Proof of concept demonstrating the utilization of the OHDSI data network to perform external model validation of existing stroke models.

**Authors:**

Jenna Reps, PhD, Janssen Research and Development

Clair Blacketer, MPH, Janssen Research and Development

…

Patrick Ryan, PhD, Janssen Research and Development

Peter Rijnbeek, PhD, Erasmus MC, Rotterdam, The Netherlands

**Date:** 2 April 2018

**Acknowledgement:** The analysis is based in part on work from the Observational Health Sciences and Informatics collaborative. OHDSI (<http://ohdsi.org>) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

The authors declare the following disclosures: Patrick Ryan, PhD, Clair Blacketer, MPH, and Jenna Reps, PhD are employees of Janssen Research & Development.

# Table of Contents

[2 List of abbreviations 3](#_Toc472329131)

[3 Abstract 3](#_Toc472329132)

[4 Amendments and Updates 3](#_Toc472329133)

[5 Milestones 4](#_Toc472329134)

[6 Rationale and Background 5](#_Toc472329135)

[7 Research Questions and Objectives 5](#_Toc472329136)

[7.1 Research Questions 5](#_Toc472329137)

[8 Research methods 6](#_Toc472329138)

[8.1 Study Design 6](#_Toc472329139)

[8.1.1 Overview 6](#_Toc472329140)

[8.1.2 Study population 6](#_Toc472329145)

[8.1.3 Additional analysis details 7](#_Toc472329146)

[8.1.4 Analysis variations 7](#_Toc472329158)

[8.2 Variables 7](#_Toc472329159)

[8.2.1 At risk 7](#_Toc472329160)

[8.2.2 Outcome 7](#_Toc472329161)

[8.3 Data Sources 8](#_Toc472329162)

[8.4 Quality control 8](#_Toc472329163)

[8.5 Strengths and Limitations of the Research Methods 8](#_Toc472329164)

[9 Protection of Human Subjects 9](#_Toc472329168)

[10 Plans for Disseminating and Communicating Study Results 9](#_Toc472329169)

[11 References 9](#_Toc472329170)

# List of abbreviations

PLP Patient Level Prediction

OHDSI Observational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

# Abstract

Observational medical datasets, such as electronic healthcare records and insurance claims databases, present the opportunity to learn about disease progression and develop patient-level prediction models. Many prediction models have been developed [1-2] but the majority fail to make clinical impacts [3]. One of the main obstacles preventing model uptake is that lack knowledge about how transportable a model is. For example, a model developed using USA claims data may perform well in the US population but may not transport to Europe or Asia. In general the population used to develop the model may not be representative of the general population, so the true performance on a wider population may differ. The type of data the model was trained on (e.g., variables available) may also limit the model’s transportability to other datasets that lack an important variable. In addition, a model may obtain optimistic performance on the data use to develop it when poor model development practices are implemented (e.g., the performance may be optimistic if some data are not left out to perform internal validation). To address these concerns it is important to validate models on new datasets and gain insight into how well the model generalizes (how well it performs on similar data, e.g., validating a model trained on one US adult EMR by applying it to a different US adult EMR) and transports (how well it performs on different data, e.g., validating a model trained on one US adult EMR by applying it to a European adult EMR or US claims data) across populations.

It has been shown that validating a model is often a slow process [4] and when independent researchers implement the existing model, they may make mistakes when processing the data or implement the existing models incorrectly. A collaborative approach to external model validation has been proposed to overcome some of these issues [5]. The Observational Healthcare Data Science and Informatics (OHDSI) community is an open group of researchers aiming to develop tools and best practices for analyzing observational healthcare data. The OHDSI network consists of a large number of researchers with access to diverse datasets from across the world. The community have developed a homogeneous format known as the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) that any observational healthcare dataset can be mapped to. This standardization of the data can then be utilized to share analysis code directly between researchers, streamlining up the analysis process. Patient level prediction is one area of OHDSIs focus and standardized processes and tools that have been developed [6]. This presents the opportunity to validate models efficiently across the OHDSI network.

This study aims to demonstrate the power of using the OHDSI network to perform efficient external validation. This will be demonstrated by replicating five well known existing stroke models using the OHDSI patient level prediction framework and then getting the network of researchers to implement the models on their datasets over a 1 month period. The prediction question investigated is to predict stroke in a target population consisting of older female patients newly diagnosed with atrial fibrillation with no history of stroke. The specific target population was picked as it was the intersection of the different models’ development target populations. In this study five existing models; ATRIA, CHADS2, CHA2DS2-VASc, Framingham and Q-Stroke were replicated using the Patient Level Prediction framework. This study will investigate how well these five models perform when applied to datasets held by OHDSI network collaborators.

The discriminative performance of the models on the new datasets will be assessed in the OHDSI data network. This study will externally validate five existing stroke prediction models across the OHDSI collaborative network datasets.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author(s) | Comments |
| 0.2 | 30 April 2018 | Jenna Reps | Prep revisions |
| 0.1 | 4 April 2018 | Jenna Reps | Initial draft |
| 0.0 | 7 October 2015 | Martijn Schuemie | Template draft |

# Milestones

|  |  |
| --- | --- |
| Milestone | Planned / Estimated Date |
| Start of analysis | 01/05/2018 |
| End of analysis | 01/06/2018 |
| Posting of results | 10/06/2018 |
| Submission of manuscript | 20/06/2018 |

# Rationale and Background

Precision medicine presents the opportunity to tailor healthcare to the individual rather than reply on population averages. The ability to learn patient-level prediction models depends on the data available. Observational healthcare data, such as insurance claims or electronic medical records, are often big and may contain unknown medical relationships that can be utilized for patient-level prediction. One key advantage of developing models using this type of data is that the variables recorded in the data are readily available, so the prediction models can be widely implemented. A model needs to be tested on numerous different datasets for an accurate estimate of the performance across diverse datasets and to gain more insight into how well it is likely to perform in a clinical setting. This is known as external validation. External validation can give insight into how dependent the model is on certain variables and patient characteristics. For example, a model that was trained using data collected for patients with employment insurance may not generalize to the general population if the employed patients differ from unemployed patients in certain characteristics. The model needs to be externally validated on diverse patient populations to learn any transportability limitations before it is used widely in a clinical setting.

