Proof of concept study for large-scale patient-level predictive modeling in the OHDSI data network.

**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

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

**Date:** 23 September 2016

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

Table of Contents

[2 List of abbreviations 3](#_Toc460883277)

[3 Abstract 3](#_Toc460883278)

[4 Amendments and Updates 3](#_Toc460883279)

[5 Milestones 4](#_Toc460883280)

[6 Rationale and Background 4](#_Toc460883281)

[7 Research Questions and Objectives 5](#_Toc460883282)

[7.1 Research Questions 5](#_Toc460883283)

[7.2 Objectives 5](#_Toc460883284)

[8 Research methods 6](#_Toc460883285)

[8.1 Study Design 6](#_Toc460883286)

[8.1.1 Overview 6](#_Toc460883287)

[8.1.2 Study population 7](#_Toc460883288)

[8.1.3 Additional analysis details 7](#_Toc460883289)

[8.1.4 Analysis variations 7](#_Toc460883290)

[8.2 Variables 8](#_Toc460883291)

[8.2.1 At risk 8](#_Toc460883292)

[8.2.2 Outcomes 8](#_Toc460883293)

[8.3 Data Sources 12](#_Toc460883300)

[8.4 Quality control 13](#_Toc460883303)

[8.5 Strengths and Limitations of the Research Methods 13](#_Toc460883304)

[9 Protection of Human Subjects 14](#_Toc460883305)

[10 Plans for Disseminating and Communicating Study Results 14](#_Toc460883306)

[11 References 14](#_Toc460883307)

# List of abbreviations

ATC Anatomic Therapeutic Chemical

CYCLOPS Cyclic coordinate descent for logistic, Poisson and survival

OHDSI Observational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

# Abstract

Clinical decision making is a complicated task in which the clinician has to infer a diagnosis or treatment pathway based on the available medical history of the patient and the current clinical guidelines. Clinical prediction models have been developed to support this decision making process and are used in clinical practice in a wide spectrum of specialties. These models predict a diagnostic or prognostic outcome based on a combination of patient characteristics, e.g. demographic information, disease history, treatment history.

Surprisingly, most of the currently used models are estimated using small datasets and contain a limited set of patient characteristics. This low sample size, and thus low statistical power, forces the data analyst to make stronger modelling assumptions. The selection of the often limited set of patient characteristics is strongly guided by the expert knowledge at hand. This contrasts sharply with the reality of modern medicine wherein patients generate a rich digital trail, which is well beyond the power of any medical practitioner to fully assimilate. Presently, health care is generating a large amount of patient-specific information contained in the Electronic Health Record (EHR). This includes structured data in the form of diagnoses, medications, laboratory test results, and unstructured data contained in clinical narratives. Currently, it is unknown how much predictive accuracy can be gained by leveraging the large amount of data originating from the complete EHR of a patient.

This proof of concept study aims to investigate the feasibility of large-scale predictive modelling using observational healthcare data in the OHDSI network. In this study patient-level prediction models will be trained and evaluated for multiple health outcomes in patients with pharmaceutically treated depression.

The models will be developed using the Patient-Level Prediction Package developed in OHDSI. The discriminative performance and calibration scores of the models will be assessed in the OHDSI data network. This study will develop prediction models for 22 outcomes across the OHDSI collaborative network datasets.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author(s) | Comments |
| 1.0 | 21 Sep 2016 | Peter Rijnbeek | Final version |
| 0.3 | 02 Sep 2016 | Jenna Reps | Revising |
| 0.2 | 19 Aug 2016 | Peter Rijnbeek | Editing |
| 0.1 | 14 Aug 2016 | 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 Health Data Sciences and Informatics (OHDSI) holds the promise of making massive-scale, patient-specific predictive modeling a reality. The OHDSI network contains longitudinal data on over 600 million patients observed for multiple years and comprising over 5 billion clinical observations. The data is stored in a common data model (CDM), enabling uniform and transparent analysis. These large standardized populations contain rich data to build highly predictive large-scale models and also provide immediate opportunity to serve large communities of patients who are in most need of improved quality of care. Effective exploitation of these massive dataset to develop patient-level prediction models demands a standardized pipeline for both model development and evaluation.

A patient level prediction model problem is defined by an ‘at risk’ cohort (the group of people we wish to do the prediction for), the ‘outcome’ cohort (the outcome we wish to predict) and the ‘at-risk’ period (time window relative to the start of the at risk cohort index date). At present only a limited number of conditions have existing patient level prediction models and little is known about the feasibility of utilizing observational databases for clinically useful patient level prediction models at scale (for all suitable ‘at risk’ and ‘outcome’ cohort pairs). It very interesting to use the observational databases to determine a very large number of ‘at-risk’ and ‘outcome’ pairs and develop prediction models for all these pairs.

