**Study Protocol**

**Development and validation of patient-level prediction models for adverse health outcomes amongst adult RA patients initiating first-line treatment of methotrexate monotherapy: a multinational real-world cohort analysis**.

**Created by:**

Cynthia Yang, Erasmus MC, Rotterdam, The Netherlands

Ross D. Williams, Erasmus MC, Rotterdam, The Netherlands

Joel Swerdel, PhD, Janssen Research and Development

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

Patrick Ryan, PhD, Janssen Research and Development

**Prepared on:** 16 January 2020

**Acknowledgement:** The analysis is performed in the context of the European Health Data and Evidence Network (EHDEN) project ([www.ehden.eu](http://www.ehden.eu)) in close collaboration with the Observational Health Sciences and Informatics collaborative (OHDSI, [http://ohdsi.org](http://ohdsi.org/)).

The authors declare the following disclosures: Joel Swerdel, PhD and Patrick Ryan, PhD are employees of Janssen Research & Development.

**Table of Contents**

[1. List of Abbreviations 3](#_Toc31012612)

[2. Executive Summary 3](#_Toc31012613)

[3. Rationale & Background 4](#_Toc31012614)

[4. Objective 5](#_Toc31012615)

[5. Methods 5](#_Toc31012616)

[5.1. Study Design 5](#_Toc31012617)

[5.2. Data Source(s) 6](#_Toc31012618)

[5.3. Study Populations 11](#_Toc31012619)

[6.3. Statistical Analysis Method(s) 16](#_Toc31012620)

[6.4. Quality Control 17](#_Toc31012621)

[6.5. Tools 17](#_Toc31012622)

[7. Diagnostics 17](#_Toc31012623)

[8. Data Analysis Plan 18](#_Toc31012624)

[8.1. Algorithm Settings 18](#_Toc31012625)

[8.2. Covariate Settings 18](#_Toc31012626)

[8.3. Model Development & Evaluation 23](#_Toc31012627)

[8.4. Analysis Execution Settings 23](#_Toc31012628)

[9. Strengths & Limitations 23](#_Toc31012629)

[10. Protection of Human Subjects 23](#_Toc31012630)

[11. Plans for Disseminating & Communicating Study Results 24](#_Toc31012631)

[12. Tables & Figures 24](#_Toc31012632)

[12.1. Incidence Rate of Target & Outcome 24](#_Toc31012633)

[13. Appendices 25](#_Toc31012634)

[13.1. Study Generation Version Information 25](#_Toc31012635)

[14. References 25](#_Toc31012636)

# List of Abbreviations

|  |  |
| --- | --- |
| Abbreviation | Phrase |
| AUROC | Area Under the Receiver Operating Characteristic Curve |
| CDM | Common Data Model |
| csDMARD | Conventional synthetic disease-modifying anti-rheumatic drug |
| CVD | Cardiovascular disease |
| DMARD | Disease-modifying anti-rheumatic drugs |
| EULAR | European League Against Rheumatism |
| MI | Myocardial infarction |
| MTX | Methotrexate |
| O | Outcome Cohort |
| OHDSI | Observational Health Data Sciences & Informatics |
| OMOP | Observational Medical Outcomes Partnership |
| RA | Rheumatoid arthritis |
| T | Target Cohort |
| TAR | Time at Risk |

# Executive Summary

European League Against Rheumatism (EULAR) guidelines recommend the early initiation of methotrexate (MTX) monotherapy as soon as possible after the diagnosis of rheumatoid arthritis (RA). However, many adverse health outcomes may occur. Evaluating the patient-level risk for adverse health outcomes would allow clinicians to provide more personalised care.

The objective of this study is to  develop and validate patient-level prediction models for adverse health outcomes including leukopenia, pancytopenia, infection (serious, opportunistic, all), cardiovascular disease (CVD) (myocardial infarction (MI), stroke), and cancer (breast, colorectal, uterus) in adult RA patients initiating first-line treatment of MTX monotherapy.

Health data from claims and electronic health records will be used including patients from 7 European countries (Spain, Estonia, Netherlands, Belgium, Germany, France, and the UK), the United States of America, Australia, and Japan. All adult RA patients initiating first-line treatment of MTX monotherapy with at least one year of prior observation will be included. Prediction models for the outcomes will be developed for a time at risk of 90 days (infections, leukopenia, pancytopenia), 2 years (MI and stroke), and 5 years (cancers) on the Optum© De-Identified Clinformatics® Data Mart Database. Models will be developed using LASSO logistic regression, evaluated using the area under the receiver operator characteristic curve (AUROC) for discrimination and graphically assessed for calibration. The models will be externally validated on all other databases.

# Rationale & Background

Rheumatoid Arthritis (RA) is a common musculoskeletal disease, affecting approximately 0.5-1.0% of the adult population in Europe and North America. The management for the condition has changed considerably over the last 35 years, with a number of therapeutic options available including short and long-term disease-modifying anti-rheumatic drugs (DMARDs). Conventional synthetic disease-modifying anti-rheumatic drugs (csDMARDs) are usually the first line of treatment in newly diagnosed RA. Among the csDMARDs, methotrexate is currently adopted as the “anchor drug” (1) either for use as monotherapy or in combination with another csDMARDs such as hydroxychloroquine, leflunomide or sulfasalazine. Guidance for use for all DMARDs has evolved with time and also differ internationally between North America and Europe (2-4).