The external validation is a fundamental prerequisite before a model can be widely implemented but it is often the bottleneck step in model development. Studies have shown that only 16% of models are independently validated within 5 years and only 9% of models are validated by overlapping researchers [4]. It can be difficult for an independent researcher to replicate a model due to heterogeneous data structures and clinical coding often causing additional work to redefine variables in the new database. As some development details may be ambiguous or missing from the model documentation, independent researchers may not accurately replicate the model or may stop part way through replicating the model due to difficulties.

The Observational Healthcare Data Science and Informatics (OHDSI) collaboration is a network of researchers working towards a common goal of standardizations and best practice frameworks for analyzing observational data in healthcare. The OHDSI collaboration relies on researchers mapping their datasets into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and this provides a homogeneous data structure that analyses frameworks can be set up for. The OHDSI network currently contains longitudinal data on over 600 million patients observed for multiple years and comprising over 5 billion clinical observations. The data consists of numerous datasets owned by different collaborators around the world.

The Patient Level Prediction framework and software have been developed by the OHDSI community specifically for developing and validating patient-level prediction models following accepted best practices. Once advantage of the prediction framework and OHDSI standardizations is the ability to directly share models that can be applied to any data in the OMOP CDM. This means that one researcher in the community can develop a model and internally validate it, then she can share the code with the community and the model can be extensively validated across the network without the independent researcher having to do any additional coding. The ease of implementing other people’s models should speed up external validation. In addition, the OHDSI network contains data all over the world, so the transportability of models will be thoroughly investigated.

In this study we wish to demonstrate how the OHDSI standardizations and toolset can be used to revolutionize prediction model validation. As a proof of concept, we focused on validating a single prediction problem that is suitable for many existing models: predicting 1-year risk of stroke within a target population consisting of females aged 65-95 with newly diagnosed atrial fibrillation and no prior stroke. We replicated five existing stroke prediction models using the OHDSI Patient Level Prediction framework and will implement these models across the OHDSI collaborator network to externally validate the models and evaluate their transportability across the world. The specific target population was chosen as each existing model was developed for slightly different target populations, but older females with no prior stroke is suitable for all five models investigated in this study (i.e., we use the female Q-Stroke model and the no prior stroke Framingham model).

# Research Questions and Objectives

## Research Questions

We want to demonstrate the advantages of utilizing the OHDSI collaborative network for performing external model validation.

Primary objective

* To demonstrate the ability of utilizing the OHDSI collaborative network and Patient Level Prediction framework to perform extensive external validation of prediction models that can help give insight into the generalizability and transportability of each model

The study is completely implemented in an R package which will be distributed to partners that are willing to participate.

Secondary objective

* The OHDSI network and tool system will enable external validation to be performed efficiently (less than 1 month)

At present the majority of models are not validated within 5 years [4]

# Research methods

## Study Design

### Overview

To externally validate existing risk prediction models using the OHDSI collaborative research network. We chose the prediction problem of: within a target population cohort consisting of females aged 65-95 with atrial fibrillation and no prior stroke, predict the recording of stroke within 1 day to 365 days after atrial fibrillation.

Five existing models will be validated: ATRIA, CHADS2, CHA2DS2-VASc, Framingham (no prior stroke model) and Q-Stroke (female model). These models have all been developed to predict stroke for various target populations. We chose the target population of older females with atrial fibrillation and no history of stroke as we are implementing the female Q-stroke model and the Framingham model for people without prior stroke. The existing models will be added in the format of the Patient-Level-Prediction framework and will be readily implementable by researchers with the Patient-Level-Prediction R package. The existing evaluation framework will be utilized to evaluate the models.

As a measure of performance we will investigate the model’s ability to discriminate between those who will have stroke and those who will not. The existing models’ discriminations will be calculated using the area under the receiver operating characteristic curve (AUROC). We will not investigate calibration as prediction models are likely to require recalibration when applied to different datasets.