In this proof on concept study we focus on a single ‘at-risk’ cohort (pharmaceutically treated depression cohort) and investigate 22 ‘outcome’ cohorts. The study will be implemented across the OHDSI collaborator network to externally validate the models and evaluate their transportability across the world.

# Research Questions and Objectives

## Research Questions

We want to investigate the performance of the various patient level prediction models to gain insight into the feasibility of large-scale patient level prediction models. Evaluation of the models both internally and externally is a key component in this assessment.

Primary hypothesis

* Observational databases can be used to develop a large number of patient level prediction models that have good discriminative abilities and are well calibrated.

Furthermore, we would like to test whether the large scale prediction package can be deployed in a distributed data network. This presents the opportunity to externally validate the models across numerous databases spanning various countries using a standardized approach. The study is completely implemented in an R package which will be distributed to partners that are willing to participate.

Secondary hypothesis 1

* Transportability assessments are feasible in a large data network.

The no free lunch theory suggests that there will be no classifier that always outperforms the others. In this study we will investigate five different classifiers and compare their performances across the ‘outcome’ cohorts and OHDSI databases.

Secondary hypothesis 2

* There will be no classifiers that outperforms the other classifiers across all outcomes investigated in this study.

# Research methods

## Study Design

### Overview

For a single ‘at risk’ cohort consisting of people with pharmaceutically treated depression we will develop prediction models that predict whether 22 different outcomes occur within a year of the start of the depression period. We will assess five different classifiers to also investigate database and method differences for each prediction problem. This will be distributed across the OHDSI network observational databases.

The prediction problem will leverage the full medical history at the start of the depression period, i.e. demographics, conditions and drugs. The classification outcome labels will be constructed by setting all the ‘at risk’ people who develop the outcome as class 1 and all the ‘at risk’ people who do not develop the outcome as class 0. The classifiers will then use the labeled data (features paired with the label) to learn patterns that predict the class of a person based on their features up to the start of the depression episode (the index date). We will use 75% of the labelled data to train the models and the remaining 25% of the data will be used to perform an internal validation of the prediction model. This 25% of the data will not be seen by the classifier when it is being trained. We will perform 3-fold cross validation when training the classifiers to identify the optimal hyper-parameters for each classifier.

The outcomes to be predicted by the models are: acute myocardial infarction, gastrointestinal hemhorrage, open-angle glaucoma, stroke, suicide and suicidal ideation, insomnia, diarrhea, nausea, hypothyroidism, constipation, seizure, delirium, alopecia, tinnitus, vertigo, hyponatremia, decreased libido, fracture, hypotension, acute liver injury, hyperprolactinemia, ventricular arrhythmia and sudden cardiac death.

The classifiers are: lasso regularized logistic regression, random forest, naïve bayes, gradient boosting machines and k-nearest neighbors.

To internally evaluate the models on the test set, the area under the receiver operating characteristic curve (AUC) is determined to evaluate the discriminative performance of the models. Additionally, model calibration is assessed by plotting the predicted risk against the observed fraction and the brier score for general prediction performance is calculated. The models are externally validated by applying each trained model to the other participating databases .

### Study population

All subjects in the database will be included who meet the following criteria: (note: the index date is the start of the depression era)

* Recorded condition of depression
* Exposure to depression therapy
* At least 365 days of observation time prior to the index date
* No diagnose of the outcome of interest preceding the index date

### Additional analysis details

The lasso logistic regression 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 3-fold cross-validation.

The random forest hyper-parameter will be selected by performing a 3-fold cross validation using a grid search on the following parameters space:

* ntrees =50, 500 or 1000
* mtries = sqaure root of number of features, 50 or 500
* max\_depth =4,10 or 17
* varImp =True or False (if True a preliminary step is included where an initial random forest is trained and the features with a feature importance score greater than the median across all features are only included in the main model)

The gradient boosting machine uses decision trees and the hyper-parameter will be selected by performing a 3-fold cross validation using a grid search on the following parameters space:

* ntrees = 50, 100 or 250
* max\_depth = 4, 7 or 16
* min\_rows = 1 or 20

The Naïve bayes classifier does not have hyper-parameters but a preliminary step of feature selection is implemented using a univariate correlation measure between each feature and the outcome label to select the top 2000 features.

The K-nearest neighbors classifier uses a set k value of 10000 as preliminary work has showed a large k value resulted in better performance.

### Analysis variations

The following variations of the analysis will be performed in future work:

Primary analysis:

* [Classification – using covariate setting 1] Using all demographics, conditions and drug records as features

Secondary analysis:

* [Classification – using covariate setting 2] Using all demographics, conditions, drugs, observations, measurements and procedures.