RA has been associated with adverse health outcomes which may or may not be related to existing treatments (5). Recent drug safety studies have focused on the risk of adverse health outcomes associated with biologic DMARDs, such as infection, cancer and cardiovascular outcomes (6-8). However, the risk of adverse health outcomes in csDMARDs continues to be conflicting. For instance, patients on methotrexate are frequently counselled regarding an increased risk of infection, while there is little good quality evidence quantifying this risk in the literature and with several studies suggesting no increased infection risk (9, 10). Since the EULAR guidelines recommend the early initiation of methotrexate monotherapy as soon as possible after the diagnosis of RA, there is need for predicting the patient-level risk of adverse health outcomes for patients initiating first-line treatment of methotrexate monotherapy. Adverse health outcomes that may occur are cytopenia, cancer, infection and cardiovascular disease. Evaluating the patient-level risk for these outcomes would allow clinicians to provide more personalized care than is possible currently. Patients could for example be targeted for recommending attendance of breast cancer screenings, adjusting the frequency of blood testing measures or developing vaccination plans to manage the risk of infections.

Previous literature addressing this issue is limited. To the best of our knowledge, there is no existing literature with respect to prediction of an individual patient’s risk for cytopenia, cancer or infections upon initiating first-line treatment with csDMARDs. For cardiovascular outcomes, the evidence of increased risk in RA patients is well-documented. Cardiovascular risk prediction tools previously developed for the general population, such as the Framingham Risk Score and the QRisk Score, generally fail to accurately predict cardiovascular risk in RA populations (11-13). For this reason, several risk prediction models have been recommended for use in RA patients. However, external validations show that these models do not predict cardiovascular risk in patients with RA more accurately than the ones developed for the general population. Therefore, more research is needed to build an accurate cardiovascular risk calculator for RA patients.

In this study we aim to predict the risk of adverse health outcomes for adult RA patients initiating treatment of methotrexate monotherapy. Outcomes of interest are myocardial infarction, stroke, infection (serious, opportunistic, all), leukopenia, pancytopenia and colorectal cancer among all patients, and breast cancer and uterus cancer among female patients. We will develop patient-level prediction models using the Observational Health Data Sciences and Informatics (OHDSI) Patient-Level Prediction framework. The 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). This along with the Patient-Level Prediction framework allows for rapid model development and validation following accepted best practices (14). We will implement the patient-level prediction models for the different outcomes in RA patients initiating first-line treatment of methotrexate monotherapy across the OHDSI collaborator network to externally validate the models and evaluate their transportability across the world.

# Objective

The objective is to develop and validate patient-level prediction models for multiple outcomes amongst adult RA patients initiating first-line treatment of methotrexate (MTX) monotherapy:

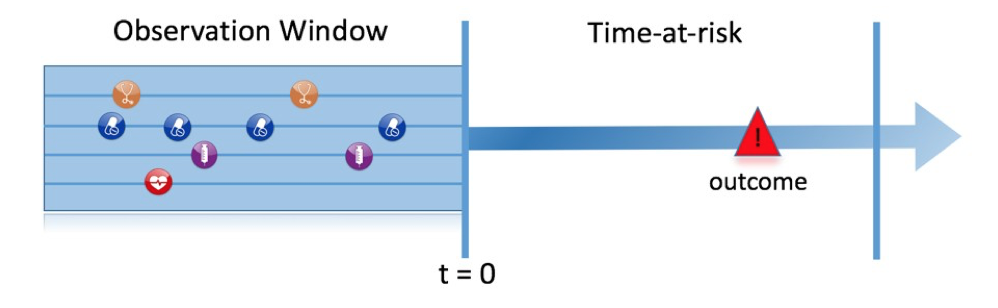
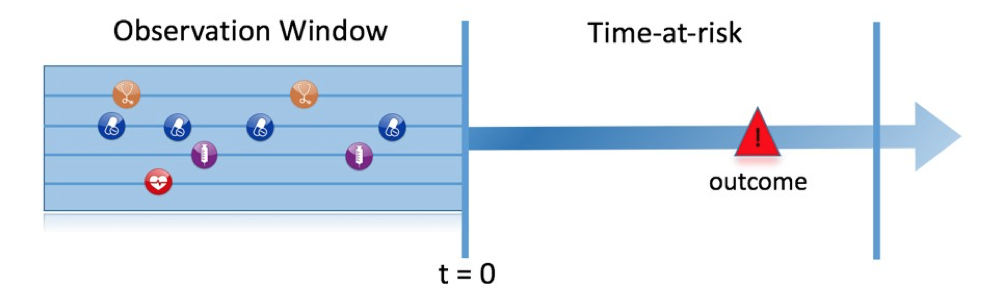
1. To predict the 2-year risk of myocardial infarction amongst adult RA patients initiating first-line treatment with MTX monotherapy
2. To predict the 2-year risk of stroke amongst adult RA patients initiating first-line treatment with MTX monotherapy
3. To predict the 90-day risk of serious infection, opportunistic infection and all infection amongst adult RA patients initiating MTX monotherapy
4. To predict the 5-year risk of colorectal cancer amongst adult RA patients initiating first-line treatment with MTX monotherapy
5. To predict the 5 year risk of breast cancer and uterus cancer amongst adult female RA patients initiating first-line treatment with MTX monotherapy
6. To predict the 90-day risk of leukopenia and pancytopenia amongst adult RA patients initiating first-line treatment with MTX monotherapy

# 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 modelling 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 (TAR). Prediction is done using only information about the patients in an observation window prior to that moment in time.



