OHDSI Predicting outcomes for new-users of Celecoxib protocol: Risk models for Myocardial Infarction, GI Hemorrhage, Acute Renal Failure, and Angioedema.

**Authors:**

Jenna Reps, PhD, Janssen Research and Development

Marc Suchard, MD, PhD, University of California, Los Angeles

Patrick Ryan, PhD, Janssen Research and Development

Martijn Schuemie, PhD, Janssen Research and Development

**Date:** 16 October 2015

**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: Drs. Ryan, Schuemie, Reps are employees of Janssen Research & Development.

Table of Contents

[1 Table of contents 2](#_Toc432765229)

[3 List of abbreviations 3](#_Toc432765230)

[4 Abstract 3](#_Toc432765231)

[5 Amendments and Updates 3](#_Toc432765232)

[6 Milestones 3](#_Toc432765233)

[7 Rationale and Background 4](#_Toc432765234)

[8 Research Questions and Objectives 4](#_Toc432765235)

[8.1 Research Questions 4](#_Toc432765236)

[8.2 Objectives 5](#_Toc432765237)

[9 Research methods 5](#_Toc432765238)

[9.1 Study Design 5](#_Toc432765239)

[9.1.1 Overview 5](#_Toc432765240)

[9.1.2 Study population 6](#_Toc432765241)

[9.1.3 Additional analysis details 6](#_Toc432765242)

[9.1.4 Analysis variations 6](#_Toc432765243)

[9.2 Variables 7](#_Toc432765244)

[9.2.1 Exposures 7](#_Toc432765245)

[9.2.2 Outcomes 7](#_Toc432765246)

[9.2.3 Potential confounders 9](#_Toc432765247)

[9.2.4 Negative controls 9](#_Toc432765248)

[9.2.5 Other variables 9](#_Toc432765249)

[9.3 Data Sources 9](#_Toc432765250)

[9.4 Sample Size and Study Power 11](#_Toc432765251)

[9.5 Quality control 12](#_Toc432765252)

[9.6 Strengths and Limitations of the Research Methods 12](#_Toc432765253)

[10 Protection of Human Subjects 12](#_Toc432765254)

[11 Plans for Disseminating and Communicating Study Results 13](#_Toc432765255)

[12 References 13](#_Toc432765256)

# List of abbreviations

ATC Anatomic Therapeutic Chemical

CYCLOPS Cyclic coordinate descent for logistic, Poisson and survival analysis

DCSI Diabetes Complications Severity Index

GI Gastro-Intestinal

MedDRA Medical Dictionary for Regulatory Activities

NSAID Non-Steroidal Anti-Inflammatory Drug

nsNSAID Non-Selective Non-Steroidal Anti-Inflammatory Drug

OHDSI Observational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

PS Propensity Scores

# Abstract

This study aims to evaluate the performance of the PatientLevelPrediction package for the purpose of developing data-driven personalized risk models. The PatientLevelPrediction implements a new-user cohort study design to create a dataset incorporating a large number of variables and then learns an easy to interpret risk prediction model via regularized logistic regression. The package directly queries data in the OMOP Common Data Model, and is therefore suited for running in a distributed research network.

Here we would like to provide a proof of principle of the PatientLevelPrediction package by determining the models’ discriminative abilities and calibration scores when predicting various health outcomes in new users of celecoxib. We will focus on personalized prediction of the following health outcomes: myocardial infarction, gastrointestinal bleed, angioedema and acute renal failure. We would also like to show that studies implemented using the PatientLevelPrediction package can be easily deployed in a distributed research network.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author(s) | Comments |
| 0.2 | 16 Oct | Jenna Reps | Followed Martijn’s suggestions and added refs/abstract |
| 0.1 | 14 October 2015 | Jenna Reps | Initial draft |
| 0.0 | 7 October 2015 | Martijn Schuemie | Template draft |

# Milestones

|  |  |
| --- | --- |
| Milestone | Planned / Estimated Date |
| Start of analysis |  |
| End of analysis |  |
| Posting of results |  |
| Submission of manuscript |  |

# Rationale and Background

Observational data present the opportunity to identify patterns that can be utilized to develop discriminative personalized health outcome risk models. The most widely implemented framework for developing health outcome risk models requires experts to define a set of independent variables, and these are then fed into a logistic or cox regression [1,2] The disadvantage of such a framework is that it requires the time consuming aspect of defining different sets of independent variables for each health outcome and dataset. Furthermore, the model is generally restricted to a small selection of variables and therefore ignores large quantities of data, whose inclusion may be able to improve risk prediction performance. For example, restricting the set of independent variables based on current expert knowledge may result in the exclusion of highly predictive independent variables because they are unknown at the time.