## Variables

### At risk

#### Pharmaceutically treated depression

Index rule defining the index date:

* First condition record of depression or first drug record of antidepressant

Inclusion rules based on the index date:

* At least 365 days of observation time prior to the index date
* At least 30 days of observation time after the index date
* Must have a condition record of depression recorded within 0 days prior until 30 days after index date
* Must have a drug record of antidepressant recorded within 0 days prior until 30 days after index date
* Must have no condition records of psychosis recorded within any days prior to index date
* Must have no condition records of mania recorded within any days prior to index date
* Must have no condition records of dementia recorded within any days prior to index date

### Outcomes

#### Acute myocardial infarction

Index rule defining the index date:

* First occurrence of an acute 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:

* None

#### Gastrointestinal hemorrhage

Index rule defining the index date:

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

Inclusion rules based on the index date:

* None

#### Open-angle glaucoma

Index rule defining the index date:

* First occurrence of an open-angle glaucoma code

Inclusion rules based on the index date:

* >= 365 days prior observation
* A distinct open-angle glaucoma code recorded within 365 days after index with a provider specialty of optician/optometry/ophthalmology

#### Stroke

Index rule defining the index date:

* First occurrence of a stoke code at an inpatient visit

Inclusion rules based on the index date:

* None

#### Suicide and suicidal ideation

Index rule defining the index date:

* First occurrence of a suicide and suicidal ideation condition or observation code

Inclusion rules based on the index date:

* None

#### Insomnia

Index rule defining the index date:

* First occurrence of an insomnia code

Inclusion rules based on the index date:

* None

#### Diarrhea

Index rule defining the index date:

* First occurrence of a diarrhea code

Inclusion rules based on the index date:

* None

#### Nausea

Index rule defining the index date:

* First occurrence of a nausea code

Inclusion rules based on the index date:

* None

#### Hypothyroidism

Index rule defining the index date – pick the first date satisfying:

* Occurrence of a hypothyroidism code

Inclusion rules based on the index date:

* One or more hypothyroidism codes occurring in the 90 days following

#### Constipation

Index rule defining the index date:

* First occurrence of a constipation code

Inclusion rules based on the index date:

* None

#### Seizure

Index rule defining the index date:

* First occurrence of a seizure and seizure disorder code with an emergency room or inpatient visit

Inclusion rules based on the index date:

* None

#### Delirium

Index rule defining the index date:

* First occurrence of a delirium code with an emergency room or inpatient visit

Inclusion rules based on the index date:

* None

#### Alopecia

Index rule defining the index date:

* First occurrence of an alopecia code

Inclusion rules based on the index date:

* None

#### Tinnitus

Index rule defining the index date:

* First occurrence of a tinnitus code

Inclusion rules based on the index date:

* None

#### Vertigo

Index rule defining the index date:

* First occurrence of a vertigo code

Inclusion rules based on the index date:

* None

#### Hyponatremia

Index rule defining the index date:

* First occurrence of a hyponatremia condition code or a measurement of serum sodium with a value less than 136 millimole per litre

Inclusion rules based on the index date:

* None

#### Decreased libido

Index rule defining the index date:

* First occurrence of a decreased libido condition code

Inclusion rules based on the index date:

* None

#### Fracture

Index rule defining the index date:

* First occurrence of a fracture condition code

Inclusion rules based on the index date:

* None

#### Hypotension

Index rule defining the index date:

* First occurrence of a hypotension condition code

Inclusion rules based on the index date:

* None

#### Acute liver injury

Index rule defining the index date:

* First occurrence of an acute livery injury condition code during an emergency or inpatient visit

Inclusion rules based on the index date:

* No acute liver injury exclusion concepts condition codes recorded 365 days before and up to 60 days after

#### Hyperprolactinemia

Index rule defining the index date:

* First occurrence of a hyperprolactinemia condition code

Inclusion rules based on the index date:

* None

#### Ventricular arrhythmia and sudden cardiac death

Index rule defining the index date:

* First occurrence of a ventricular arrhythmia and sudden cardiac death condition code with a primary/1st position condition type and a visit occurrence of emergency room or inpatient visit.

Inclusion rules based on the index date:

* None

## 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>.

## Quality control

We will investigate the risk model by

* Calculating the calibration and discrimination measures and comparing against any existing model (identified using a literature search) benchmarks
* Comparison of performance between methods to identify any outlier models. Any outliers will be inspected to check for potential errors.
* Determining the robustness of outcome specific models across the OHDSI datasets
* Inspection of the fitted lasso regularized logistic regression outcome model for large coefficients
* Inspection of all the models for 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 analysis can help us gain insight into the pipeline required to do the all-by-all analysis
* The analysis can help gain insight into to suitability of each classifier when applied for prediction modelling using observational data.
* The analysis can be readily implemented across the OHDSI network – this will give us insight into how much models vary across datasets and how transportable models are.

Limitations

* Not all outcomes will occur a sufficient number of times within the prediction risk period. It may not be possible to develop prediction models for rarely occurring outcomes.
* The classifiers 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 and may not be generalizable to the wider population.
* Some of the classifiers have a large number of hyper-parameters and performing a large search for the optimal hyper-parameter may not be possible due to time limits. This may result in sub-optimal performance for certain classifiers.
* 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

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

None.