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

We follow the PROGRESS best practice recommendations for model development and the TRIPOD guidance for transparent reporting of the model results (15, 16).

## Data Source(s)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Source Full Name** | **Country Code** | **Data Provenance** | **Source Short Name** | **Patient Count** | **History** | **Patient Type** | **Data collection** |
| IQVIA US Ambulatory EMR | US | EMR | IQVIA Amb - EMR, US | 49m | 2006 – | Outpatient / General population | Dataset consists of longitudinal, de-identified ambulatory EHR data |
| IQVIA Australia EMR | AU | EMR | IQVIA IMS - EMR, AU | 6m | 2006 – | Outpatient / General population | Anonymized patient records collected from Patient Management software used by GPs during an office visit to document patients’ clinical records |
| IQVIA LPD Belgium EMR | BE | EMR | IQVIA LPD - EMR, BE | 2m | 2005 – | Outpatient / General population | Medical contacts and diagnoses, test results and drugs associated with them. Only outpatient |
| IQVIA LPD France | FR | EMR | IQVIA LPD - EMR, FR | 7.8m | 1994- | Outpatient / General population | Anonymized patient records collected from Patient Management software used by GPs and select specialists during an office visit to document patients’ clinical records |
| IQVIA Disease Analyser France EMR | FR | EMR | IQVIA DA - EMR, FR | 10m | 1997 – | Outpatient / General population Patients seen in the primary care setting | Anonymized patient records collected from Patient Management software used by GPs during an office visit to document patients’ clinical records |
| IQVIA Disease Analyser Germany EMR | DE | EMR | IQVIA DA - EMR, DE | 37m | 1992 – | Outpatient / General population Public and private insurance | Anonymized patient records collected from Patient Management software used by GPs and selected specialists to document patients’ medical records within their office-based practice during a visit |
| IQVIA Hospital US Charge Master | US | Hospital | IQVIA Hospital, US | 86m | 2007 – | Inpatient & outpatient hospital encounters, including Emergency Room visits / General population | 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 UK THIN IMRD EMR | UK | EMR | IQVIA THIN - EMR, UK | 15m | 1989 – | General population / Primary care records with hospitalization / referral information | Pseudonymized Electronic Medical Records collected from Patient Management software used within UK Primary Care |
| Estonian Health Information System | EE | EMR | EHIS - EMR, EE | 1.4m | 2012-2016 | All inpatient and outpatient discharge summaries, general population | Pseudonymized patient level health records from central e-health database where submitting the records is mandatory for all healthcare service providers in Estonia |
| Integrated Primary Care Information | NL | EMR | IPCI - EMR, NL | 2.5m | 1996 | Patients seen in Primary Care setting | The Integrated Primary Care Information (IPCI) database is a Dutch database containing the complete medical record of more than 2.5 million patients provided by more than 450 GPs geographically spread over the Netherlands. (12) In the Netherlands, all citizens are registered with a GP practice which acts as a gatekeeper in a two-way exchange of information with secondary care. |
| SIDIAP | ES | EMR | SIDIAP - EMR, ES | 5.8m | 2006 | Primary care linked (partially) to inpatient data | The Information System for Research in Primary Care (SIDIAP; www.sidiap.org) is a primary care records database that covers >80% of the population of Catalonia, North-East Spain. Healthcare is universal and tax-payer funded in the region, and primary care physicians are gatekeepers for all care and responsible for repeat prescriptions. |
| Japan Medical Data  Center (JMDC) | JP | Claims | JMDC - Claims, JP | 7.2m | 2005- | Claims / Society-Managed Health Insurance | Japan Medical Data Center (JDMC) database consists of data from 60 Society-Managed  Health Insurance plans covering workers aged 18 to 65 and their dependents  (family member with no income or small income, e.g. children or house  wives, can join your insurance as family member with no extra fee). JMDC  data includes membership status of the insured people and claims data  provided by insurers under contract (e.g. patient-level  demographic information, inpatient and outpatient data inclusive of diagnosis and procedures,  and prescriptions as dispensed claims information). Claims data are derived from monthly  claims issued by clinics, hospitals and community pharmacies; for claims only the month and year are  provided however prescriptions, procedures, admission, discharge, and start of medical  care as associated with a full date.    All diagnoses are coded using ICD-10. All prescriptions refer to national Japanese  drug codes, which have been linked to ATC. Procedures are encoded using local  procedure codes, which the vendor has mapped to ICD-9 procedure codes. The  annual health checkups report a standard battery of measurements (e.g. BMI),  which are not coded but clearly described. |
| Optum® De-Identified  Clinformatics® Data Mart  Database – Socio-Economic Status (SES) | US | Claims | Optum SES - Claims, US | 88m | 2001- | Claims / Privately Insured | Optum© De-Identified Clinformatics® Data Mart Database (OptumInsight, Eden Prairie, MN) 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 at the US Census Division level.    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). |
| Optum® de-identified Electronic  Health Record Dataset | US | EMR | Optum EHR - EMR, US | 96m | 2006- | EHR / Privately Insured | Optum© de-identified Electronic Health Record Dataset represents Humedica’s Electronic Health Record data a medical records database. The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using Natural Language Processing (NLP). |
| IBM MarketScan® Commercial Database | US | Claims | IBM CCAE - Claims, US | 150m | 2000- | Claims / Privately Insured | IBM MarketScan® Commercial 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). |
| IBM MarketScan® Multi-State  Medicaid Database | US | Claims | IBM MDCD - Claims, US | 28.5m | 2006- | Claims / Medicaid | IBM 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. |
| IBM MarketScan® Medicare  Supplemental Database | US | Claims | IBM MDCR - Claims, US | 10m | 2000- | Claims / Retiree Supplemental | IBM MarketScan® Medicare Supplemental 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). |

