Study Protocol

Development and validation of patient-level prediction models for hospitalization and death amongst young patients presenting with a clinical diagnosis or positive test for COVID-19: a rapid network study to inform the management of COVID-19

Created by:

Sara Khalid, PhD, University of Oxford, UK
Jenna Reps, PhD, Janssen Research and Development
Daniel Prieto-Alhambra, MD PhD, University of Oxford, UK
Patrick Ryan, PhD, Columbia University, USA
Edward Burns, PhD, University of Oxford, UK
Peter Rijnbeek, PhD, Erasmus MC, Rotterdam, The Netherlands

Prepared on: 23 September 2020

Contact Person: Sara Khalid – sara.khalid@ndorms.ox.ac.uk

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

The authors declare the following disclosures: Jenna Reps, PhD is employee of Janssen Research and Development.

Table of Contents

1. List of Abbreviations

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

2. Executive Summary

The objective of this study is to develop and validate patient-level prediction models for patients younger than 50 years old who visit a general practitioner (GP), the emergency room (ER), or other outpatient care (OP) with a clinical diagnosis of Covid-19 or positive test of Covid-19. and who had no symptoms of pneumonia 60 days prior to the visit.

Four different outcomes are predicted, including, 1) hospitalizations with pneumonia, 2) hospitalizations with pneumonia or ARDS, sepsis, or AKI, 3) hospitalizations with pneumonia or ARDS, sepsis, or AKI requiring intensive services or resulting in death, 4) patient mortality. All with a time of risk of 30 days from the initial visit. These four prediction models will be implemented using Lasso Logistic Regression.

3. Rationale & Background

General description of COVID-19

The Corona Virus Disease 2019 (COVID-19), which started in late 2019 as an epidemic in Wuhan, Hubei Province, China, has been declared a pandemic and a public health emergency of international concern by the World Health Organization (WHO) in January 2020 (1). The growing number of infections by COVID-19 has resulted in an unprecedented pressure on healthcare systems worldwide, and a large number of casualties at a global scale. Diagnosis of COVID-19 currently relies on the detection of SARS-CoV-2 nucleic acid (2), but no medical treatment or vaccine is available yet. Common symptoms presented by patients include fever, cough, and dyspnea, signaling the onset of pneumonia (3). Although the majority of people have uncomplicated or mild illness (81%), some will develop severe illness requiring hospitalization and oxygen support (14%) or intensive care unit treatment (5%) (4).

Problem definition

Countries around the world have begun to experience a second wave of Covid-19, and this wave is accompanied by an over-representation of younger people amongst those infected. The current WHO Risk Communication Guidance distinguishes two distinct categories of patients at high risk of

severe disease: those older than 60 years and those with "underlying medical conditions" which is non-specific (5). There is little to no information on the risk categories in younger patients, aged 50 years or less. Early identification of younger patients who will require hospital care or are at high risk of death will ensure these patients have the best chance of receiving optimal care and surviving. Further, early intervention can reduce the severity of symptoms and as such reduce the resources required for each patient. Moreover, reducing hospital admissions that are not strictly necessary avoids burden on the already stressed healthcare system and prevents unnecessary medical interventions.
Previous work has been done to develop and validate models for predicting the risks of the aforementioned outcomes in the overall population, based on data representing the first wave. It is expected that predictors of the outcomes in a younger population in the second wave may be different, hence new models should be developed and validated for this cohort.

Study aims

The objective of this study is to inform the triage and early management of patients with clinically diagnosed or positive-tested COVID-19 by developing and validating patient-level prediction models. In particular, we aim to 1) identify adult patients aged 50 years or younger who are at risk of hospitalization or death after presenting for the first time with a clinical diagnoses or positive-test of COVID-19 at a GP/OP or ER visit.

Clinical use case

These models identify the short-term risk of hospitalization and death due to secondary infections amongst young patients with clinically diagnosed or positive-tested COVID-19. There is an over-representation of young people amongst those affected in the upcoming/ongoing second wave, whereas most model developed during and after the first wave are based on data from older patients; secondly predictors of outcomes amongst younger patients are likely to be different to those for older patients.

Description of previous literature

We reviewed previous literature on pneumonia/ARDS severity prediction. Most papers studying outcomes for patients presenting with flu, flu-like symptoms or pneumonia focused on mortality, admission to intensive care units and other adverse outcomes (e.g. septic shock or mechanical ventilation need) while hospitalized. Most of these previous studies focused on target cohorts of patients already admitted to hospital. Additionally, we reviewed recent COVID-19 studies that considered prediction modelling (8).

