Version: 0.1.0

# About

## List of Abbreviations

|  |  |
| --- | --- |
| AUC | Area Under the receiver-operator Curve |
| CCAE | IBM MarketScan Commercial Claims and Encounters |
| CDM | Common Data Model |
| CRAN | Comprehensive R Archive Network |
| EHR | Electronic Health Record |
| IPCI | Integrated Primary Care Information |
| IRB | Institutional review board |
| MDCR | IBM MarketScan Medicare Supplemental Database |
| MDCD | IBM MarketScan Multi-State Medicaid Database |
| OHDSI | Observational Health Data Science and Informatics |
| OMOP | Observational Medical Outcomes Partnership |

## Responsible Parties

### Investigators

| Investigator | Institution/Affiliation |
| --- | --- |
| Jenna Reps \* | Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA |
| Ross Williams \* | Erasmus University |
| Solomon Ioannou | Erasmus University |
| Egill Fridgeirsson | Erasmus University |
| Anthony Sena | Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA |
| Other contributors | cohort definition creators |

### Disclosures

This study is undertaken within Observational Health Data Sciences and Informatics (OHDSI), an open collaboration. **JMR** is an employee of Janssen Research and Development and shareholders in Johnson & Johnson.

# Abstract

**Background and Significance**

Predictive analytics in healthcare hold immense promise to aid medical decision making. Creating a set of reference points against which novel methods can be evaluated, provides a transparent and accountable environment to ensure consistency in fostering improvements, innovation and refinement of modeling techniques, propelling the field of clinical prediction towards heightened accuracy and efficacy.

**Study Aims**

1. Identify a set of candidate prediction models to include in the study.
2. Following state-of-the-science model development methods, compare performance of combinations of features.

**Study Description** For explicit probem definitions with target and outcome populations, and time-at-risk, see [Table 1](#tbl-problems).

* Design: Retrospecive cohort with external validation
* Covariates
  1. Demographics (age in 5 year buckets + gender)
  2. Demographics + Conditions
  3. Demographics + Drugs
  4. Demographics + Conditions + Drugs
* Models: LASSO Logistic Regression, Ridge
* Internal Metrics:
  + Area Under the receiver-operator Curve (AUC).
  + Calibration plots.
  + Likelihood ratio for model comparison.
  + Net Benefit.
* External Metrics:
  + Area Under the receiver-operator Curve (AUC)
  + Calibration Plots (pre and post recalibration)
  + Model evaluation metrics?
  + Net Benefit (pre and post recalibration)

# Amendments and Updates

| Number | Date | Section of study protocol | Amendment or update | Reason |
| --- | --- | --- | --- | --- |
| None |  |  |  |  |

# Milestones

| Milestone | Planned / actual date |
| --- | --- |
| Start of analysis |  |
| End of analysis |  |
| Results presentation |  |

# Rationale and Background

Predictive analytics hold immense promise for augmenting medical decision making in domains such as diagnostic precision, prognostic insights, and treatment strategies. As the demand for data-driven approaches in healthcare intensifies, the imperative to maintain statistical and model quality is pronounced. Evaluating a predictive model, usually requires comparisons against a benchmark. To this end, 1) identifying a set of clinical predictions problems that hold some clinical plausibility, 2) standardizing processes like sample selection, data curation, model specification (TAR), and 3) recording predictive performance, can aid in fostering continuous development and improvement of prediction models. We refer to the aforementioned process as benchmarking and suggest that it is a vital step towards transparency and accountability, that can ultimately form a pillar for continuous improvements in prediction models.  
A standard set of benchmarks can aid in question dealing with, but not limited to:  
1) Performance evaluation: Comparisons between model performance can be done where parameters that affect model performance are standardized, e.g. population specification, population samples, data collection, time-at-risk, etc. Comparing a novel method against a benchmark can provide structured means to make direct comparisons under a framework that experimental conditions are standardized. 2) Best practice research: There is often a set of decisions that need to be made when developing a predictive model. Best practices can be evaluated against a benchmark with known performance. 3) Clinical utility: Evaluating models through benchmarking delineates their performance in real-world scenarios, offering critical insights that empower clinicians in making informed decisions regarding the integration and utilization of predictive models in clinical practice.  
By creating reference points against which novel models are juxtaposed, researchers and clinicians gain a metric for evaluating the performance of their models, nurturing an environment conducive to iterative refinement and ensuring consistency in the evaluation and comparison of predictive models across diverse datasets and clinical contexts.

Related/previous work:

ADD

# Study Objectives

* Create a set of clinically relevant and diverse prediction models
* What is the impact on performance amongst feature combinations from OMOP-CDM tables?