## Study Populations

### Target Cohort(s) [T]

|  |  |  |
| --- | --- | --- |
| Cohort ID | Cohort Name | Description |
| 257 | [EHDEN RA] New users of methotrexate monotherapy used for PLP | **Index event:** have a drug utilisation record of MTX as well |
| 258 | [EHDEN RA] Female new users of methotrexate monotherapy used for PLP | Identical to the new users except limited to female patients |

[EHDEN RA] New users of methotrexate monotherapy used for PLP

The full specification can be found at: <http://atlas-demo.ohdsi.org/#/cohortdefinition/1773112>

Initial Event Cohort

People having any of the following:

* a drug exposure of [EHDEN RA] Methotrexate9
  + for the first time in the person's history

with continuous observation of at least 365 days prior and 0 days after event index date, and limit initial events to: **earliest event per person.**

Inclusion Rules

Inclusion Criteria #1: age >= 18

Having all of the following criteria:

* with the following event criteria:
  + with age >= 18

Inclusion Criteria #2: has RA in 5 year prior

Having any of the following criteria:

* at least 1 occurrences of a condition occurrence of [EHDEN RA] Rheumatoid arthritis with history10
* where event starts between 1826 days Before and 0 days After index start date
* or at least 1 occurrences of an observation of [EHDEN RA] Rheumatoid arthritis with history10
* where event starts between 1826 days Before and 0 days After index start date

Inclusion Criteria #3: does not have any DMARD any time prior

Having all of the following criteria:

* exactly 0 occurrences of a drug exposure of [EHDEN RA] DMARDs broad list4
* where event starts between all days Before and 1 days Before index start date

Inclusion Criteria #4: exactly 1 DMARD on or 7d after

Having all of the following criteria:

* exactly 1 distinct occurrences of a drug exposure of [EHDEN RA] DMARDs in ACR guideline5
* where event starts between 0 days Before and 7 days After index start date

Inclusion Criteria #5: no prior cancer and no history of cancer observation

Having all of the following criteria:

* exactly 0 occurrences of a condition occurrence of [EHDEN RA] Malignant neoplasms excluding non-melanoma skin cancer8
* where event starts between all days Before and 0 days After index start date
* and exactly 0 occurrences of an observation of [EHDEN RA] History of cancer observations6
* where event starts between all days Before and 0 days After index start date

Inclusion Criteria #6: no prior inflammatory arthropathy

Having all of the following criteria:

* exactly 0 occurrences of a condition occurrence of [EHDEN RA] Inflammatory arthropathies7
* where event starts between all days Before and 0 days After index start date

Limit qualifying cohort to **earliest event per person.**

[EHDEN RA] Female new users of methotrexate monotherapy used for PLP

The full specification can be found at:

<http://atlas-demo.ohdsi.org/#/cohortdefinition/1773113>

This is identical to the first target cohort except for the initial inclusion criteria of:

Initial Event Cohort

People having any of the following:

1. a drug exposure of [EHDEN RA] Methotrexate9
   1. for the first time in the person's history
   2. gender is any of: FEMALE

### Outcome Cohorts(s) [O]

|  |  |
| --- | --- |
| Cohort Name | Description |
| [EHDEN RA] Stroke (ischemic or haemorrhagic) events (any visit) | A condition occurrence of stroke |
| [EHDEN RA] Acute myocardial infarction events (in any visit) | A condition occurrence of myocardial infarction |
| [EHDEN RA] Pancytopenia events using diagnoses and measurements | A condition occurrence of Pancytopenia. We include patients with relevant diagnosis and measurements. |
| [EHDEN RA] Opportunistic Infections | A condition occurrence of an opportunistic infection. |
| [EHDEN RA] Serious Infection events | A condition occurrence of a serious infection.The condition occurrence of a serious infection occurring on a day where there was a visit occurrence for inpatient hospitalisation or emergency room presentation also recorded, or where death occurred up to 30 days following the condition occurrence of a serious infection. |
| [EHDEN RA] Persons with a Malignant neoplasm of breast 1 dx | A condition occurrence of breast cancer |
| [EHDEN RA] Persons with a Malignant neoplasm of uterus 1 dx | A condition occurrence of uterine cancer |
| [EHDEN RA] Persons with a Malignant neoplasm of colon and rectum 1 dx | A condition occurrence of colorectal cancer |
| [EHDEN RA] Serious Infection, opportunistic infections and other infections of interest event | A serious infection (as defined above), an opportunistic infection (as defined above), or a condition occurrence of a other infection of interest. |
| [EHDEN RA] Leukopenia events using diagnoses and measurements | A condition occurrence of Leukopenia. |
| [EHDEN RA] Pancytopenia or leukopenia events using diagnoses and measurements | A condition occurrence of either Pancytopenia or Leukopenia as defined previously. |

**Full descriptions:**

The JSON files describing for all the outcome cohorts are available at : <https://github.com/ohdsi-studies/EhdenRaPrediction/tree/master/inst/cohorts>

In order to convert these to a human readable form, import the json into a new cohort definition in any instance of ATLAS and reload.