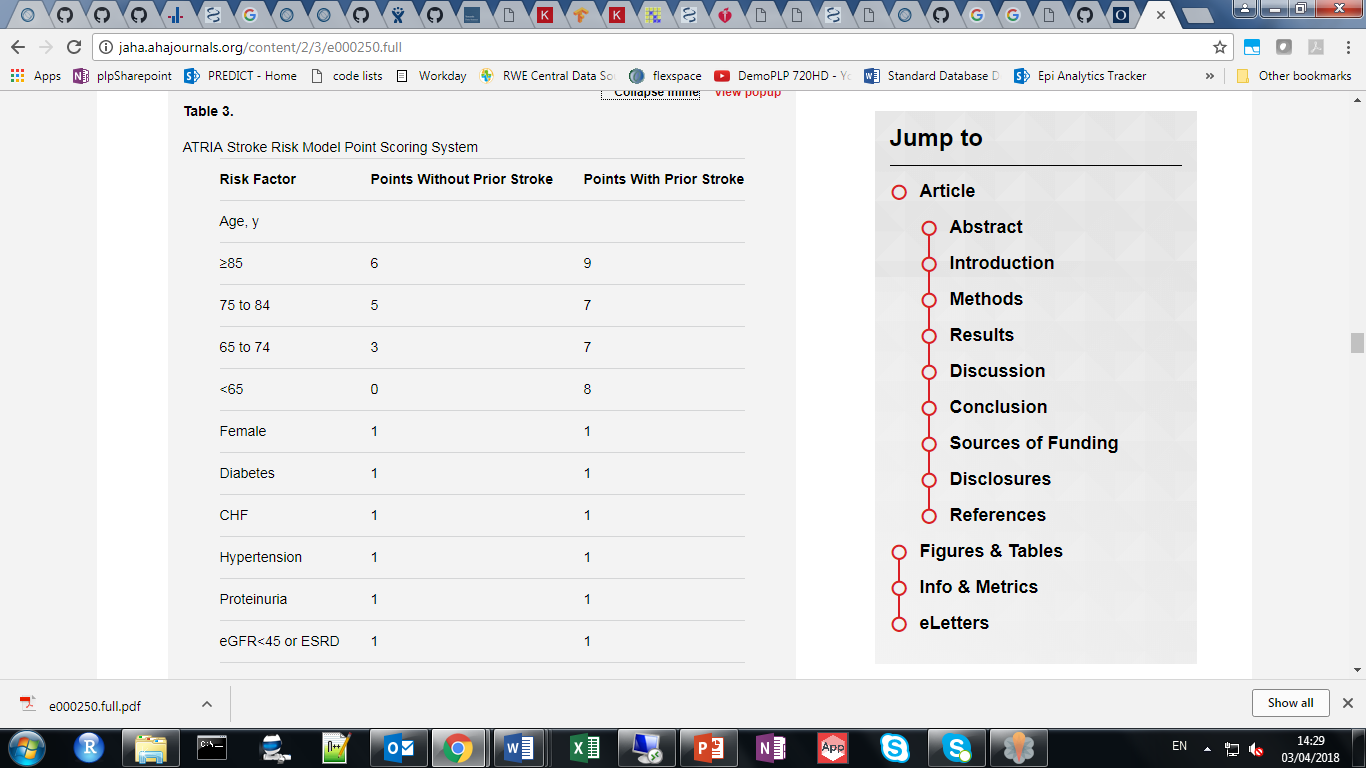
### Prediction Problem

The existing models will be evaluated by predicting stroke during 1 day to 365 days following the first record of atrial fibrillation in the database for females ages 65-95 with no prior stroke.

## Existing Models

The concept sets used to define each variable in the existing models can be found in the Appendix. Where possible we used medical definitions included in the paper with the model published.

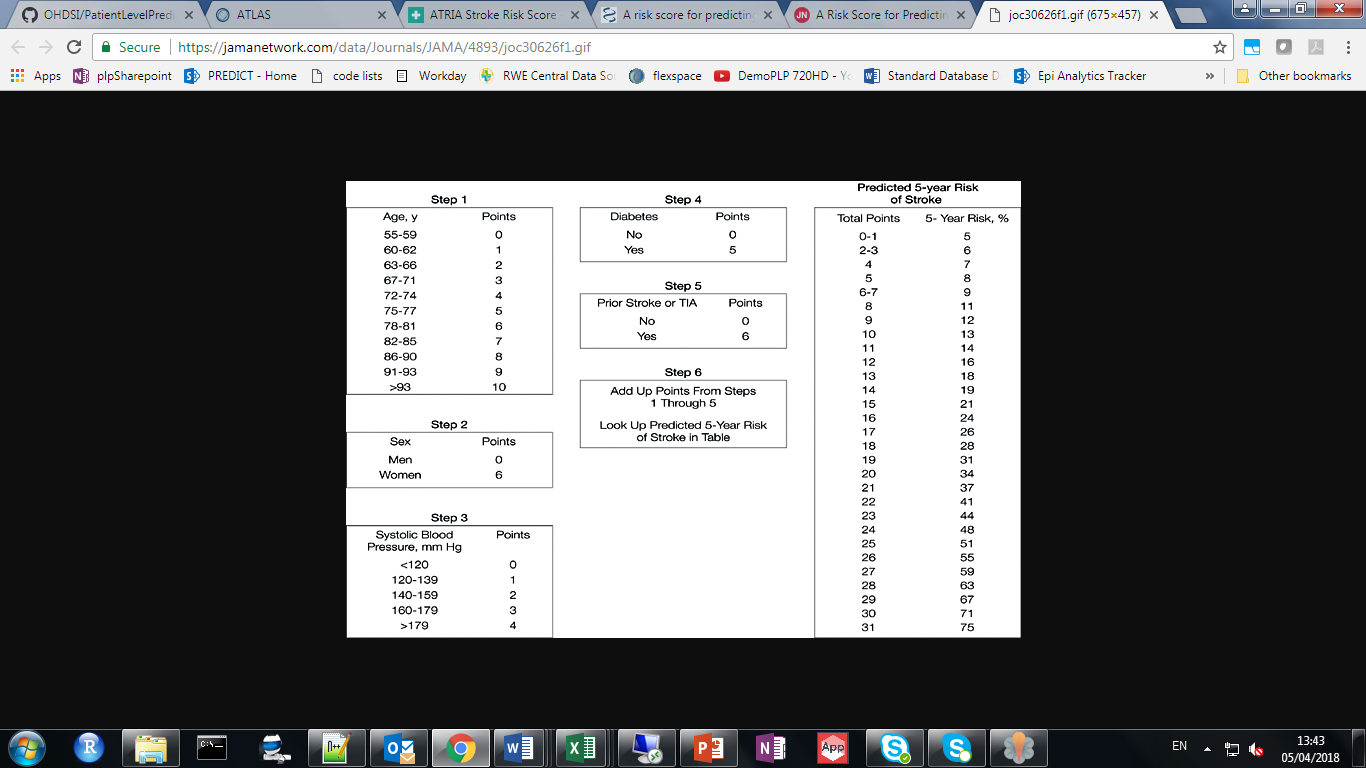
### ATRIA



The atria model obtained an internal validation AUROC of 0.72. When externally validated by the people who developed the model the AUROC was 0.7 (Singer 2013).