The PatientLevelPrediction package developed in OHDSI (Observational Health Data Science and Informatics) uses an adaptive framework approach, where a data-driven method explores all the data to find the independent variables that are predictive of the outcome. This is accomplished by the logistic regression model including a large number of independent variables but each independent variable’s coefficient has a Laplace prior [3], which acts as a type of regularization and results in many coefficients being shrunk to zero to limit model overfitting [4]. The independent variables with non-zero coefficients are selected by the model as they are predictive of the health outcome. If such a framework is able to perform well, then it could be efficiently applied, across the network of observational data available to the OHDSI community, to develop risk models for many health outcomes. The added advantage of the lasso logistic regression is that is learns an easy to interpret sparse model, so it may also be used to gain new medical insight by clearly highlighting unknown risk factors.

In this study described here we would like to provide a proof of principle of the PatientLevelPrediction package, and evaluate the performance and robustness of the risk prediction models across a range of health outcomes and datasets. We would also like to show that the PatientLevelPrediction package can be easily deployed in a distributed research network to develop risk models using different observational datasets.

# Research Questions and Objectives

## Research Questions

We will investigate whether a Bayesian lasso logistic regression model that considers a large number of independent variables can obtain a consistently high performance in predicting the risk of various health outcomes. We believe PatientLevelPrediction is capable of selecting the important independent variables from the data and will therefore consistently obtain good discriminative performances (AUC>0.7) across the outcomes and datasets.

Primary hypothesis

* Using regularized logistic regression with a large set of features derived from all data elements in the observational databases we can construct models that are predictive of the outcomes of interest.

The model performance is likely to depend on attributes of the data (e.g., average observation time per patient, type of observational data, and data size). We believe there will be a key set of predictive independent variables for each outcome that are consistently selected by models developed using different data.

Secondary hypotheses 1

* The risk models developed for the same outcome across different datasets will consistently select the same set of key independent variables

Furthermore, we would like to test whether the PatientLevelPrediction package can be deployed in a distributed data network. The study described here will be implemented into an R package, and this package will be distributed to partners willing to execute it.

Secondary hypothesis 2

* The code implementing the study described here can be executed in a distributed network against a range of diverse databases.

## Objectives

Primary objective

* Determine the predictive performance of the OHDSI PatientLevelPrediction using standard measures including model calibration and model discrimination.

Secondary objectives

* Investigate the robustness of the model across different datasets and identify any limitations that are specific to certain dataset attributes.
* Show feasibility of running PatientLevelPrediction in a distributed data network

# Research methods

## Study Design

### Overview

This study will be developing logistic regression models with regularization via a Bayesian lasso for the purpose of health outcome risk prediction. This will be an observational study that involves creating a retrospective new-user cohort of patients who are prescribed celecoxib, with an index date corresponding to the first time prescription of celecoxib. The cohort will then be followed-up for 1-years. By ‘retrospective’ we mean the study will use data already collected at the start of the study. By ‘observational’ we mean no intervention will take place in the course of this study. By ‘new-user’ we mean we will only analyze the first exposure of a subject to the drugs of interest.

The outcomes to be predicted by the models are Myocardial Infarction, GI Hemorrhage, Acute Renal Failure, and Angioedema. We will create labelled data for each person in the cohort, where their feature vector will consist of demographic variables and indicator variables for each medical condition, drug prescription, measurement and observation that was recorded prior to index and the label will correspond to whether they experienced the outcome of interest during follow-up.

To evaluate the models, we will partition the labelled data into a test set (25%) and train set (75%). The hyper parameters for the models will be selected using 10-fold cross validation on the train set and a final model will be trained using the train set and chosen hyper parameters. The performance of the models will be evaluated using 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.

### Study population

All subjects in the database will be included who meet the following criteria: (note: the index date is the start of the first exposure to celecoxib)

* Exposure to celecoxib
* At least 183 days of observation time prior to the index date
* No diagnose of the outcome of interest preceding the index date

### Additional analysis details

The risk models will be fitted using a regularized regression with a LaPlace prior. The regularization hyperparameter will be selected by optimizing the likelihood in a 10-fold cross-validation.

### Analysis variations

The following variations of the analysis will be performed:

Primary analysis:

* [Classification – exclude people with prior outcome] Logistic regression with laplace prior – we will exclude patients who have the outcome recorded prior to their index data.

Secondary analysis:

* [Classification – include people with prior outcome]] Logistic regression with laplace prior – we will exclude patients who have the outcome recorded prior to their index data.

## Variables

### Exposures

#### Celecoxib

Index rule defining the index date:

* First exposure to any drug containing the RxNorm ingredient celecoxib

Inclusion rules based on the index date:

* At least 183 days of observation time prior to the index date

### Outcomes

#### Myocardial infarction

Index rule defining the index date:

* Occurrence of a myocardial infarction code (excluding codes referring to an old myocardial infarction) as a primary diagnosis in an inpatient or emergency room setting

Inclusion rules based on the index date:

* Cannot have a myocardial infarction code (excluding codes referring to an old myocardial infarction) in the 90 days preceding the index date.