Specific aims:

ADD

# Research Methods

## Target Populations, Outcome Population and Time-at-risk

New names:  
• `` -> `...1`

Table 1: Clinical problems to include in the study.

| No | Problem Spec | Cohort Name\_ Target | Cohort Name\_ Outcome |
| --- | --- | --- | --- |
| 1 | Heart Failure in Patients with T2DM who were treated with metformin and who became new adult users of one of sulfonylureas, thiazolidinediones, dipeptidyl peptidase-4 inhibitors, glucagon-like peptide-1 receptor agonists, or sodium-glucose co-transporter-2 inhibitors (date of secondary drug). Patients with HF or patients treated with insulin on or prior to the index date were excluded from the analysis. Patients were required to have been enrolled for at least 365 days before cohort entry. with TAR: 1 to 365 days | [IPEV]T2dm\_second\_line | [IPEV] Heart\_Failure\_\_revised\_no\_hosp |
| 2 | Colorectal cancer in Patients with anemia (date of first anemia record) with TAR: 1 day - 1-year | [PL]Anemia | [PL]ClrctlCancer |
| 3 | Cardiovascular event and death in Patients with COPD (first record) with TAR: 1-30 days | [PL]COPD | [PL]CVDeath |
| 4 | Cardiovascular event and death in Patients with COPD (first record) with TAR: 1-90 days | [PL]COPD | [PL]CVDeath |
| 5 | Lung cancer in Patients with an outpatient visit in 2017 with no prior cancer (first visit in 2017) with TAR: 1 day - 3 years after index | [PL]Non\_urgent\_outpatient\_visit | lung\_cancer |
| 6 | Earliest of AMI cardiac arrest or death (MACE) in Patients undergoing elective major non-cardiac surgery (date of surgery) with Creatinine measurements. Patients were required to have been enrolled for at least 365 days before cohort entry. with TAR: 0 to 30 days | [RCRI] Composite MNCS 004.2 with Cr | [RCRI] Earliest of AMI cardiac arrest or death |
| 7 | Earliest of AMI cardiac arrest or death (MACE) in Patients undergoing elective major non-cardiac surgery (date of surgery) without creatinine measurements. Patients were required to have been enrolled for at least 365 days before cohort entry. with TAR: 0 to 30 days | [RCRI] Composite MNCS 004.2 | [RCRI] Earliest of AMI cardiac arrest or death |
| 8 | Hospitalization with pneumonia in Patients with an outpatient visit and a positive COVID test with TAR: 1 day - 30 days after index | [seek cover] covid outpatient visit | [seek cover] hospitalization with pneumonia |
| 9 | Hospitalization with pneumonia that required intensive services (ventilation, intubation, tracheotomy, or extracorporeal membrane oxygenation) or death in Patients with an outpatient visit and a positive COVID test with TAR: 1 day - 30 days after index | [seek cover] covid outpatient visit | [seek cover] hosp with broad pneumonia and intensive care or death |
| 10 | Death in Patients with an outpatient visit and a positive COVID test with TAR: 1 day - 30 days after index | [seek cover] covid outpatient visit | [seek cover] death |
| 11 | Ischemic stroke in Patients newly diagnosed with atrial fibrilation (date of initial afib record) with TAR: 1 to 365 days | afib | non-hem-stroke |
| 12 | Bipolar in Patients newly diagnosed with major depressive disorder (date of first record) with TAR: 1 day - 365 day after index | mdd | bipolar |
| 13 | Dementia in Patients with an outpatient visit in 2019 with TAR: 1 day - 3 years after index | dementiaTarget | dementiaOutcome |

### Cohort Definitions

* Cohort definition authors

## Candidate Covariates

* Demographics.
* Conditions.
* Drugs

## Modelling

In this study we will model the data using the following methods:

* Standard LASSO model

## Sampling

We will consider a sample of the target cohort for each problem. [[**John2022?**](#ref-John2022)] has shown through learning curves that they tend to reach a plateau above a certain number of positive event counts (report results from the study). The number of event counts required may vary for each problem (mean, s.d.) however, a stratified sample of the original population maintaining the outcome rate with adequate event counts, is adequate to capture the variability in the data and generate similar out-of-sample c-statistic as using the whole population. Use of a larger sample size to enhance discriminative performance comes with small gains in discrimination for the cost of longer running times.

The sample size for each problem will be determined by first estimating the study population for each problem (e.g. ensure outcome occurred within the specified time-at-risk). From this population we will estimate the outcome rate (outcome cases/all people in target cohort) and use this cohort to draw a stratified sample with a pre-specified number of outcomes included. For the datasets that outcomes are less than the required number of positive events will be included as a whole.

## Statistical Analysis Methods

* How predictors were handled
* Type of model, model building procedures, internal validation method.
* Measures for model performance evaluation
* Measures for model comparison

## Metrics

We will evaluate the performance by calculating the area under the receiver operating characteristic curve for overal discrimination and Eavg metric for calibration. We will also calculate the area under the precision recall curve and net benefit plots. Threshold dependent performance will be displayed using the probability threshold plot.

Performance will be plotted as a function of number of features in the model to observed the relationship between simplicity and performance.

## Data sources

We will execute the study as an OHDSI network study. All data partners within OHDSI are encouraged to participate voluntarily and can do so conveniently, because of the community’s shared Observational Medical Outcomes Partnership (OMOP) common data model (CDM) and OHDSI tool-stack.