### Framingham

The Framingham risk score is (Wang 2003):

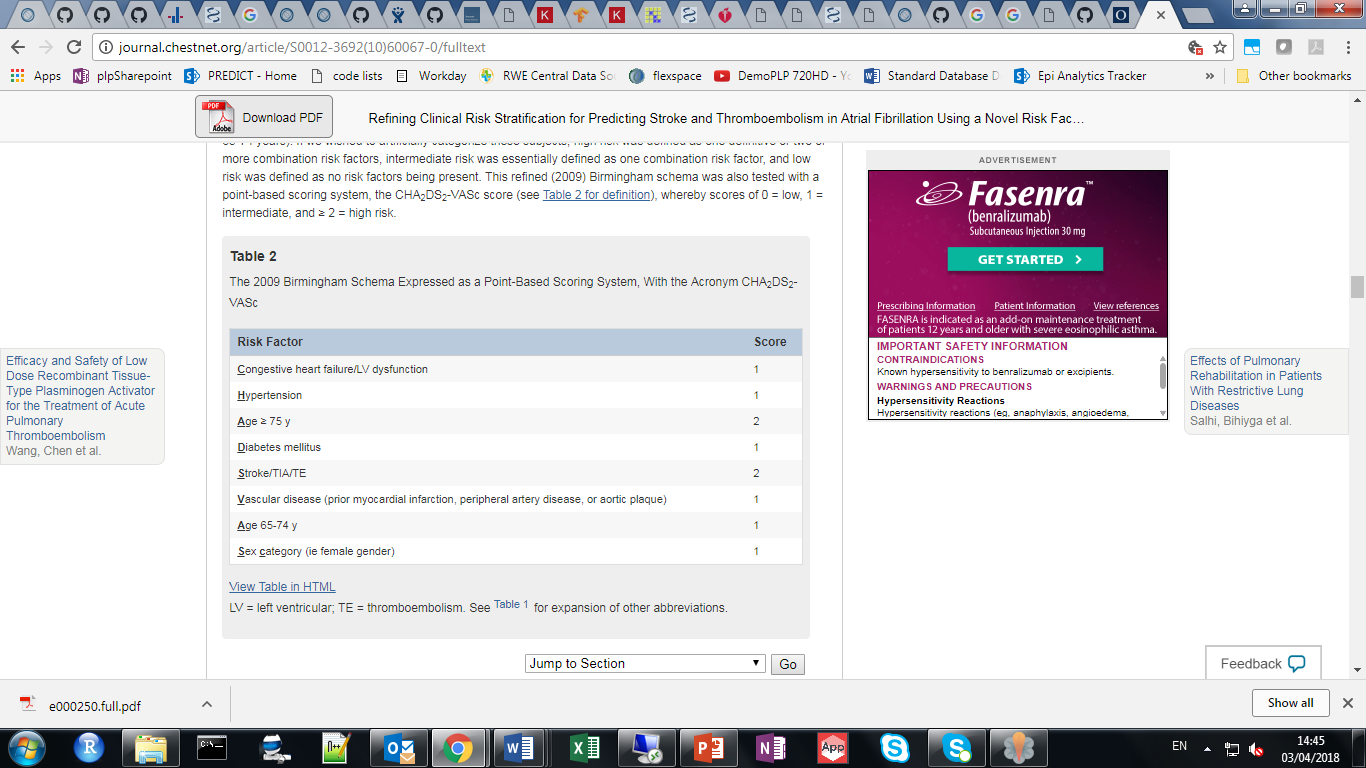


### CHADS2

|  |  |
| --- | --- |
| Risk factor | Points |
| congestive heart failure | 1 |
| hypertension | 1 |
| age 75 years or older | 1 |
| diabetes mellitus | 1 |
| history of stroke or transient ischemic attack | 2 |

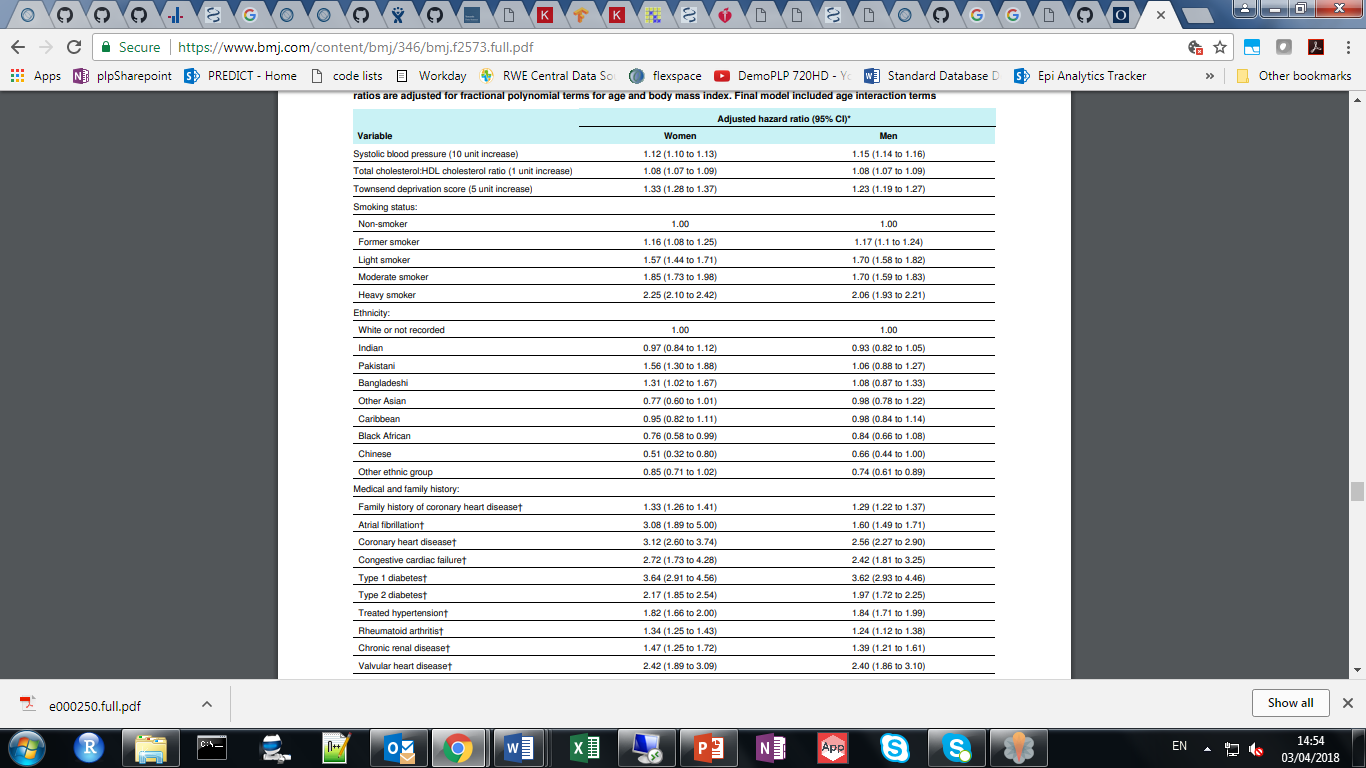
The CHADS2 was formed by combining various other predictive models and obtained an AUROC of 0.82 on the internal validation (Gage 2001).

