**Patient-Level Prediction: Developing and validating models *predicting fatality in patients hospitalized for COVID-19***

Prepared on: 2020-10-01

Created by: Jenna Reps

**Table of Contents**

# List of Abbreviations

| Abbreviation | Phrase |
| --- | --- |
| AUC | Area Under the Receiver Operating Characteristic Curve |
| CDM | Common Data Model |
| O | Outcome Cohort |
| OHDSI | Observational Health Data Sciences &amp; Informatics |
| OMOP | Observational Medical Outcomes Partnership |
| T | Target Cohort |
| TAR | Time at Risk |

# Responsible Parties

Jenna Reps, Sara Khalid, Cynthia Yang, Clair Blacketer, Peter Rijnbeek, Seng Chan You, Chungsoo Kim, Talitia Duarte Salles, Sergio Fernandez-Bertolin

# Executive Summary

The objective of this study is to develop and validate patient-level prediction models that can predict the risk of death at the point a patient is hospitalized with COVID-19 using their claims or medical records prior to index. The study is demonstrating the OHDSI pipeline that can be followed to develop credible, transparent and reproducible prediction models.

We will develop models in the largest COVID database that and then externally validate the model across suitable databases in the OHDSI network. Based on best practices, we will train multiple classifiers: LASSO Logistic Regression, Random forest, Gradient boosting machine, AdaBoost, DecisionTree, Neural network.

We will use the OHDSI skeleton packages, dev: https://github.com/OHDSI/SkeletonPredictionStudy and val: https://github.com/OHDSI/SkeletonPredictionValidationStudy, that wrap around the PatientLevelPrediction R package (<https://github.com/OHDSI/PatientLevelPrediction>) to transparently perform the analyses. These the populated skeleton packages will be shared online at <https://github.com/ohdsi-studies>. This enables others to reproduce the analysis on their data (model development and validation).

# Rational & Background

COVID-19 is a novel disease that has impacted the whole world. Observational data are readily collected and may be analyzed to extract insight into the disease. However, care needs to be taken when using observational data, as identifying people with a specific disease is often not a trivial task and mistakes are often made. It is well known that junk into a classifier will result in junk coming out.

OHDSI have been developing a pipeline for analyses such as developing prediction models using observational data. The pipeline includes various checks that aim to check that the design is correctly answering the desired prediction question. These checks include understanding the quality of the data and how suitable it is for the prediction question and ensuring the analysis phenotypes appear valid and transport across numerous datasets.

In this study, we will demonstrate the full pipeline by developing and validating models that can predict a patient’s risk of death at the point she is hospitalized for COVID-19. We will show that it is important to investigate the validity of phenotypes, especially ones for novel illnesses, such as COVID-19 as this novel disease may be coded different across databases and the code practices within one database may be very unstable overtime. Failure to correctly identify patients with COVID-19 or death would likely impact the clinical utility of a prediction model.

# Objective

The objective is to use the full OHDSI pipeline (Data Quality Dashboard, Problem definition, Protocol Development, Phenotype Development, Machine learning, Standardized evaluation) to develop and validate patient-level prediction models for the following prediction problem:

| Target Cohorts | Outcome Cohorts | Time at Risk |
| --- | --- | --- |
| [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | [Time at Risk Settings #1] Risk Window Start: 0, Add Exposure Days to Start: FALSE, Risk Window End: 30, Add Exposure Days to End: FALSE |

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

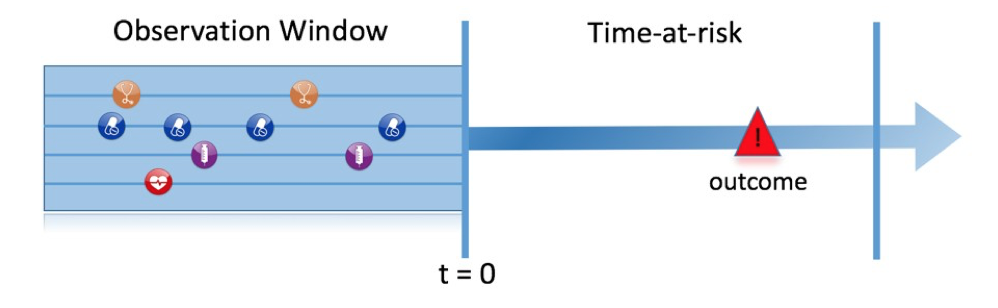


Figure 1: The prediction problem

*Citation:*

We follow the PROGRESS best practice recommendations for model development and the TRIPOD guidance for transparent reporting of the model results.