### Time at Risk

The table below describes the Time at Risk (TAR) window start and end for each of the analyses that are executed.

|  |
| --- |
| Time at Risk |
| [Time at Risk Settings #1] Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 730, 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 |
| [Time at Risk Settings #3] Risk Window Start: 365, Add Exposure Days to Start: FALSE, Risk Window End: 1826, Add Exposure Days to End: FALSE |

### Additional Population Settings

The final study population in which we will develop our model is a subset of the target cohort, because we may for example apply criteria that are dependent on the outcome, or we want to perform sensitivity analyses with sub-populations of the target cohort. For this we have to answer the following questions:

* ***What is the minimum amount of observation time we require before the start of the target cohort?***This choice could depend on the available patient time in the training data, but also on the time we expect to be available in the data sources we want to apply the model on in the future. The longer the minimum observation time, the more baseline history time is available for each person to use for feature extraction, but the fewer patients will qualify for analysis. Moreover, there could be clinical reasons to choose a short or longer look-back period.
* ***Can patients enter the target cohort multiple times?*** In the target cohort definition, a person may qualify for the cohort multiple times during different spans of time, for example if they had different episodes of a disease or separate periods of exposure to a medical product. The cohort definition does not necessarily apply a restriction to only let the patients enter once, but in the context of a particular patient-level prediction problem we may want to restrict the cohort to the first qualifying episode. In our example, a person can only enter the target cohort once since our criteria was based on first use of an ACE inhibitor.
* ***Do we allow persons to enter the cohort if they experienced the outcome before?*** Do we allow persons to enter the target cohort if they experienced the outcome before qualifying for the target cohort? Depending on the particular patient-level prediction problem, there may be a desire to predict incident first occurrence of an outcome, in which case patients who have previously experienced the outcome are not at risk for having a first occurrence and therefore should be excluded from the target cohort. In other circumstances, there may be a desire to predict prevalent episodes, whereby patients with prior outcomes can be included in the analysis and the prior outcome itself can be a predictor of future outcomes.
* ***How do we define the period in which we will predict our outcome relative to the target cohort start?***We have to make two decisions to answer this question. First, does the time-at-risk window start at the date of the start of the target cohort or later? Arguments to make it start later could be that we want to avoid outcomes that were entered late in the record that actually occurred before the start of the target cohort or we want to leave a gap where interventions to prevent the outcome could theoretically be implemented. Second, we need to define the time-at-risk by setting the risk window end, as some specification of days offset relative to the target cohort start or end dates.
* ***Do we require a minimum amount of time-at-risk?*** We have to decide if we want to include patients that did not experience the outcome but did leave the database earlier than the end of our time-at-risk period. These patients may experience the outcome when we no longer observe them. For our prediction problem we decide to answer this question with “yes,” requiring a minimum time-at-risk for that reason. Furthermore, we have to decide if this constraint also applies to persons who experienced the outcome, or we will include all persons with the outcome irrespective of their total time at risk.

In our study three population settings are defined as described below:

***Population Settings #1***

|  |  |
| --- | --- |
| Item | Settings |
| minTimeAtRisk | 1 |
| requireTimeAtRisk | TRUE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 1 |
| washoutPeriod | 365 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 90 |
| binary | TRUE |
| removeSubjectsWithPriorOutcome | TRUE |
| riskWindowEnd | 730 |
| firstExposureOnly | TRUE |

***Population Settings #2***

|  |  |
| --- | --- |
| Item | Settings |
| minTimeAtRisk | 1 |
| requireTimeAtRisk | TRUE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 1 |
| washoutPeriod | 365 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 90 |
| binary | TRUE |
| removeSubjectsWithPriorOutcome | TRUE |
| riskWindowEnd | 90 |
| firstExposureOnly | TRUE |

***Population Settings #3***

|  |  |
| --- | --- |
| Item | Settings |
| minTimeAtRisk | 1 |
| requireTimeAtRisk | TRUE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 365 |
| washoutPeriod | 365 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 9999 |
| binary | TRUE |
| removeSubjectsWithPriorOutcome | TRUE |
| riskWindowEnd | 1826 |
| firstExposureOnly | TRUE |

## Statistical Analysis Method(s)

### Algorithms

In this study we will apply a 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 quantiles 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 trade-off 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 colour 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. More information can be found in the Book of OHDSI at: <https://ohdsi.github.io/TheBookOfOhdsi/SoftwareValidity.html>

## Tools

To create the study package, ATLAS will be used to specify the cohorts, time-at-risk, covariate and population settings as well as which models will be analysed. Information on this is available in the Book of OHDSI at: <https://ohdsi.github.io/TheBookOfOhdsi/OhdsiAnalyticsTools.html#atlas>

The package developed in ATLAS will utilise the Patient-Level Prediction R package to run the analysis. More information on this is available at: <https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html>

This study will be designed using OHDSI tools and run with R (17). 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.