### CHA2DS2-VASc



The CHA2DS2-VASc model obtained an internal validation AUROC of 0.61 (Lip 2010).

### Q-Stroke



The Q-Stroke model obtains an internal validation AUROC of 0.81 when applied to females aged 35-75 and 0.81 when applied to males aged 35-75. Increasing the age range increased performance (Hippisley-Cox 2013).

### External Validation of Models

The performance comparison between the existing models for some external datasets are presented below. In general the external validation of the models is approximately 0.65-0.70 for all five existing models.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Year | Population | Outcomes | ATRIA | CHADS2 | CHADS2VAS | Framingham | QStroke |
| UK EMR | 2015 <https://doi.org/10.1016/j.jacc.2015.08.033> | 60,594 | 2.99% per year | 0.7 (0.69-0.71) | 0.68 (0.67-0.69) | 0.68 (0.67-0.69) | - | - |
| Swedish EMR | 2016 <https://doi.org/10.1093/eurheartj/ehw077> | 152,153 | 11,053 (3.2% per year) | 0.71 (0.70-0.71) | 0.69 (0.69-0.70) | 0.69 (0.69-0.70) | - | - |
| Taiwan | 2016 <https://doi.org/10.1093/eurheartj/ehw077> | 186,570 | 23,723 | - | 0.66 | 0.70 | - | - |
| New Zealand, Russia and the Netherlands | 2014 <https://doi.org/10.1111/ijs.12411> | 9,501 | 752 | - | - | - | 0.70 (0.68-0.73) | 0.71 (0.69-0.73) |
| UK EMR | 2010 <https://doi.org/10.1111/j.1538-7836.2010.04085.x> | 79.844 |  | - | 0.66 (0.64–0.68) | 0.67 (0.65–0.69) | 0.65 (0.63–0.68) | - |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

## Cohorts

### Target Population

#### Atrial Fib – published def:

To qualify as an incident case, an AF diagnosis must have occurred between January 1, 2002, and June 30, 2007 (index date being the first date of diagnosis within this observation period) and met the following inclusion criteria: (1) an AF diagnosis (International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] code 427.31) on ≥2 outpatient visits on different dates, and/or (2) an AF diagnosis (ICD-9-CM code 427.31) on 1 outpatient visit, with the presence of electrocardiogram testing on the same date or within 30 days before the visit, and/or (3) an AF diagnosis for an emergency department visit or hospitalization Without the additional conditions but we also excluded general stroke <https://www.ajconline.org/article/S0002-91491301288-5/fulltext>

<https://epi.jnj.com/atlas/#/cohortdefinition/6623>

Atrial Fibrillation

As defined here: https://www.ajconline.org/article/S0002-91491301288-5/fulltext excluding the additional conditions but removing prior broad stroke

Initial Event Cohort

People having any of the following:

* a condition occurrence of Atrial fibrillation for plp stroke network study2
  + for the first time in the person's history
  + with age between 65 and 95 (inclusive)
  + gender is any of: FEMALE

with continuous observation of at least 730 days prior and 0 days after event index date, and limit initial events to: **earliest event per person.**

For people matching the Primary Events, include:

Having any of the following criteria:

* at least 1 occurrences of a condition occurrence of Atrial fibrillation for plp stroke network study2

starting between 1 days After and all days After event index date

* or at least 1 occurrences of a condition occurrence of Atrial fibrillation for plp stroke network study2
  + visit occurrence is any of: Emergency Room Visit, Emergency Room and Inpatient Visit, Inpatient Visit

starting between 0 days Before and 0 days After event index date

* or at least 1 occurrences of a measurement of electrocardiogram measurement plp stroke network study1

starting between 30 days Before and 0 days After event index date

Limit cohort of initial events to: **earliest event per person.**

Inclusion Rules

Inclusion Criteria #1: No prior stroke -broad definition

*Broad stroke definition: Ramalle-Gomara E, Ruiz E, Serrano M, Bartulos M, Gonzalez M-A, Matute B. Validity of Discharge Diagnoses in the Surveillance of Stroke. Neuroepidemiology. 2013;41: 185–188. doi:10.1159/000354626*

Having all of the following criteria:

* exactly 0 occurrences of a condition occurrence of Broad stroke: Ramalle-Gomara E, Ruiz E, Serrano M, Bartulos M, Gonzalez M-A, Matute B. Validity of Discharge Diagnoses in the Surveillance of Stroke. Neuroepidemiology. 2013;41: 185–188. doi:10.1159/0003546263

starting between all days Before and 0 days Before event index date

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.