*Citation: Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. doi: 10.1371/journal.pmed.1001381. Epub 2013 Feb 5. Review. PubMed PMID: 23393430; PubMed Central PMCID: PMC3564751.*

*Citation: Collins, G., et al. (2017.02.01). 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.' from https://www.equator-network.org/reporting-guidelines/tripod-statement/*

## Data Source(s)

We will use the following databases mapped to the OMOP CDM:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Database full name | Database acronym | Country | Data type | Time period covered |
| Optum© De-Identified Clinformatics® Data Mart Database | Optum DOD | USA | Claims | 2000 - present  COVID-19: January 2020 - present |
| Optum© De-identified Electronic Health Record Dataset | Optum EHR | USA | EHR | 2006- present  COVID-19: January 2020 - present |
| The Information System for Research in Primary Care | SIDIAP | Spain | GP and EHR linked | 2007 – present  COVID-19: January 2020 - present |
| Health Insurance and Review Assessment | HIRA | South Korea | Claims | 2015 - present  COVID-19: January 2020 - present |

## Study Populations

### Target Cohort(s) [T]

| Cohort ID | Cohort Name | Description |
| --- | --- | --- |
| 17780 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | TBD |

Cohort Entry Events

People may enter the cohort when observing any of the following:

1. visit occurrences of 'Inpatient Visit', starting after December 1, 2019.

Restrict entry events to with any of the following criteria:

1. having at least 1 measurement of 'SARS-CoV-2 positive test measurement pre-coordinated', starting 21 days before cohort entry start date and starting anytime on or before cohort entry end date.
2. having at least 1 measurement of 'SARS-CoV-2 test measurement', starting 21 days before cohort entry start date and starting anytime on or before cohort entry end date; with value as concept: "detected", "detected", "positive", "positive", "present" or "present".
3. having at least 1 observation of 'SARS-CoV-2 test measurement', starting 21 days before cohort entry start date and starting anytime on or before cohort entry end date; with value as concept: "detected", "detected", "positive", "positive", "present" or "present".
4. having at least 1 condition occurrence of 'COVID-19 conditions', starting 21 days before cohort entry start date and starting anytime on or before cohort entry end date.

Limit these restricted entry events to the earliest event per person.

Inclusion Criteria

1. has >=365 prior observation

Entry events having at least 1 observation period, starting anytime up to 365 days before cohort entry start date and ending between 0 days before and all days after cohort entry start date.

### Outcome Cohorts(s) [O]

| Cohort ID | Cohort Name | Description |
| --- | --- | --- |
| 16741 | [COVID ID164 V1] Death | TBD |

Cohort Entry Events

People enter the cohort when observing any of the following:

1. death of any form.

Limit cohort entry events to the earliest event per person.

### Time at Risk

| Time at Risk |
| --- |
| [Time at Risk Settings #1] Risk Window Start: 0, Add Exposure Days to Start: FALSE, Risk Window End: 30, Add Exposure Days to End: FALSE |

### Additional Population Settings

***Population Settings #1***

| Item | Settings |
| --- | --- |
| minTimeAtRisk | 364 |
| requireTimeAtRisk | FALSE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 0 |
| washoutPeriod | 365 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 99999 |
| binary | TRUE |
| removeSubjectsWithPriorOutcome | TRUE |
| riskWindowEnd | 30 |
| firstExposureOnly | TRUE |