We identified one study that is similar to our first prediction problem (9). To date, this seems to be the only study with similar inclusion criteria of flu or flu-like symptoms in adults without a diagnosis of pneumonia or hospitalization in the target cohort definition. Moreover, the outcome cohort of this study is also similarly defined as hospital admission or readmission without a focus on mortality or ICU admission. There are some differences in the study design as the prediction is made in a target cohort of severe influenza patients presenting at the hospital emergency department, predicting inclusion in the outcome cohort of hospital admission/readmission. The AUROC was 0.84 and the key discriminators identified by decision tree classification were underlying illness, age, vaccination history, and influenza viral load. The study was run from EHR data in a small cohort of 184 patients. The authors suggest the model can be used for further investigation of possible hospitalization of patients with confirmed influenza in the ER.

There does not seem to be an available model to predict risk of hospitalization in younger patients due to secondary infections for COVID-19 patients. The closest such model is the COVID-19

Vulnerability Index (10) built from a 5% sample of Medicare claims data from 2015-2016 (1.85M people), using a proxy for COVID-19: hospitalization in patients diagnosed with pneumonia (except when caused by tuberculosis), influenza, acute bronchitis, or other specified upper respiratory infections. The model performs with an AUROC of 0.731 and has not been validated externally either against a COVID-19 cohort nor against data with an available hospitalization outcome.

How the study is performed

In this study we aim to identify, based on the medical history prior to the first encounter (GP/OP or ER visit), which patients with a clinical diagnosis or positive test of Covid-19 after presenting for the first time are likely going to need hospitalization due to secondary infections. Known complications in COVID-19 patients include hospitalization due to pneumonia, ARDS, sepsis or acute kidney injury (4).

We will develop patient-level prediction models using the Observational Health Data Sciences and Informatics (OHDSI) Patient-Level Prediction framework. The OHDSI collaboration is a network of researchers working towards a common goal of standardizations and best practice frameworks for analysing observational data in healthcare. The OHDSI collaboration relies on researchers mapping their datasets into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). This, along with the Patient-Level Prediction framework, allows for rapid model development and validation following accepted best practices (11), which provides a unique opportunity to make a difference in the current crisis. We will train the patient-level prediction models across databases in the OHDSI collaborator network and perform external validation of each model across the OHDSI network. The main advantage of our approach is that we have access to some COVID-19 data and will externally validate the models on recent COVID-19 data to evaluate whether the models transport to this similar patient population.

4. Objective

The objective is to develop and validate patient-level prediction models for hospitalization amongst adults aged 50 years or less, with a clinical diagnosis or positive test for Covid-19:

1) To predict the 30-day risk of hospitalization due to secondary infections (pneumonia, ARDS, sepsis or acute kidney injury) amongst adults aged 50 years or less, with a clinical diagnosis or positive test for Covid-19 after presenting at a GP/OP or ER visit for the first time.
2) To predict the 30-day risk of death due to secondary infections (pneumonia, ARDS, sepsis or acute kidney injury) amongst adults aged 50 years or less, with a clinical diagnosis or positive test for Covid-19 after presenting at a GP/OP or ER visit for the first time.

Target Cohorts	Outcome Cohorts	Time at Risk
Cohort #1270 [COVID PLP training] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID25 V1] Hospitalizations with pneumonia	[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
Cohort #1270 [COVID PLP training] Persons with COVID without	[COVID19 ID26 V1] Hospitalizations with	[Time at Risk Settings #1] Risk Window Start: 0, Add Exposure Days to Start:

inpatient or intensive services, >365d prior observation and <50yo	pneumonia or ARDS or sepsis or AKI	FALSE, Risk Window End: 30, Add Exposure Days to End: FALSE
Cohort #1270 [COVID PLP training] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID27 V1] Hospitalizations with pneumonia or ARDS or sepsis or AKI requiring intensive services or resulting in death in 30d	[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
Cohort #1270 [COVID PLP training] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID28 v1] persons who die	[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
Cohort #1271 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID25 V1] Hospitalizations with pneumonia	[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
Cohort #1271 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID26 V1] Hospitalizations with pneumonia or ARDS or sepsis or AKI	[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
Cohort #1271 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID27 V1] Hospitalizations with pneumonia or ARDS or sepsis or AKI requiring intensive services or resulting in death in 30d	[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
Cohort #1271 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	[COVID19 ID28 v1] persons who die	[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
Cohort #1273 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <30yo	[COVID19 ID25 V1] Hospitalizations with pneumonia	[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
Cohort #1273 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <30yo	[COVID19 ID26 V1] Hospitalizations with pneumonia or ARDS or sepsis or AKI	[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
Cohort #1273	[COVID19 ID27 V1] Hospitalizations with	[Time at Risk Settings #1] Risk Window Start: 0, Add

[COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <30yo	pneumonia or ARDS or sepsis or AKI requiring intensive services or resulting in death in 30d	Exposure Days to Start: FALSE, Risk Window End: 30, Add Exposure Days to End: FALSE
Cohort #1273 [COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <30yo	[COVID19 ID28 v1] persons who die	[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

5. Methods

5.1.Study Design

This study will follow a retrospective, observational, patient-level prediction design. We define 'retrospective' to mean the study will be conducted using data already collected prior to the start of the study. We define 'observational' to mean there is no intervention or treatment assignment imposed by the study. We define 'patient-level prediction' as a modelling process wherein an outcome is predicted within a time at risk relative to the target cohort start and/or end date. Prediction is performed using a set of covariates derived using data prior to the start of the target cohort.

Figure 1 illustrates the prediction problem we will address. Among a population at risk, we aim to predict which patients at a defined moment in time (t = 0) will experience some outcome during a time-at-risk (TAR). Prediction is done using only information about the patients in an observation window prior to that moment in time.

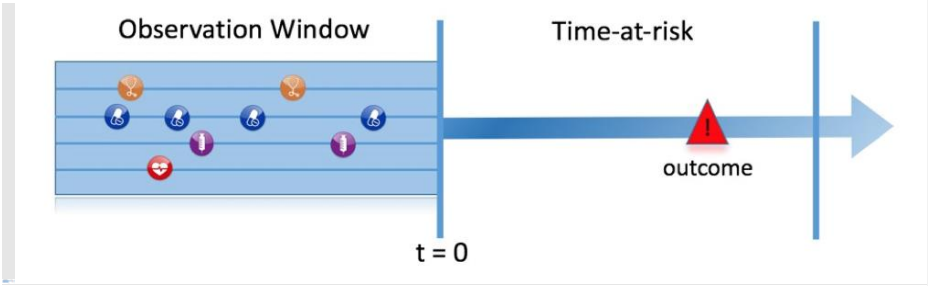

Figure 1: The prediction problem

We follow the PROGRESS best practice recommendations for model development and the TRIPOD guidance for transparent reporting of the model results (12, 13).

5.2. Data Source(s)

Commented [SK1]: This has to be discussed and updated

Source Full Name	Country Code	Data Provenance	Source Short Name	Patient Count	History	Patient Type	Data collection
Optum® de-identified Electronic Health Record Dataset	US	EMR	Optum EHR - EMR, US	96m	2006-	EHR / Privately Insured	Optum© de-identified Electronic Health Record Dataset represents Humedica's Electronic Health Record data a medical records database. The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using Natural Language Processing (NLP).

5.3.Study Populations

5.3.1. Target Cohort(s) [T]

Cohort ID	Cohort Name	Description
1270	[COVID PLP training] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	Patients with either a covid19 diagnosis OR positive test, AND no inpatient or intensive care in prior 30d AND no visit within 30d prior or on diagnosis
1271	[COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	Patients with either a covid19 diagnosis OR positive test, AND no inpatient or intensive care in prior 30d
1273	[COVID PLP training] Persons with COVID without inpatient or intensive services, >365d prior observation and <30yo	Patients with either a covid19 diagnosis OR positive test, AND no inpatient or intensive care in prior 30d AND no visit within 30d prior or on diagnosis

5.3.2. Validation Cohorts

Cohort ID	Cohort Name	Description
1270	[COVID PLP training] Persons with COVID without inpatient or intensive	Patients with either a covid19 diagnosis OR positive test, AND no inpatient or intensive care in

	services, >365d prior observation and <50yo	prior 30d AND no visit within 30d prior or on diagnosis
1271	[COVID PLP training v2] Persons with COVID without inpatient or intensive services, >365d prior observation and <50yo	Patients with either a covid19 diagnosis OR positive test, AND no inpatient or intensive care in prior 30d
1273	[COVID PLP training] Persons with COVID without inpatient or intensive services, >365d prior observation and <30yo	Patients with either a covid19 diagnosis OR positive test, AND no inpatient or intensive care in prior 30d AND no visit within 30d prior or on diagnosis