Appendix 1: Concept Set Definitions

1. electrocardiogram measurement plp stroke network study

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 36302161 | Cardiac and vascular investigations (excl enzyme tests) | Measurement | MedDRA | YES | NO | NO |
| 36303102 | ECG investigations | Measurement | MedDRA | NO | YES | NO |

2. Atrial fibrillation for plp stroke network study

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 313217 | Atrial fibrillation | Condition | SNOMED | NO | YES | NO |
| 4108832 | Atrial fibrillation and flutter | Condition | SNOMED | YES | NO | NO |

3. Broad stroke: Ramalle-Gomara E, Ruiz E, Serrano M, Bartulos M, Gonzalez M-A, Matute B. Validity of Discharge Diagnoses in the Surveillance of Stroke. Neuroepidemiology. 2013;41: 185–188. doi:10.1159/000354626

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 374060 | Acute ill-defined cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 372924 | Cerebral artery occlusion | Condition | SNOMED | NO | YES | NO |
| 376713 | Cerebral hemorrhage | Condition | SNOMED | NO | YES | NO |
| 381591 | Cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 439847 | Intracranial hemorrhage | Condition | SNOMED | NO | YES | NO |
| 434056 | Late effects of cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 4162038 | Occlusion of artery | Condition | SNOMED | NO | YES | NO |
| 432923 | Subarachnoid hemorrhage | Condition | SNOMED | NO | YES | NO |
| 373503 | Transient cerebral ischemia | Condition | SNOMED | NO | YES | NO |

### Outcome

#### Stroke outcome 1

[OHDSI Europe tutorial] Graham replication: outcome cohort #1 - incident ischemic stroke, observed in inpatient setting

Ischemic stroke, as defined in Graham et al, Circulation, 2015: http://www.ncbi.nlm.nih.gov/pubmed/25359164

Initial Event Cohort

People having any of the following:

* a condition occurrence of Ischemic stroke1
  + condition type is any of: Inpatient detail - primary, Inpatient header - primary, Primary Condition, Inpatient detail - 1st position, Inpatient header - 1st position
  + visit occurrence is any of: Emergency Room Visit, Inpatient Visit

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event 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.

Appendix 1: Concept Set Definitions

1. Ischemic stroke

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 374060 | Acute ill-defined cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 4108356 | Cerebral infarction due to embolism of cerebral arteries | Condition | SNOMED | NO | YES | NO |
| 4110192 | Cerebral infarction due to thrombosis of cerebral arteries | Condition | SNOMED | NO | YES | NO |
| 4043731 | Infarction - precerebral | Condition | SNOMED | NO | YES | NO |

#### Stroke outcome 2 – any form of Cerebrovascular Disease

Stroke defined by ICD-9-CM 430-438, primary diagnosis obtained a sensitivity of 96% and positive predictive value of 83% [12-13]

Initial Event Cohort

People having any of the following:

* a condition occurrence of Broad stroke: Ramalle-Gomara E, Ruiz E, Serrano M, Bartulos M, Gonzalez M-A, Matute B. Validity of Discharge Diagnoses in the Surveillance of Stroke. Neuroepidemiology. 2013;41: 185–188. doi:10.1159/0003546261
  + for the first time in the person's history

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event 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.

Appendix 1: Concept Set Definitions

1. Broad stroke: Ramalle-Gomara E, Ruiz E, Serrano M, Bartulos M, Gonzalez M-A, Matute B. Validity of Discharge Diagnoses in the Surveillance of Stroke. Neuroepidemiology. 2013;41: 185–188. doi:10.1159/000354626

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 374060 | Acute ill-defined cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 372924 | Cerebral artery occlusion | Condition | SNOMED | NO | YES | NO |
| 376713 | Cerebral hemorrhage | Condition | SNOMED | NO | YES | NO |
| 381591 | Cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 439847 | Intracranial hemorrhage | Condition | SNOMED | NO | YES | NO |
| 434056 | Late effects of cerebrovascular disease | Condition | SNOMED | NO | YES | NO |
| 4162038 | Occlusion of artery | Condition | SNOMED | NO | YES | NO |
| 432923 | Subarachnoid hemorrhage | Condition | SNOMED | NO | YES | NO |
| 373503 | Transient cerebral ischemia | Condition | SNOMED | NO | YES | NO |

#### Stroke outcome 3 – Hemorrhagic Stroke

ICD-9-CM 430-432 – sensitivity 91% and positive predictive value 75% [12-13]

Initial Event Cohort

People having any of the following:

* a condition occurrence of hemorrhagic stroke plp network1
  + for the first time in the person's history

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event 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.

Appendix 1: Concept Set Definitions

1. hemorrhagic stroke plp network

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 376713 | Cerebral hemorrhage | Condition | SNOMED | NO | YES | NO |
| 432923 | Subarachnoid hemorrhage | Condition | SNOMED | NO | YES | NO |
| 439847 | Intracranial hemorrhage | Condition | SNOMED | NO | YES | NO |

#### Stroke outcome 4 – Ischemic Stroke

ICD-9-CM 433&434, primary diagnosis obtained a sensitivity of 84% and positive predictive value of 81% [12-13] but removing the primary diagnosis constraint lead to a sensitivity of 97% and positive predictive value of 88% in a different study [14].

Initial Event Cohort

People having any of the following:

* a condition occurrence of Ischaemic stroke plp network outcome1
  + for the first time in the person's history

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event 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.

Appendix 1: Concept Set Definitions

1. Ischaemic stroke plp network outcome

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 372924 | Cerebral artery occlusion | Condition | SNOMED | NO | YES | NO |
| 4162038 | Occlusion of artery | Condition | SNOMED | NO | YES | NO |

## Data Sources

The analyses will be performed across a network of observational healthcare databases. All databases have been transformed into the OMOP Common Data Model version 5. The complete specification for OMOP Common Data Model, version 5 is available at: <https://github.com/OHDSI/CommonDataModel>.

Prior to the study we will not know which collaborators will be running the analysis, so we are unable to specify the full datasets.

## Statistical Methods

The discriminative performance of each model will be calculated using the area under the receiver operating characteristic curve (AUROC). This corresponds to the probability that a random person who has the stroke during the 1 year follow-up is assigned a higher risk by the model than a random person who does not have the stroke during the 1 year follow-up. An AUROC value of 0.5 is equivalent to random guessing and an AUROC of 1 means perfect discrimination.

## Quality control

We will investigate the external validation of each risk model by

* Calculating the discrimination measures on various OMOP CDM datasets
* Determining the robustness of each model’s performance across the OHDSI datasets

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

## Strengths and Limitations of the Research Methods

Strength

* The analysis can help us gain insight into using the OHDSI network for efficient external validation.
* The analysis can help gain insight into the clinical usefulness of each stroke model by identifying whether it is transportable and/or reproducible.