#### Myocardial infarction and ischemic death

Index rule defining the index date:

* Occurrence of a myocardial infarction code (excluding codes referring to an old myocardial infarction) as a primary diagnosis in an inpatient or emergency room setting, or
* Any occurrence of death

Inclusion rules based on the index date:

Either

* Occurrence of a myocardial infarction code (excluding codes referring to an old myocardial infarction) on the index date as a primary diagnosis in an inpatient or emergency room setting, and
* No myocardial infarction code (excluding codes referring to an old myocardial infarction) in the 30 days preceding the index date

Or

* Any occurrence of death on the index date, and
* A diagnose code indicating angina pectoris, coronary occlusion, preinfarction syndrome, or myocardial infarction (excluding codes referring to an old myocardial infarction) on the index date or on the day before the index date

#### Gastrointestinal hemorrhage

Index rule defining the index date:

* Occurrence of a gastrointestinal hemorrhage code as a primary diagnosis in an inpatient or emergency room setting

Inclusion rules based on the index date:

* Cannot have a gastrointestinal hemorrhage diagnose code in the 30 days preceding the index date.

#### Angioedema

Index rule defining the index date:

* Any occurrence of an angioedema diagnose code

Inclusion rules based on the index date:

* Cannot have an angioedema diagnose code in the 30 days preceding the index date.

#### Acute liver failure

Index rule defining the index date:

* Occurrence of an acute nephropathy code as a primary diagnosis in an inpatient or emergency room setting

Inclusion rules based on the index date:

* Cannot have an acute nephropathy diagnose code in the 30 days preceding the index date.

#### Drug-induced liver injury

Index rule defining the index date:

* Occurrence of a measurement of aminotransferase (aspartate or alanine) with a value at least three times the upper bound of the normal range

Inclusion rules based on the index date:

* Must have a measurement of bilirubin with a value at least twice the upper bound of the normal range in the 30 days before to the 30 days after the index date
* Cannot have a diagnose of viral hepatitis any time before the index date to 30 days after the index date
* Must have a measurement of alkaline phosphatase with a value at least twice the upper bound of the normal range any time before the index date to 30 days after the index date
* Cannot have a diagnose if liver damage any time before the index date

#### Heart failure

Index rule defining the index date:

* Occurrence of a heart failure code as a primary diagnosis in an inpatient or emergency room setting

Inclusion rules based on the index date:

* Cannot have a heart failure code in the 30 days preceding the index date

### Potential confounders

NA

### Negative controls

NA

### Other variables

NA

## 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 4 or OMOP Common Data Model, version 5. The complete specification for OMOP Common Data Model, version 4 is available at: <http://omop.org/cdm>. The complete specification for OMOP Common Data Model, version 5 is available at: <https://github.com/OHDSI/CommonDataModel>. The following databases will be included in this analysis:

* Truven MarketScan Commercial Claims and Encounters (CCAE)
* Truven MarketScan Medicare Supplemental Beneficiaries (MDCR)
* Truven MarketScan Multi-state Medicaid (MDCD)
* Optum ClinFormatics (Optum)
* Clinical Practice Research Datalink (CPRD)
* <<add others who agree to participate>>

Truven MarketScan Commercial Claims and Encounters (CCAE)

CCAE is an administrative health claims database for active employees, early retirees, COBRA continues, and their dependents insured by employer-sponsored plans (individuals in plans or product lines with fee-for-service plans and fully capitated or partially capitated plans). As of 30November2014, CCAE contained 117m patients with patient-level observations from Jan2000 through Jul2014. Source codes used in CCAE include: conditions- ICD-9-CM; drugs: NDC, HCPCS, ICD-9-CM; procedures: CPT-4, HCPCS, ICD-9-CM; lab: LOINC.

The ETL specification for transforming CCAE into the OMOP CDM is available at: <http://omop.org/cdm>.

ACHILLES has been used to characterize the database and provide a data quality assessment. The ACHILLES summary is available internally within Janssen at: <http://hix.jnj.com/achilles/#/truven_ccae/dashboard>.

Truven MarketScan Medicare Supplemental Beneficiaries (MDCR)

MDCR is an administrative health claims database for Medicare-eligible active and retired employees and their Medicare-eligible dependents from employer-sponsored supplemental plans (predominantly fee-for-service plans). Only plans where both the Medicare-paid amounts and the employer-paid amounts were available and evident on the claims were selected for this database.. As of 30November2014, MDCR contained 9m patients with patient-level observations from Jan2000 through Jul2014. Source codes used in MDCR include: conditions- ICD-9-CM; drugs: NDC, HCPCS, ICD-9-CM; procedures: CPT-4, HCPCS, ICD-9-CM; lab: LOINC.