# 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 | TRUE |
| ConditionEraStartLongTerm | FALSE |
| temporal | FALSE |
| DemographicsIndexMonth | TRUE |
| ConditionOccurrencePrimaryInpatientLongTerm | FALSE |
| ConditionEraAnyTimePrior | FALSE |
| addDescendantsToInclude | FALSE |
| ConditionGroupEraStartMediumTerm | FALSE |
| ProcedureOccurrenceLongTerm | FALSE |
| DrugExposureLongTerm | FALSE |
| DrugEraStartShortTerm | FALSE |
| DistinctIngredientCountMediumTerm | FALSE |
| DistinctMeasurementCountShortTerm | FALSE |
| MeasurementRangeGroupLongTerm | FALSE |
| ConditionGroupEraOverlapping | FALSE |
| MeasurementRangeGroupMediumTerm | FALSE |
| DrugGroupEraStartMediumTerm | FALSE |
| MeasurementAnyTimePrior | FALSE |
| MeasurementMediumTerm | FALSE |
| includedCovariateIds |  |
| ConditionOccurrenceAnyTimePrior | TRUE |
| DistinctConditionCountLongTerm | FALSE |
| MeasurementValueLongTerm | FALSE |
| DrugEraShortTerm | FALSE |
| DrugGroupEraAnyTimePrior | FALSE |
| DrugEraOverlapping | FALSE |
| ConditionOccurrencePrimaryInpatientAnyTimePrior | FALSE |
| ConditionEraMediumTerm | FALSE |
| ConditionEraOverlapping | FALSE |
| ConditionEraStartShortTerm | FALSE |
| ObservationAnyTimePrior | FALSE |
| VisitConceptCountShortTerm | FALSE |
| DemographicsEthnicity | FALSE |
| 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 | FALSE |
| ProcedureOccurrenceAnyTimePrior | FALSE |
| DistinctObservationCountLongTerm | FALSE |
| ProcedureOccurrenceMediumTerm | FALSE |
| CharlsonIndex | TRUE |
| DemographicsPriorObservationTime | FALSE |
| MeasurementShortTerm | FALSE |
| DistinctProcedureCountMediumTerm | FALSE |
| ConditionEraLongTerm | FALSE |
| DrugGroupEraStartLongTerm | FALSE |
| DemographicsGender | TRUE |
| DeviceExposureAnyTimePrior | FALSE |
| ObservationLongTerm | FALSE |
| DemographicsIndexYearMonth | FALSE |
| ConditionOccurrenceMediumTerm | FALSE |
| longTermStartDays | -365 |
| DemographicsAge | TRUE |
| DrugGroupEraOverlapping | FALSE |
| DistinctMeasurementCountLongTerm | FALSE |
| MeasurementRangeGroupAnyTimePrior | FALSE |
| DistinctConditionCountMediumTerm | FALSE |
| DrugGroupEraMediumTerm | FALSE |
| ProcedureOccurrenceShortTerm | FALSE |
| ObservationMediumTerm | FALSE |
| ConditionGroupEraAnyTimePrior | FALSE |
| Chads2 | FALSE |
| DrugExposureAnyTimePrior | FALSE |
| DeviceExposureLongTerm | FALSE |
| DemographicsTimeInCohort | FALSE |
| DistinctMeasurementCountMediumTerm | FALSE |
| MeasurementValueShortTerm | FALSE |
| DeviceExposureMediumTerm | FALSE |
| ConditionGroupEraStartShortTerm | FALSE |
| ConditionOccurrencePrimaryInpatientMediumTerm | FALSE |
| MeasurementLongTerm | FALSE |
| DemographicsIndexYear | FALSE |
| MeasurementValueMediumTerm | FALSE |
| DrugEraStartMediumTerm | FALSE |
| MeasurementValueAnyTimePrior | FALSE |
| DistinctObservationCountShortTerm | FALSE |
| DrugEraMediumTerm | FALSE |
| ConditionGroupEraLongTerm | FALSE |
| 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 |