Limitations

* The external validation datasets may not have a sufficient number of stroke outcomes to be used in the analysis
* The target and outcome cohort definitions used may need to be revised due to differences in the recording of data across the OHDSI network. For example, some databases do not have inpatient records specified. As a first step we will run the cohort creation across the network to check for issues that may require revising the cohorts from these original definitions.
* Although the CDM standardizes the vocabularies of the datasets, the concept recording distributions are likely to differ between databases and it is unknown how much this will limit model transportability.
* In addition, the external validation datasets may be missing important features that are used in the existing models, this may cause the model to perform poorly. It is worth highlighting that a model that doesn’t transport to diverse datasets may still be useful for predictions on populations matching the development dataset, so lack of transportability does not mean the model is bad. Qstroke contains many UK specific variables and may not transport due to this.
* It is sometimes difficult to correctly define the existing models’ variables how the authors intended.

# Protection of Human Subjects

The study is using only de-identified data. Confidentiality of patient records will be maintained at all times. All study reports will contain aggregate data only and will not identify individual patients or physicians.

# 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.

# References

[1] - Wilson, P.W., D’Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B., 1998. Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), pp.1837-1847.

[2] - [Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010 Feb;137(2):263-72. doi: 10.1378/chest.09-1584. Epub 2009 Sep 17. PubMed PMID: 19762550.](http://www.ncbi.nlm.nih.gov/pubmed/19762550)

[3] - Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ 1995;311:1539-41.

[4] - Siontis, G.C., Tzoulaki, I., Castaldi, P.J. and Ioannidis, J.P., 2015. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of clinical epidemiology*, *68*(1), pp.25-34.

[5] - Altman DG, Vergouwe Y, Royston P, Moons KGM: Prognosis and prognostic research: validating a prognostic model. BMJ. 2009, 338: b605-10.1136/bmj.b605. (message: need external validation to see whether overfit or optimistic in training set!)

[6] - Hripcsak, G., Ryan, P.B., Duke, J.D., Shah, N.H., Park, R.W., Huser, V., Suchard, M.A., Schuemie, M.J., DeFalco, F.J., Perotte, A. and Banda, J.M., 2016. Characterizing treatment pathways at scale using the OHDSI network.*Proceedings of the National Academy of Sciences*, p.201510502.

[7] - Singer, D.E., Chang, Y., Borowsky, L.H., Fang, M.C., Pomernacki, N.K., Udaltsova, N., Reynolds, K. and Go, A.S., 2013. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *Journal of the American Heart Association*, *2*(3), p.e000250

[8] - Wang, T.J., Massaro, J.M., Levy, D., Vasan, R.S., Wolf, P.A., D'agostino, R.B., Larson, M.G., Kannel, W.B. and Benjamin, E.J., 2003. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. *Jama*, *290*(8), pp.1049-1056.

[9] - Gage, B.F., Waterman, A.D., Shannon, W., Boechler, M., Rich, M.W. and Radford, M.J., 2001. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *Jama*, *285*(22), pp.2864-2870.

[10] - Lip, G.Y., Nieuwlaat, R., Pisters, R., Lane, D.A. and Crijns, H.J., 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, *137*(2), pp.263-272.

[11] - Hippisley-Cox, J., Coupland, C. and Brindle, P., 2013. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *Bmj*, *346*, p.f2573.

[12] McCormick, N., Bhole, V., Lacaille, D. and Avina-Zubieta, J.A., 2015. Validity of diagnostic codes for acute stroke in administrative databases: a systematic review. *PLoS One*, *10*(8), p.e0135834

[13] Ramalle-Gomara, E., Ruiz, E., Serrano, M., Bartulos, M., Gonzalez, M.A. and Matute, B., 2013. Validity of discharge diagnoses in the surveillance of stroke. *Neuroepidemiology*, *41*(3-4), pp.185-188.

[14] Hsieh C-Y, Chen C-H, Li C-Y, Lai M-L. Validating the diagnosis of acute ischemic stroke in a National Health Insurance claims database. J Formos Med Assoc Taiwan Yi Zhi. 2013; doi:10.1016/j.jfma.2013.09.009

# Appendix- Concept sets used in models

atria stroke model for PLP Network Study

Initial Event Cohort

People having any of the following: 

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event per person.**

Inclusion Rules

Inclusion Criteria #1: Diabetes +1

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Diabetes2

starting between all days Before and 0 days After event index date

Inclusion Criteria #2: CHF +1

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Congestive heart failure1

starting between all days Before and 0 days After event index date

Inclusion Criteria #3: Hypertension +1

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Hypertension4

starting between all days Before and 0 days After event index date

Inclusion Criteria #4: Proteinuria +1

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Proteinuria5

starting between all days Before and 0 days After event index date

Inclusion Criteria #5: End Stage Renal Disease +1

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of End stage renal disease3

starting between all days Before and 0 days After event index date

Inclusion Criteria #6: >=85 years +6

Having all of the following criteria:

Inclusion Criteria #7: 75-84 years +5

Having all of the following criteria:

Inclusion Criteria #8: 65-74 years +3

Having all of the following criteria:

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.