5.3.3. Outcome Cohorts(s) [O]

Cohort ID	Cohort Name	Description
5889	[COVID19 ID27 V1] Hospitalizations with pneumonia or ARDS or sepsis or AKI requiring intensive services or resulting in death in 30d	TBD
5890	[COVID19 ID28 v1] persons who die	All Cause mortality
5892	[COVID19 ID25 V1] Hospitalizations with pneumonia	
5893	[COVID19 ID26 V1] Hospitalizations with pneumonia or ARDS or sepsis or AKI	TBD

Full descriptions:

The JSON files describing for all the outcome cohorts are available at:

- https://github.com/ohdsi-studies/Covid19PredictionStudies/tree/master/HospitalizationInSymptomaticPatients/inst/cohorts
- https://github.com/ohdsi-studies/Covid19PredictionStudies/tree/master/HospitalizationInSentHomePatients/inst/cohorts

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

5.3.4. Time at Risk

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

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

5.3.5. Additional Population Settings

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

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

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

Population Settings #1

Item	Settings
minTimeAtRisk	364
requireTimeAtRisk	FALSE
addExposureDaysToStart	FALSE

riskWindowStart	0
washoutPeriod	365
addExposureDaysToEnd	FALSE
includeAllOutcomes	TRUE
priorOutcomeLookback	99999
binary	TRUE
removeSubjectsWithPriorOutcome	FALSE
riskWindowEnd	30
firstExposureOnly	FALSE

Population Settings #2

Item	**Settings**
minTimeAtRisk	364
requireTimeAtRisk	FALSE
addExposureDaysToStart	FALSE
riskWindowStart	2
washoutPeriod	365
addExposureDaysToEnd	FALSE
includeAllOutcomes	TRUE
priorOutcomeLookback	99999
binary	TRUE
removeSubjectsWithPriorOutcome	FALSE
riskWindowEnd	30
firstExposureOnly	FALSE

5.4.Statistical Analysis Method(s)

5.4.1. Algorithms

In this study we will apply a Lasso Logistic Regression. Lasso logistic regression belongs to the family of generalized linear models, where a linear combination of the variables is learned and finally a logistic function maps the linear combination to a value between 0 and 1. The lasso regularization adds a cost based on model complexity to the objective function when training the model. This cost is the sum of the absolute values of the linear combination of the coefficients. The model automatically performs feature selection by minimizing this cost. We use the Cyclic coordinate descent for logistic, Poisson and survival analysis (Cyclops) package to perform large-scale regularized logistic regression: https://github.com/OHDSI/Cyclops.

5.4.2. Model Evaluation

The following evaluations will be performed on the model:

Evaluation	Description
Box Plots	The prediction distribution boxplots are box plots for the predicted risks of the people in the test set with the outcome (class 1: blue) and without the outcome (class 0: red).
Calibration Plot	The calibration plot shows how close the predicted risk is to the observed risk. The diagonal dashed line thus indicates a perfectly calibrated model. The ten (or fewer) dots represent the mean predicted values for each quantile plotted against the observed fraction of people in that quantile who had the

	outcome (observed fraction). The straight black line is the linear regression using these 10 plotted quantiles mean predicted vs observed fraction points. The two blue straight lines represented the 95% lower and upper confidence intervals of the slope of the fitted line.
Demographic Summary Plot	This plot shows for females and males the expected and observed risk in different age groups together with a confidence area.
Precision Recall Plot	The precision-recall curve is valuable for dataset with a high imbalance between the size of the positive and negative class. It shows the trade-off between precision and recall for different threshold. High precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). A high area under the curve represents both high recall and high precision.
Prediction Distribution Plots	The preference distribution plots are the preference score distributions corresponding to i) people in the test set with the outcome (red) and ii) people in the test set without the outcome (blue).
ROC Plot	The ROC plot plots the sensitivity against 1-specificity on the test set. The plot shows how well the model is able to discriminate between the people with the outcome and those without. The dashed diagonal line is the performance of a model that randomly assigns predictions. The higher the area under the ROC plot the better the discrimination of the model.
Smooth Calibration Plot	Similar to the traditional calibration shown above the Smooth Calibration plot shows the relationship between predicted and observed risk. the major difference is that the smooth fit allows for a more fine-grained examination of this. Whereas the traditional plot will be heavily influenced by the areas with the highest density of data the smooth plot will provide the same information for this region as well as a more accurate interpretation of areas with lower density. the plot also contains information on the distribution of the outcomes relative to predicted risk. However, the increased information game comes at a computational cost. It is recommended to use the traditional plot for examination and then to produce the smooth plot for final versions.
Test-Train Similarity Plot	The test-train similarity is presented by plotting the mean covariate values in the train set against those in the test set for people with and without the outcome.
Variable Scatter Plot	The variable scatter plot shows the mean covariate value for the people with the outcome against the mean covariate value for the people without the outcome. The size and colour of the dots correspond to the importance of the covariates in the trained model (size of beta) and its direction (sign of beta with green meaning positive and red meaning negative), respectively.