## Statistical Analysis Method(s)

### Algorithms

| Algorithm | Description |
| --- | --- |
| AdaBoost | AdaBoost is a boosting ensemble technique. Boosting works by iteratively adding decision trees but adds more weight to the data-points that are misclassified by prior decision trees in the cost function when training the next tree. We use the sklearn 'AdaboostClassifier' implementation in Python. |
| DecisionTree | A decision tree is a classifier that partitions the variable space using individual tests selected using a greedy approach. It aims to find partitions that have the highest information gain to separate the classes. The decision tree can easily overfit by enabling a large number of partitions (tree depth) and often needs some regularization (e.g., pruning or specifying hyper-parameters that limit the complexity of the model). We use the sklearn 'DecisionTreeClassifier' implementation in Python. |
| Gradient boosting machine | Gradient boosting machines is a boosting ensemble technique and in our framework it combines multiple decision trees. Boosting works by iteratively adding decision trees but adds more weight to the data-points that are misclassified by prior decision trees in the cost function when training the next tree. We use Extreme Gradient Boosting, which is an efficient implementation of the gradient boosting framework implemented in the xgboost R package available from CRAN. |
| 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 |
| Neural network | Neural networks contain multiple layers that weight their inputs using an non-linear function. The first layer is the input layer, the last layer is the output layer the between are the hidden layers. Neural networks are generally trained using feed forward back-propagation. This is when you go through the network with a data-point and calculate the error between the true label and predicted label, then go backwards through the network and update the linear function weights based on the error. This can also be performed as a batch, where multiple data-points are feed through the network before being updated. We use the sklearn 'MLPClassifier' implementation in Python. |
| Random forest | Random forest is a bagging ensemble technique that combines multiple decision trees. The idea behind bagging is to reduce the likelihood of overfitting, by using weak classifiers, but combining multiple diverse weak classifiers into a strong classifier. Random forest accomplishes this by training multiple decision trees but only using a subset of the variables in each tree and the subset of variables differ between trees. Our packages uses the sklearn learn implementation of Random Forest in python. |

### 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 quantile 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 tradeoff 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 color 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.

*Citation:*

## Tools

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

*Citation: R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*

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. The full table can be found below in the 'Table and Figures' section under 'Characterization'.

# Data Analysis Plan

## Algorithm Settings

***Model Settings Settings #1 - LassoLogisticRegressionSettings***

| Covariates | Settings |
| --- | --- |
| seed |  |
| variance | 0.01 |

***Model Settings Settings #2 - RandomForestSettings***

| Covariates | Settings |
| --- | --- |
| maxDepth | 4, 10, 17 |
| varImp | TRUE |
| seed |  |
| mtries | -1 |
| ntrees | 1000 |

***Model Settings Settings #3 - GradientBoostingMachineSettings***

| Covariates | Settings |
| --- | --- |
| maxDepth | 4, 6, 17 |
| minRows | 20, 10, 5 |
| seed |  |
| nthread | 20 |
| ntrees | 10, 100, 500 |
| learnRate | 0.01, 0.1, 0.9, 0.98 |

***Model Settings Settings #4 - AdaBoostSettings***

| Covariates | Settings |
| --- | --- |
| nEstimators | 50, 10, 100 |
| seed |  |
| learningRate | 1, 0.9, 0.1 |

***Model Settings Settings #5 - DecisionTreeSettings***

| Covariates | Settings |
| --- | --- |
| maxDepth | 10, 5 |
| minSamplesLeaf | 10, 2 |
| seed |  |
| minImpurityDecrease | 1e-07 |
| classWeight | None |
| plot | FALSE |
| minSamplesSplit | 2, 10 |

***Model Settings Settings #6 - MLPSettings***

| Covariates | Settings |
| --- | --- |
| seed |  |
| size | 4, 40, 20 |
| alpha | 1e-05, 0.1, 0.9, 0.005 |