Appendix 1: Concept Set Definitions

1. Congestive heart failure

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 319835 | Congestive heart failure | Condition | SNOMED | NO | YES | NO |

2. Diabetes

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 201820 | Diabetes mellitus | Condition | SNOMED | NO | YES | NO |

3. End stage renal disease

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 193782 | End stage renal disease | Condition | SNOMED | NO | YES | NO |

4. Hypertension

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 320128 | Essential hypertension | Condition | SNOMED | NO | YES | NO |

5. Proteinuria

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 75650 | Proteinuria | Condition | SNOMED | NO | YES | NO |

Framingham stroke model for PLP Network Study

https://jamanetwork.com/journals/jama/fullarticle/197176

Initial Event Cohort

People having any of the following: 

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event per person.**

Inclusion Rules

Inclusion Criteria #1: Age 60-62 + 1

Having all of the following criteria:

Inclusion Criteria #2: Age 63-66 +2

Having all of the following criteria:

Inclusion Criteria #3: Age 67-71 +3

Having all of the following criteria:

Inclusion Criteria #4: Age 72-74 +4

Having all of the following criteria:

Inclusion Criteria #5: Age 75-77 +5

Having all of the following criteria:

Inclusion Criteria #6: Age 78-81 +6

Having all of the following criteria:

Inclusion Criteria #7: Age 82-85 +7

Having all of the following criteria:

Inclusion Criteria #8: Age 86-90 + 8

Having all of the following criteria:

Inclusion Criteria #9: Age 91-93 +9

Having all of the following criteria:

Inclusion Criteria #10: Age >93 +10

Having all of the following criteria:

Inclusion Criteria #11: Female + 6

Having all of the following criteria:

Inclusion Criteria #12: Diabetes +5

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Diabetes1

starting between all days Before and 0 days After event index date

Inclusion Criteria #13: Stroke or TIA + 6

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Stroke or TIA2

starting between all days Before and 0 days After event index date

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.

Appendix 1: Concept Set Definitions

1. Diabetes

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 201820 | Diabetes mellitus | Condition | SNOMED | NO | YES | NO |

2. Stroke or TIA

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 373503 | Transient cerebral ischemia | Condition | SNOMED | NO | YES | NO |
| 4043731 | Infarction - precerebral | Condition | SNOMED | NO | YES | NO |
| 4110192 | Cerebral infarction due to thrombosis of cerebral arteries | Condition | SNOMED | NO | YES | NO |
| 4108356 | Cerebral infarction due to embolism of cerebral arteries | Condition | SNOMED | NO | YES | NO |
| 374060 | Acute ill-defined cerebrovascular disease | Condition | SNOMED | NO | YES | NO |

qstroke stroke model for PLP Network Study

Didn't include measurements like Systolic blood pressure, total cholesterol hdl ratio or townsend deprivation score as not available

Initial Event Cohort

People having any of the following: 

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **earliest event per person.**

Inclusion Rules

Inclusion Criteria #1: Ethnicity Indian + log(0.97)

Having all of the following criteria:

Inclusion Criteria #2: Ethnicity Pakistani + log(1.56)

Having all of the following criteria:

Inclusion Criteria #3: Ethnicity Bangladeshi +log(1.31)

Having all of the following criteria:

Inclusion Criteria #4: Ethnicity Other Asian +log(0.77)

Having all of the following criteria:

Inclusion Criteria #5: Ethnicity Caribbean +log(0.95)

Having all of the following criteria:

Inclusion Criteria #6: Ethnicity Black African +log(0.76)

Having all of the following criteria:

Inclusion Criteria #7: Ethnicity Chinese +log(0.51)

Having all of the following criteria:

Inclusion Criteria #8: Ethnicity Other ethnic group +log(0.85)

Having all of the following criteria:

Inclusion Criteria #9: Family history of coronary heart disease + log(1.33)

Having all of the following criteria:

* at least 1 occurrences of an observation of Family history of CHD5

starting between all days Before and 0 days After event index date

Inclusion Criteria #10: Atrial fibrillation +log(3.08)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Atrial fibrillation1

starting between all days Before and 0 days After event index date

Inclusion Criteria #11: Coronary heart disease +log(3.12)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Coronary heart disease 4

starting between all days Before and 0 days After event index date

Inclusion Criteria #12: Congestive cardiac failure +log (2.72)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of congestive heart failure3

starting between all days Before and 0 days After event index date

Inclusion Criteria #13: Type 1 diabetes +log(3.64)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Type I diabetes9

starting between all days Before and 0 days After event index date

Inclusion Criteria #14: Type 2 diabetes +log(2.17)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Type II Diabetes10

starting between all days Before and 0 days After event index date

Inclusion Criteria #15: Treated hypertension +log(1.82)

Having all of the following criteria:

Inclusion Criteria #16: Rheumatoid arthritis +log(1.34)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Rheumatoid Arthritis8

starting between all days Before and 0 days After event index date

Inclusion Criteria #17: Chronic renal disease +log(1.47)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Chronic renal failure2

starting between all days Before and 0 days After event index date

Inclusion Criteria #18: Valvular heart disease +log(2.42)

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Heart Valve Disorder 7

starting between all days Before and 0 days After event index date

Inclusion Criteria #19: Smoking status: Former smoker +log(1.16)

*Can only tell former - don't know current status*

Having all of the following criteria:

* at least 1 occurrences of a condition occurrence of Former smoker6

starting between all days Before and all days After event index date

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.

Appendix 1: Concept Set Definitions

1. Atrial fibrillation

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 313217 | Atrial fibrillation | Condition | SNOMED | NO | YES | NO |

2. Chronic renal failure

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 46271022 | Chronic kidney disease | Condition | SNOMED | NO | YES | NO |

3. congestive heart failure

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 319835 | Congestive heart failure | Condition | SNOMED | NO | YES | NO |

4. Coronary heart disease

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 4187067 | Disorder of coronary artery | Condition | SNOMED | NO | YES | NO |

5. Family history of CHD

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 44789401 | Family history of coronary heart disease | Observation | SNOMED | NO | YES | NO |

6. Former smoker

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 437264 | Tobacco dependence syndrome | Condition | SNOMED | NO | YES | NO |

7. Heart Valve Disorder

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 4281749 | Heart valve disorder | Condition | SNOMED | NO | YES | NO |

8. Rheumatoid Arthritis

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 80809 | Rheumatoid arthritis | Condition | SNOMED | NO | YES | NO |

9. Type I diabetes

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 201254 | Type 1 diabetes mellitus | Condition | SNOMED | NO | YES | NO |

10. Type II Diabetes

| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| --- | --- | --- | --- | --- | --- | --- |
| 201826 | Type 2 diabetes mellitus | Condition | SNOMED | NO | YES | NO |