The ETL specification for transforming MDCR into the OMOP CDM is available at: <http://omop.org/cdm>.

ACHILLES has been used to characterize the database and provide a data quality assessment. The ACHILLES summary is available internally within Janssen at: <http://hix.jnj.com/achilles/#/truven_mdcr/dashboard>.

Truven MarketScan Multi-state Medicaid (MDCD)

MDCD is an administrative health claims database for the pooled healthcare experience of Medicaid enrollees from multiple states. As of 30November2014, MDCD contained 16m patients with patient-level observations from Jan2006 through Dec2012. Source codes used in MDCD include: conditions- ICD-9-CM; drugs: NDC, HCPCS, ICD-9-CM; procedures: CPT-4, HCPCS, ICD-9-CM; lab: LOINC.

The ETL specification for transforming MDCD into the OMOP CDM is available at: <http://omop.org/cdm>.

ACHILLES has been used to characterize the database and provide a data quality assessment. The ACHILLES summary is available internally within Janssen at: <http://hix.jnj.com/achilles/#/truven_mdcd/dashboard>.

Optum ClinFormatics (Optum)

Optum is an administrative health claims database for members of United Healthcare, who enrolled in commercial plans (including ASO, 36.31M), Medicaid (prior to July 2010, 1.25M) and Legacy Medicare Choice (prior to January 2006, 0.36M) with both medical and prescription drug coverage. As of 30November2014, Optum contained 38m patients with patient-level observations from Oct2005 through Dec2013. Source codes used in Optum include: conditions- ICD-9-CM; drugs: NDC, HCPCS, ICD-9-CM; procedures: CPT-4, HCPCS, ICD-9-CM; lab: LOINC.

The ETL specification for transforming Optum into the OMOP CDM is available at: <http://omop.org/cdm>.

ACHILLES has been used to characterize the database and provide a data quality assessment. The ACHILLES summary is available internally within Janssen at: <http://hix.jnj.com/achilles/#/optum/dashboard>.

Clinical Practice Research Datalink (CPRD)

CPRD is an anonymized longitudinal electronic health records from primary care practices in UK. Patient management system with many aspects of patient care covered, including diagnoses, prescriptions, signs and symptoms, procedures, labs, lifestyle factors, clinical and administrative/social data. As of 30November2014, CPRD contained 11m patients with patient-level observations from Jan1988 through Nov2013. Source codes used in CPRD include: conditions- Read; drugs: Multilex; procedures: OPCS.

The ETL specification for transforming CPRD into the OMOP CDM is available at: <http://omop.org/cdm>.

ACHILLES has been used to characterize the database and provide a data quality assessment. The ACHILLES summary is available internally within Janssen at: <http://hix.jnj.com/achilles/#/cprd/dashboard>.

*Database X*

*Database X description*

The ETL specification for transforming *Database X* into the OMOP CDM is available at:  *ETL\_specification\_URL*

ACHILLES has been used to characterize the database and provide a data quality assessment. The ACHILLES summary is available at: *URL to ACHILLES*.

## Sample Size and Study Power

NA

## Quality control

We will investigate the risk model by

* Calculating the calibration and discrimination measures and comparing against existing model benchmarks
* Determining the robustness of outcome specific models across the datasets
* Inspection of the fitted outcome model for large coefficients and predictors that we cannot explain (post-hoc).

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 framework can overcome data sparseness
* The framework utilizes all the data and may identify new risk factors
* The framework produces easy to interpret models
* The framework can be readily implemented for any outcome to efficiently develop a risk model
* The framework can be readily implemented across the OHDSI network

Limitations

* The generalized linear models included only consider linear combinations of the variables in the prediction, this may miss out on utilizing highly predictive interactions.
* The logistic regression cannot include censored patients, so any person who does not get observed for the complete follow-up period or experiences the outcome during follow-up is excluded. This can cause a bias in the population used to develop and model and may prevent the model being applicable to the general population.
* Not all medical events are recorded into the observational datasets and some recordings can be incorrect. This results in a noisy dataset with potential outcome misclassification. It is unknown to what extent misclassification of any of the outcomes occurs.
* The risk models are only applicable to the population of patients represented by the data used to train the model.

# 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] Hippisley-Cox, J., Coupland, C., Robson, J., & Brindle, P. (2010). Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *Bmj*, *341*.

[2] North, R. A., McCowan, L. M., Dekker, G. A., Poston, L., Chan, E. H., Stewart, A. W., ... & Kenny, L. C. (2011). Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort.*Bmj*, *342*.

[3] Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681-686.

[4] Ye, J., & Liu, J. (2012). Sparse methods for biomedical data. *ACM SIGKDD Explorations Newsletter*, *14*(1), 4-15.