5.5.Quality Control

The PatientLevelPrediction package itself, as well as other OHDSI packages on which PatientLevelPrediction depends, use unit tests for validation. More information can be found in the Book of OHDSI at: https://ohdsi.github.io/TheBookOfOhdsi/SoftwareValidity.html

5.6.Tools

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

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

This study will be designed using OHDSI tools and run with R (1). More information about the tools can be found in the Appendix 'Study Generation Version Information'.

6. 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.

7. Data Analysis Plan

7.1.Algorithm Settings

Model Settings Settings #1 - LassoLogisticRegressionSettings

Covariates	Settings
seed	
variance	0.01

7.2.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	TRUE
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	TRUE
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	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

7.3.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. The external validity of the models will be assessed on recent COVID-19 data. We will use the area under the receiver operating characteristic curve (AUROC) 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.

7.4.Analysis Execution Settings

For the first prediction model there is 1 target cohort evaluated for 4 outcomes over 1 model over 2 covariates settings and over 1 population setting. For the second prediction model there are 2 target cohorts evaluated for 2 outcomes over 1 model over 2 covariates settings and over 1 population setting. In total there are 16 analyses performed.

8. Strengths & Limitations

Strength

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

Limitations

- The external validation datasets may not have a sufficient number of (some of) the different outcomes to be used in the analysis.
- Although the CDM standardizes the vocabularies of the datasets, the concept recording distributions are likely to differ between databases and it is unknown how much this will limit model transportability.

9. Protection of Human Subjects

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

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

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

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

10. Plans for Disseminating & Communicating Study Results

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

11. Tables & Figures

11.1. Incidence Rate of Target & Outcome

Feasibility assessment training data

T	O	TAR start	TAR end	T size	O count	O incidence

Feasibility assessment validation data

Database	O	TAR	T size	O count	O incidence

12. Appendices

12.1. Study Generation Version Information

Skeleton Version: PatientLevelPredictionStudy - v0.0.1
Identifier / Organization: OHDSI

13. References

1. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19. 2020 March 11 [cited 2020 March 28]. Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.
2. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. JAMA. 2020.
3. Yang X, Yu Y, Xu J, Shu H, Liu H, Wu Y, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. The Lancet Respiratory Medicine. 2020.
4. World Health Organization. Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020. Geneva: World Health Organization; 2020.

5. World Health Organization. Coronavirus disease 2019 (COVID-19) Situation Report – 51 2020 March 11 [cited 2020 March 28]. Available from: ttps://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_4.
6. Heymann DL, Shindo N. COVID-19: what is next for public health? The Lancet. 2020;395(10224):542-5.
7. Ranieri VM, Pettilä V, Karvonen MK, Jalkanen J, Nightingale P, Brealey D, et al. Effect of Intravenous Interferon β-1a on Death and Days Free From Mechanical Ventilation Among Patients With Moderate to Severe Acute Respiratory Distress Syndrome: A Randomized Clinical Trial. JAMA. 2020;323(8):725-33.
8. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection. medRxiv. 2020:2020.03.24.20041020.
9. Chen K-F, Hsieh Y-H, Gaydos CA, Valsamakis A, Rothman RE. Derivation of a clinical prediction rule to predict hospitalization for influenza in EDs. The American journal of emergency medicine. 2013;31(3):529-34.
10. DeCaprio D, Gartner J, Burgess T, Kothari S, Sayed S. Building a COVID-19 Vulnerability Index. arXiv preprint arXiv:200307347. 2020.
11. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association. 2018 Aug;25(8):969-75.
12. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
13. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
14. Team RC. R: A language and environment for statistical computing. 2013.