Many OHDSI community data partners have already committed to participate and we will recruit further data partners through OHDSIs standard recruitment process, which includes protocol publication on OHDSIs GitHub, an announcement in OHDSIs research forum, presentation at the weekly OHDSI all-hands-on meeting and direct requests to data holders.

Table @ref(tab:data-sources) lists the 6 already committed data sources; these sources encompass a large variety of practice types and populations. For each data source, we report a brief description and size of the population it represents. All data sources will receive institutional review board approval or exemption for their participation before executing the study.

Table 2: Committed data sources and the populations they cover.

| Data source | Population | Patients | History | Data capture process and short description |
| --- | --- | --- | --- | --- |
| IBM MarketScan Commercial Claims and Encounters (CCAE) | Commercially insured, < 65 years | 142M | 2000 – | Adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) from large employers and health plans who provide private healthcare coverage to employees, their spouses and dependents. |
| IBM MarketScan Medicare Supplemental Database (MDCR) | Commercially insured, 65 years | 10M | 2000 – | Adjudicated health insurance claims of retirees with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service or capitated health plans. |
| IBM MarketScan Multi-State Medicaid Database (MDCD) | Medicaid enrollees, racially diverse | 26M | 2006 – | Adjudicated health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims. |
| Optum Clinformatics Data Mart (Optum) | Commercially or Medicare insured | 85M | 2000 – | Inpatient and outpatient healthcare insurance claims. |
| Optum Electronic Health Records (OptumEHR) | US, general | 93M | 2006 – | Clinical information, prescriptions, lab results, vital signs, body measurements, diagnoses and procedures derived from clinical notes using natural language processing. |
| Integrated Primary Care Information (IPCI) | Netherlands, general practitioner | 2.5M | 2006 – | General practitioner data including demographic information, patient complaints and symptoms, diagnoses, lab results, lifestyle factors, referral notes from consultants, and hospital admissions. |

# Strengths and Limitations

## Strengths

* We follow the PatientLevelPrediction framework for developing and evaluating models to ensure standard practices are applied
* Our standardized framework enables us to externally validate our models across a large number of external databases
* The fully specified study protocol is being published before analysis begins.
* All analytic methods have previously been verified on real data.
* All software is freely available as open source.
* Use of a common data model allows extension of the experiment to future databases and allows replication of these results on licensee databases that were used in this experiment, while still maintaining patient privacy on patient-level data.

## Limitations

* We are unable to test our models in a clinical setting
* In a clinically setting predictors may be self reported and these may differ from our covariate definitions
* The electronic health record databases may be missing care episodes for patients due to care outside the respective health systems.

# Protection of Human Subjects

This study does not involve human subjects research. The project does, however, use de-identified human data collected during routine healthcare provision. All data partners executing the study within their data sources will have received institutional review board (IRB) approval or waiver for participation in accordance to their institutional governance prior to execution (see Table @ref(tab:irb)). This study executes across a federated and distributed data network, where analysis code is sent to participating data partners and only aggregate summary statistics are returned, with no sharing of patient-level data between organizations.

Table 3: IRB approval or waiver statement from partners.

| Data source | Statement |
| --- | --- |
| IBM MarketScan Commercial Claims and Encounters (CCAE) | New England Institutional Review Board and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research. |
| IBM MarketScan Medicare Supplemental Database (MDCR) | New England Institutional Review Board and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research. |
| IBM MarketScan Multi-State Medicaid Database (MDCD) | New England Institutional Review Board and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research. |
| Japan Medical Data Center (JMDC) | New England Institutional Review Board and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research. |
| Optum Clinformatics Data Mart (Optum) | New England Institutional Review Board and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research. |
| Optum Electronic Health Records (OptumEHR) | New England Institutional Review Board and was determined to be exempt from broad IRB approval, as this research project did not involve human subject research. |

# Plans for Disseminating and Communicating Study Results

Open science aims to make scientific research, including its data process and software, and its dissemination, through publication and presentation, accessible to all levels of an inquiring society, amateur or professional [[**Woelfle2011-ss?**](#ref-Woelfle2011-ss)] and is a governing principle of this study. Open science delivers reproducible, transparent and reliable evidence. All aspects of this study (except private patient data) will be open and we will actively encourage other interested researchers, clinicians and patients to participate. This differs fundamentally from traditional studies that rarely open their analytic tools or share all result artifacts, and inform the community about hard-to-verify conclusions at completion.

## Transparent and re-usable research tools

We will publicly register this protocol and announce its availability for feedback from stakeholders, the OHDSI community and within clinical professional societies. This protocol will link to open source code for all steps to generate and evaluate prediction models, figures and tables. Such transparency is possible because we will construct our studies on top of the OHDSI toolstack of open source software tools that are community developed and rigorously tested [[**Schuemie2020-wx?**](#ref-Schuemie2020-wx)]. We will publicly host the source code at ADD, allowing public contribution and review, and free re-use for anyone’s future research.

## Scientific meetings and publications

We will deliver multiple presentations at scientific venues and will also prepare multiple scientific publications for clinical, informatics and statistical journals.

## General public

We believe in sharing our findings that will guide clinical care with the general public. We will use social-media (Twitter) to facilitate this.