## 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 | FALSE |
| ConditionGroupEraStartLongTerm | FALSE |
| ConditionEraShortTerm | FALSE |
| Dcsi | FALSE |
| DrugGroupEraLongTerm | TRUE |
| DrugGroupEraShortTerm | FALSE |
| ConditionEraStartLongTerm | FALSE |
| temporal | FALSE |
| DemographicsIndexMonth | FALSE |
| 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 | FALSE |
| DistinctConditionCountLongTerm | FALSE |
| MeasurementValueLongTerm | FALSE |
| DrugEraShortTerm | FALSE |
| DrugGroupEraAnyTimePrior | TRUE |
| DrugEraOverlapping | FALSE |
| ConditionOccurrencePrimaryInpatientAnyTimePrior | FALSE |
| ConditionEraMediumTerm | FALSE |
| ConditionEraOverlapping | FALSE |
| ConditionEraStartShortTerm | FALSE |
| ObservationAnyTimePrior | TRUE |
| 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 | TRUE |
| DemographicsRace | FALSE |
| ProcedureOccurrenceAnyTimePrior | TRUE |
| DistinctObservationCountLongTerm | FALSE |
| ProcedureOccurrenceMediumTerm | FALSE |
| CharlsonIndex | FALSE |
| DemographicsPriorObservationTime | FALSE |
| MeasurementShortTerm | FALSE |
| DistinctProcedureCountMediumTerm | FALSE |
| ConditionEraLongTerm | FALSE |
| DrugGroupEraStartLongTerm | FALSE |
| DemographicsGender | TRUE |
| DeviceExposureAnyTimePrior | TRUE |
| 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 | TRUE |
| Chads2 | FALSE |
| DrugExposureAnyTimePrior | FALSE |
| DeviceExposureLongTerm | FALSE |
| DemographicsTimeInCohort | FALSE |
| DistinctMeasurementCountMediumTerm | FALSE |
| MeasurementValueShortTerm | FALSE |
| DeviceExposureMediumTerm | FALSE |
| ConditionGroupEraStartShortTerm | FALSE |
| ConditionOccurrencePrimaryInpatientMediumTerm | FALSE |
| MeasurementLongTerm | TRUE |
| DemographicsIndexYear | FALSE |
| MeasurementValueMediumTerm | FALSE |
| DrugEraStartMediumTerm | FALSE |
| MeasurementValueAnyTimePrior | FALSE |
| DistinctObservationCountShortTerm | FALSE |
| DrugEraMediumTerm | FALSE |
| ConditionGroupEraLongTerm | TRUE |
| DrugExposureShortTerm | FALSE |
| DistinctIngredientCountShortTerm | FALSE |
| DeviceExposureShortTerm | FALSE |
| mediumTermStartDays | -180 |
| DemographicsPostObservationTime | FALSE |
| VisitConceptCountLongTerm | FALSE |
| VisitConceptCountMediumTerm | FALSE |
| excludedCovariateConceptIds |  |
| ConditionGroupEraMediumTerm | FALSE |
| DrugExposureMediumTerm | FALSE |
| DistinctProcedureCountLongTerm | FALSE |
| DrugEraAnyTimePrior | FALSE |
| endDays | -1 |
| ConditionOccurrenceShortTerm | FALSE |

***Covariate Settings #2***

| Covariates | Settings |
| --- | --- |
| VisitCountMediumTerm | FALSE |
| ObservationShortTerm | FALSE |
| shortTermStartDays | -30 |
| MeasurementRangeGroupShortTerm | FALSE |
| ConditionOccurrenceLongTerm | FALSE |
| DrugEraStartLongTerm | FALSE |
| VisitCountShortTerm | FALSE |
| Chads2Vasc | FALSE |
| ConditionGroupEraStartLongTerm | FALSE |
| ConditionEraShortTerm | FALSE |
| Dcsi | FALSE |
| DrugGroupEraLongTerm | FALSE |
| DrugGroupEraShortTerm | FALSE |
| ConditionEraStartLongTerm | FALSE |
| temporal | FALSE |
| DemographicsIndexMonth | FALSE |
| 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 | FALSE |
| 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 | FALSE |
| 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 1 target cohorts evaluated for 1 outcomes over 6 models over 2 covariates settings and over 1 population settings. In total there are 12 analysis performed. For a full list refer to appendix 'Complete Analysis List'.