***Covariate Settings #2***

|  |  |
| --- | --- |
| 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 | FALSE |
| DrugGroupEraShortTerm | FALSE |
| ConditionEraStartLongTerm | FALSE |
| temporal | FALSE |
| DemographicsIndexMonth | TRUE |
| ConditionOccurrencePrimaryInpatientLongTerm | FALSE |
| ConditionEraAnyTimePrior | FALSE |
| addDescendantsToInclude | FALSE |
| ConditionGroupEraStartMediumTerm | FALSE |
| ProcedureOccurrenceLongTerm | FALSE |
| DrugExposureLongTerm | FALSE |
| DrugEraStartShortTerm | FALSE |
| DistinctIngredientCountMediumTerm | FALSE |
| DistinctMeasurementCountShortTerm | FALSE |
| MeasurementRangeGroupLongTerm | FALSE |
| ConditionGroupEraOverlapping | FALSE |
| MeasurementRangeGroupMediumTerm | FALSE |
| DrugGroupEraStartMediumTerm | FALSE |
| MeasurementAnyTimePrior | FALSE |
| MeasurementMediumTerm | FALSE |
| includedCovariateIds |  |
| ConditionOccurrenceAnyTimePrior | TRUE |
| DistinctConditionCountLongTerm | FALSE |
| MeasurementValueLongTerm | FALSE |
| DrugEraShortTerm | FALSE |
| DrugGroupEraAnyTimePrior | FALSE |
| DrugEraOverlapping | FALSE |
| ConditionOccurrencePrimaryInpatientAnyTimePrior | FALSE |
| ConditionEraMediumTerm | FALSE |
| ConditionEraOverlapping | FALSE |
| ConditionEraStartShortTerm | FALSE |
| ObservationAnyTimePrior | FALSE |
| VisitConceptCountShortTerm | FALSE |
| DemographicsEthnicity | FALSE |
| 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 | FALSE |
| ProcedureOccurrenceAnyTimePrior | FALSE |
| DistinctObservationCountLongTerm | FALSE |
| ProcedureOccurrenceMediumTerm | FALSE |
| CharlsonIndex | TRUE |
| DemographicsPriorObservationTime | FALSE |
| MeasurementShortTerm | FALSE |
| DistinctProcedureCountMediumTerm | FALSE |
| ConditionEraLongTerm | FALSE |
| DrugGroupEraStartLongTerm | FALSE |
| DemographicsGender | TRUE |
| DeviceExposureAnyTimePrior | FALSE |
| ObservationLongTerm | FALSE |
| DemographicsIndexYearMonth | FALSE |
| ConditionOccurrenceMediumTerm | FALSE |
| longTermStartDays | -365 |
| DemographicsAge | FALSE |
| DrugGroupEraOverlapping | FALSE |
| DistinctMeasurementCountLongTerm | FALSE |
| MeasurementRangeGroupAnyTimePrior | FALSE |
| DistinctConditionCountMediumTerm | FALSE |
| DrugGroupEraMediumTerm | FALSE |
| ProcedureOccurrenceShortTerm | FALSE |
| ObservationMediumTerm | FALSE |
| ConditionGroupEraAnyTimePrior | FALSE |
| Chads2 | FALSE |
| DrugExposureAnyTimePrior | FALSE |
| DeviceExposureLongTerm | FALSE |
| DemographicsTimeInCohort | FALSE |
| DistinctMeasurementCountMediumTerm | FALSE |
| MeasurementValueShortTerm | FALSE |
| DeviceExposureMediumTerm | FALSE |
| ConditionGroupEraStartShortTerm | FALSE |
| ConditionOccurrencePrimaryInpatientMediumTerm | FALSE |
| MeasurementLongTerm | FALSE |
| DemographicsIndexYear | FALSE |
| MeasurementValueMediumTerm | FALSE |
| DrugEraStartMediumTerm | FALSE |
| MeasurementValueAnyTimePrior | FALSE |
| DistinctObservationCountShortTerm | FALSE |
| DrugEraMediumTerm | FALSE |
| ConditionGroupEraLongTerm | FALSE |
| 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 2 target cohorts evaluated for 11 outcomes over 1 model over 2 covariates settings and over 3 population settings. In total there are 132 analysis performed.

# Strengths & Limitations

Strength

* The analysis can help gain insight into the clinical usefulness of each developed model by identifying whether it is transportable.

Limitations

* The external validation datasets may not have a sufficient number of (some of) the different outcomes to be used in the analysis.
* 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.

# Protection of Human Subjects

For this study, participants from various countries will process personal data from individuals which is collected in national/regional electronic health record databases. Due to the sensitive nature of this personal medical data, it is important to be fully aware of ethical and regulatory aspects and to strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy.

In agreement with these regulations, rather than combining person level data and performing only a central analysis, local analyses will be run, which generate non-identifiable aggregate summary results.

All the databases used in this study have a well-developed mechanism to ensure that regulations dealing with ethical use of the data and adequate privacy control are adhered to.

If required, the protocol has been reviewed by the Institutional Review Boards of the respective databases (IPCI, SIDIAP).

# Plans for Disseminating & Communicating Study Results

Dissemination activities to be undertaken will have mainly, although not exclusively, a scientific nature (articles, presentations at conferences, etc.).

# Tables & Figures

## Incidence Rate of Target & Outcome

***Feasibility assessment training data***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| T | O | TAR start | TAR end | T size | O count | O incidence |
| New MTX user | Acute MI | 1 | 730 | 5308 | 98 | 1.85 |
| New MTX user | All infections | 1 | 90 | 4750 | 476 | 10.02 |
| New MTX user | Colorectal cancer | 365 | 1826 | 3812 | 12 | 0.31 |
| New MTX user | Colorectal cancer | 365 | 1826 | 3812 | 12 | 0.31 |
| New MTX user | Leukopenia | 1 | 90 | 5306 | 21 | 0.40 |
| New MTX user | Opportunistic Infections | 1 | 90 | 5293 | 36 | 0.68 |
| New MTX user | Pancytopenia | 1 | 90 | 5317 | 8 | 0.15 |
| New MTX user | Serious Infection events | 1 | 90 | 5251 | 79 | 1.50 |
| New MTX user | Serious Infection events | 1 | 90 | 5251 | 79 | 1.50 |
| New MTX user | Stroke | 1 | 730 | 5301 | 127 | 2.40 |
| New MTX user | Stroke | 1 | 730 | 5301 | 127 | 2.40 |

