
| 4110190 | Cerebral infarction due to embolism of precerebral arteries | Condition | SNOMED | NO | YES | NO |
| 46273649 | Cerebral infarction due to occlusion of basilar artery | Condition | SNOMED | NO | YES | NO |
| 46270031 | Cerebral infarction due to occlusion of precerebral artery | Condition | SNOMED | NO | YES | NO |
| 4110192 | Cerebral infarction due to thrombosis of cerebral arteries | Condition | SNOMED | NO | YES | NO |
| 45767658 | Cerebral infarction due to thrombosis of middle cerebral artery | Condition | SNOMED | NO | YES | NO |
| 40479572 | Infarct of cerebrum due to iatrogenic cerebrovascular accident | Condition | SNOMED | NO | YES | NO |
| 4043731 | Infarction - precerebral | Condition | SNOMED | NO | YES | NO |
| 4046360 | Lacunar infarction | Condition | SNOMED | NO | YES | NO |
| 4045737 | Pure motor lacunar infarction | Condition | SNOMED | NO | YES | NO |
| 4045738 | Pure sensory lacunar infarction | Condition | SNOMED | NO | YES | NO |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| O1 - Hemorrhagic Stroke |
| T1 - Anticoagulation (or antiplatelet) and ischemic stroke |

***Concept Set #2 - Qiong-Drug(Warfarin)***

Qiong-Drug(Warfarin)

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 1310149 | Warfarin | Drug | RxNorm | NO | YES | NO |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T1 -PLP-T(Cerebral infarction+Warfarin) |

***Concept Set #3- Qiong-Drug(Dabigatran)***

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 40228152 | dabigatran etexilate | Drug | RxNorm | NO | YES | NO |
| 35606208 | dabigatran etexilate 110 MG Oral Capsule | Drug | RxNorm | NO | YES | NO |
| 40228154 | dabigatran etexilate 150 MG Oral Capsule | Drug | RxNorm | NO | YES | NO |
| 40228158 | dabigatran etexilate 150 MG Oral Capsule [Pradaxa] | Drug | RxNorm | NO | YES | NO |
| 40228161 | dabigatran etexilate 75 MG Oral Capsule | Drug | RxNorm | NO | YES | NO |
| 40228162 | dabigatran etexilate 75 MG Oral Capsule [Pradaxa] | Drug | RxNorm | NO | YES | NO |
| 40228164 | dabigatran etexilate Oral Capsule | Drug | RxNorm | NO | YES | NO |
| 40228165 | dabigatran etexilate Oral Capsule [Pradaxa] | Drug | RxNorm | NO | YES | NO |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T2 -PLP-T(Cerebral infarction+Dabigatran) |

***Concept Set #4- Qiong-Drug(Aspirin)***

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 1112807 | Aspirin | Drug | RxNorm | NO | YES | NO |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T3 -PLP-T(Cerebral infarction+Aspirin) |

***Concept Set #4- Qiong-Drug(Clopidogrel)***

Qiong-Drug(Clopidogrel)

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 1322184 | clopidogrel | Drug | RxNorm | NO | YES | NO |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T4 -PLP-T(Cerebral infarction+Clopidogrel) |

***Concept Set #4- Qiong-Drug(Rivearoxaban)***

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 40241331 | rivaroxaban | Drug | RxNorm | NO | YES | NO |
| 40241332 | rivaroxaban 10 MG | Drug | RxNorm | NO | YES | NO |
| 40241333 | rivaroxaban 10 MG Oral Tablet | Drug | RxNorm | NO | YES | NO |
| 40241334 | rivaroxaban 10 MG Oral Tablet [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 40241335 | rivaroxaban 10 MG [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 40244443 | rivaroxaban 15 MG | Drug | RxNorm | NO | YES | NO |
| 40244444 | rivaroxaban 15 MG Oral Tablet | Drug | RxNorm | NO | YES | NO |
| 40244445 | rivaroxaban 15 MG Oral Tablet [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 40244446 | rivaroxaban 15 MG [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 40244447 | rivaroxaban 20 MG | Drug | RxNorm | NO | YES | NO |
| 40244448 | rivaroxaban 20 MG Oral Tablet | Drug | RxNorm | NO | YES | NO |
| 40244449 | rivaroxaban 20 MG Oral Tablet [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 40244450 | rivaroxaban 20 MG [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 36227606 | rivaroxaban Oral Product | Drug | RxNorm | NO | YES | NO |
| 40241336 | rivaroxaban Oral Tablet | Drug | RxNorm | NO | YES | NO |
| 40241337 | rivaroxaban Oral Tablet [Xarelto] | Drug | RxNorm | NO | YES | NO |
| 36227607 | rivaroxaban Pill | Drug | RxNorm | NO | YES | NO |
| 45777060 | {42 (rivaroxaban 15 MG Oral Tablet [Xarelto]) / 9 (rivaroxaban 20 MG Oral Tablet [Xarelto]) } Pack [Xarelto Kit] | Drug | RxNorm | NO | YES | NO |
| 45777059 | {42 (rivaroxaban 15 MG Oral Tablet) / 9 (rivaroxaban 20 MG Oral Tablet) } Pack | Drug | RxNorm | NO | YES | NO |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T5 -PLP-T(Cerebral infarction+Rivaroxaban) |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T1 - Anticoagulation (or antiplatelet) and ischemic stroke |

Cohorts that use this Concept Set:

|  |
| --- |
| **Cohort Name** |
| T1 - Anticoagulation (or antiplatelet) and ischemic stroke |

## Complete Analysis List

Below is a complete list of analysis that will be performed. Definitions for the column 'Covariate Settings ID' can be found above in the 'Covariate Settings' section. Definitions for the 'Population Settings Id' can be found above in the 'Additional Population Settings' section.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Target Cohort Name** | **Outcome Cohort Name** | **Model Settings Id** | **Model Settings Description** | **Covariate Settings ID** | **Population Settings ID** |
| 1 | T1 - Anticoagulation (or antiplatelet) and ischemic stroke | O1 - Hemorrhagic  Event | 1 | Lasso Logistic Regression | 1 | 1 |
| 2 | T1 - Anticoagulation (or antiplatelet) and ischemic stroke | O1 - Hemorrhagic  Event | 1 | Lasso Logistic Regression | 1 | 2 |

*<< add models here >>*

# References

To be completed.