# Strengths & Limitations

Strengths:

* Using the OHDSI network to perform external validation
* The pipeline records all settings making it possible to reproduce the analysis
* The OHDSI tools: Data Quality Dashboard and CohortDiagnostic investigate the phenotypes across the network
* We are training multiple classifiers
* We use skeleton analysis packages – this reduces time it takes to write the codes as it is mostly done and reduces errors are the code has already been tested
* We use the PLP package – this has test cases to ensure it does what is expected

Limitations

* Observational data can have unknown bias (which is why it is important to run analyses across multiple databases)
* Phenotype sensitivity/specificity is generally unknown

# Protection of Human Subjects

Confidentiality of patient records will be maintained always. All study reports will contain aggregate data only and will not identify individual patients or physicians. At no time during the study will the sponsor receive patient identifying information except when it is required by regulations in case of reporting adverse events.

# Plans for Disseminating & Communicating Study Results

Full results will be published at data.ohdsi.org and a paper will be written and submitted to a high impact journal. A conference may also be targeted.

# Tables & Figures

## Incidence Rate of Target & Outcome

*<< add incidence here. >>*

## Characterization

*<< add characterization table here. >>*

*<< add results here. >>*

# Appendices

## Study Generation Version Information

Skeleton Version: PatientLevelPredictionStudy - v0.0.1

Identifier / Organization: Janssen Research and Development

## Code List

***Concept Set #1 - SARS-CoV-2 positive test measurement pre-coordinated***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 37310282 | 1240581000000104 | 2019 novel coronavirus detected | Measurement | Standard | Clinical Finding | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 17780 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation |

***Concept Set #2 - SARS-CoV-2 test measurement***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 756055 | OMOP4873969 | Measurement of severe acute respiratory syndrome coronavirus 2 | Measurement | Standard | Procedure | V | FALSE | TRUE |
| Valid | 37310281 | 1240591000000102 | 2019 novel coronavirus not detected | Measurement | Standard | Clinical Finding | V | TRUE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 17780 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation |

***Concept Set #3 - COVID-19 conditions***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 37311061 | 840539006 | Disease caused by severe acute respiratory syndrome coronavirus 2 | Condition | Standard | Clinical Finding | V | FALSE | TRUE |
| Valid | 4100065 | 27619001 | Disease due to Coronaviridae | Condition | Standard | Clinical Finding | V | FALSE | TRUE |
| Valid | 439676 | 186747009 | Coronavirus infection | Condition | Standard | Clinical Finding | V | FALSE | TRUE |
| Valid | 37311060 | 840544004 | Suspected disease caused by severe acute respiratory coronavirus 2 | Observation | Standard | Context-dependent | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 17780 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation |

***Concept Set #4 - Inpatient Visit***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 262 | ERIP | Emergency Room and Inpatient Visit | Visit | Standard | Visit | V | FALSE | TRUE |
| Valid | 9201 | IP | Inpatient Visit | Visit | Standard | Visit | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 17780 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation |

## Complete Analysis List

Below is a complete list of analysis that will be performed. Definitions for the column 'Covariate Settings ID' can be found above in the 'Covariate Settings' section. Definitions for the 'Population Settings Id' can be found above in the 'Additional Population Settings' section.

| ID | Target Cohort Name | Outcome Cohort Name | Model Settings Id | Model Settings Description | Covariate Settings ID | Population Settings ID |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 1 | Lasso Logistic Regression | 1 | 1 |
| 2 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 2 | Random forest | 1 | 1 |
| 3 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 3 | Gradient boosting machine | 1 | 1 |
| 4 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 4 | AdaBoost | 1 | 1 |
| 5 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 5 | DecisionTree | 1 | 1 |
| 6 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 6 | Neural network | 1 | 1 |
| 7 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 1 | Lasso Logistic Regression | 2 | 1 |
| 8 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 2 | Random forest | 2 | 1 |
| 9 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 3 | Gradient boosting machine | 2 | 1 |
| 10 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 4 | AdaBoost | 2 | 1 |
| 11 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 5 | DecisionTree | 2 | 1 |
| 12 | [COVID ID134 V1] Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation | [COVID ID164 V1] Death | 6 | Neural network | 2 | 1 |

*<< add models here >>*

# References

*<< To be completed outside of ATLAS. >>*