***Feasibility assessment validation data***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Database | O | TAR | T size | O count | O incidence |
| PanTher | Acute MI | 2y | 41496 | 716 | 1.73 |
| IQVIA\_AMBEMR | Acute MI | 2y | 32524 | 115 | 0.35 |
| CCAE | Acute MI | 2y | 28084 | 173 | 0.62 |
| IQVIA\_GERMANY | Acute MI | 2y | 7401 | 38 | 0.51 |
| IQVIA\_THIN | Acute MI | 2y | 6935 | 44 | 0.63 |
| MDCR | Acute MI | 2y | 6613 | 210 | 3.18 |
| IQVIA\_HOSPITAL | Acute MI | 2y | 4140 | 191 | 4.61 |
| MDCD | Acute MI | 2y | 3427 | 78 | 2.28 |
| JMDC | Acute MI | 2y | 3299 | 9 | 0.27 |
| IQVIA\_LPDFRANCE | Acute MI | 2y | 3258 | 7 | 0.21 |
| Estonia | Acute MI | 2y | 1465 | 18 | 1.23 |
| IQVIA\_AUS | Acute MI | 2y | 560 | 14 | 2.50 |
| IPCI | Acute MI | 2y | 556 | 7 | 1.26 |
| SIDIAP | Acute MI | 2y | 3614 | 15 | 0.42 |
| PanTher | Serious Infection | 3m | 41072 | 397 | 0.97 |
| CCAE | Serious Infection | 3m | 27877 | 216 | 0.77 |
| MDCR | Serious Infection | 3m | 6533 | 152 | 2.33 |
| IQVIA\_HOSPITAL | Serious Infection | 3m | 3703 | 746 | 20.15 |
| MDCD | Serious Infection | 3m | 3355 | 111 | 3.31 |
| JMDC | Serious Infection | 3m | 3278 | 11 | 0.34 |
| Estonia | Serious Infection | 3m | 1464 | 8 | 0.55 |
| PanTher | Stroke | 2y | 41404 | 868 | 2.10 |
| IQVIA\_AMBEMR | Stroke | 2y | 32561 | 131 | 0.40 |
| CCAE | Stroke | 2y | 28082 | 243 | 0.87 |
| IQVIA\_GERMANY | Stroke | 2y | 7416 | 37 | 0.50 |
| IQVIA\_THIN | Stroke | 2y | 6937 | 21 | 0.30 |
| MDCR | Stroke | 2y | 6609 | 297 | 4.49 |
| IQVIA\_HOSPITAL | Stroke | 2y | 4127 | 199 | 4.82 |
| MDCD | Stroke | 2y | 3415 | 108 | 3.16 |
| JMDC | Stroke | 2y | 3299 | 21 | 0.64 |
| Estonia | Stroke | 2y | 1464 | 24 | 1.64 |
| SIDIAP | Stroke | 2y | 3615 | 7 | 0.19 |

# Appendices

## Study Generation Version Information

Skeleton Version: PatientLevelPredictionStudy - v0.0.1

Identifier / Organization: OHDSI

# References

1. Visser K, van der Heijde D. Optimal dosage and route of administration of methotrexate in rheumatoid arthritis: a systematic review of the literature. Ann Rheum Dis. 2009;68(7):1094-9.

2. Smolen JS, Landewe R, Breedveld FC, Buch M, Burmester G, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2013 update. Ann Rheum Dis. 2014;73(3):492-509.

3. Smolen JS, Landewe R, Bijlsma J, Burmester G, Chatzidionysiou K, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann Rheum Dis. 2017;76(6):960-77.

4. Singh JA, Saag KG, Bridges SL, Jr., Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol. 2016;68(1):1-26.

5. Costello R, David T, Jani M. Impact of Adverse Events Associated With Medications in the Treatment and Prevention of Rheumatoid Arthritis. Clin Ther. 2019;41(7):1376-96.

6. Dixon WG, Watson K, Lunt M, Hyrich KL, Silman AJ, Symmons DP, et al. Rates of serious infection, including site-specific and bacterial intracellular infection, in rheumatoid arthritis patients receiving anti-tumor necrosis factor therapy: results from the British Society for Rheumatology Biologics Register. Arthritis Rheum. 2006;54(8):2368-76.

7. Mercer LK, Galloway JB, Lunt M, Davies R, Low AL, Dixon WG, et al. Risk of lymphoma in patients exposed to antitumour necrosis factor therapy: results from the British Society for Rheumatology Biologics Register for Rheumatoid Arthritis. Ann Rheum Dis. 2017;76(3):497-503.

8. Low ASL, Symmons DPM, Lunt M, Mercer LK, Gale CP, Watson KD, et al. Relationship between exposure to tumour necrosis factor inhibitor therapy and incidence and severity of myocardial infarction in patients with rheumatoid arthritis. Annals of the rheumatic diseases. 2017;76(4):654-60.

9. Salliot C, van der Heijde D. Long-term safety of methotrexate monotherapy in patients with rheumatoid arthritis: a systematic literature research. Annals of the rheumatic diseases. 2009;68(7):1100-4.

10. Lacaille D, Guh DP, Abrahamowicz M, Anis AH, Esdaile JM. Use of nonbiologic disease-modifying antirheumatic drugs and risk of infection in patients with rheumatoid arthritis. Arthritis and rheumatism. 2008;59(8):1074-81.

11. Crowson CS, Matteson EL, Roger VL, Therneau TM, Gabriel SE. Usefulness of risk scores to estimate the risk of cardiovascular disease in patients with rheumatoid arthritis. Am J Cardiol. 2012;110(3):420-4.

12. Arts EEA, Popa C, Den Broeder AA, Semb AG, Toms T, Kitas GD, et al. Performance of four current risk algorithms in predicting cardiovascular events in patients with early rheumatoid arthritis. Annals of the rheumatic diseases. 2015;74(4):668-74.

13. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart. 2008;94(1):34-9.

14. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-75.

15. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.

16. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

17. Team RC. R: A language and environment for statistical computing. 2013.