Computer Methods and Programs in Biomedicine 211 (2021) 106394

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data

Sara Khalid[a,1], Cynthia Yang[b,1], Clair Blacketer[c], Talita Duarte-Salles[d], Sergio Fernández-Bertolín[d], Chungsoo Kim[e], Rae Woong Park[e,f], Jimyung Park[e], Martijn J. Schuemie[c], Anthony G. Sena[b,c], Marc A. Suchard[g], Seng Chan You[h], Peter R. Rijnbeek[b,2], Jenna M. Reps[c,2,1,*]

[a] Botnar Research Centre, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, UK.
[b] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands
[c] Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA
[d] Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
[e] Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea
[f] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[g] Departments of Biomathematics, University of California, Los Angeles, USA
[h] Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, Republic of Korea

ARTICLE INFO

Article history:
Received 18 March 2021
Accepted 30 August 2021

Keywords:
COVID-19
Data harmonization
Data quality control
Distributed data network
Machine learning
Risk prediction

ABSTRACT

Background and objective: As a response to the ongoing COVID-19 pandemic, several prediction models in the existing literature were rapidly developed, with the aim of providing evidence-based guidance. However, none of these COVID-19 prediction models have been found to be reliable. Models are commonly assessed to have a risk of bias, often due to insufficient reporting, use of non-representative data, and lack of large-scale external validation. In this paper, we present the Observational Health Data Sciences and Informatics (OHDSI) analytics pipeline for patient-level prediction modeling as a standardized approach for rapid yet reliable development and validation of prediction models. We demonstrate how our analytics pipeline and open-source software tools can be used to answer important prediction questions while limiting potential causes of bias (*e.g.*, by validating phenotypes, specifying the target population, performing large-scale external validation, and publicly providing all analytical source code).

Methods: We show step-by-step how to implement the analytics pipeline for the question: 'In patients hospitalized with COVID-19, what is the risk of death 0 to 30 days after hospitalization?'. We develop models using six different machine learning methods in a USA claims database containing over 20,000 COVID-19 hospitalizations and externally validate the models using data containing over 45,000 COVID-19 hospitalizations from South Korea, Spain, and the USA.

Results: Our open-source software tools enabled us to efficiently go end-to-end from problem design to reliable Model Development and evaluation. When predicting death in patients hospitalized with COVID-19, AdaBoost, random forest, gradient boosting machine, and decision tree yielded similar or lower internal and external validation discrimination performance compared to L1-regularized logistic regression, whereas the MLP neural network consistently resulted in lower discrimination. L1-regularized logistic regression models were well calibrated.

Conclusion: Our results show that following the OHDSI analytics pipeline for patient-level prediction modelling can enable the rapid development towards reliable prediction models. The OHDSI software tools and pipeline are open source and available to researchers from all around the world.

* Corresponding author.
E-mail address: jreps@its.jnj.com (J.M. Reps).

[1] These authors contributed equally to this work as co-first authors.
[2] These authors contributed equally to this work as co-last authors.

https://doi.org/10.1016/j.cmpb.2021.106394

1. Introduction

THE COVID-19 pandemic continues to cause unprecedented pressure on healthcare systems worldwide, and many casualties at a global scale [1]. Due to the urgency of the COVID-19 pandemic there was increased pressure to efficiently develop COVID-19 prediction models. Unfortunately, model reliability was often compromised in order to rapidly develop models. Despite guidelines on developing and reporting of prediction models [2], there are several common problems identified in published COVID-19 prediction models including uncertain data quality, unclear target setting, lack of large-scale external validation, and insufficient reporting [3–6].

This motivates the need for a standardized approach for rapid yet reliable development and validation of prediction models, one that allows researchers to address various sources of bias. For instance, such an approach should ensure that the data used are representative of the population for which the developed prediction model is intended to be used in clinical practice, and that the quality of the phenotypes used is investigated and transparently reported.

Observational Health Data Sciences and Informatics (OHDSI) is an international, multi-stakeholder collaboration that has developed open-source solutions for large-scale analytics [7]. The OHDSI community has used these open-source solutions to generate observational evidence for COVID-19 and has impacted international clinical guidelines [8–22].

In this paper, we demonstrate the OHDSI analytics pipeline for patient-level prediction modeling (henceforth also referred to as "pipeline" or "prediction pipeline") as a standardized approach for reliable and rapid development and validation of prediction models. We show that our pipeline makes it possible to develop prediction models rapidly without compromising model reliability. The main contributions of our work are summarized as follows:

(1) *Reliable and rapid research.*

OHDSI implements a distributed data network strategy where no patient-level data are shared. Instead, the analytical source code is shared publicly, run by data partners on their data, and only aggregated results are shared with the study coordinator. Such a strategy has proven to enable international collaborative research while providing various advantages [23]. Key advantages of OHDSI's distributed data network strategy that are particularly relevant for the current pandemic are:

- *Reliable research* The use of open-source software tools and publicly shared analytical source code, along with extensive documentation makes studies conducted with the same analysis within this distributed data network fully *reproducible* (*i.e.*, same data, same results), as well as *replicable* (*i.e.*, similar data, similar results).
- *Rapid research* To improve the interoperability of originally heterogenous observational data sources (*e.g.*, electronic healthcare records (EHRs), administrative claims), they are mapped to a common data model (CDM). The use of an established CDM enables standardized approaches for data curation and enables standardized analytics pipelines to generate results much faster [14,24,25]. In addition, the data standardization enables the ability to externally validate models at scale to investigate how reliable the models are across different case-mixes.

(2) *Analysis of COVID-19 data from multiple countries around the world.*

In March 2020, the OHDSI community began contributing to generating observational evidence for COVID-19 with data from around the world. By November 2020, there were 22 databases (including EHRs, administrative claims, primary and secondary care databases) in the OHDSI network that incorporated COVID-19 data - 11 from North America, 8 from Europe, and 3 representing Asia-Pacific (Fig. 1). In total, the mapped data included:

- 7.4 million patients tested for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).
- 1.6 million patients diagnosed or tested positive for COVID-19.
- 300,000 patients hospitalized with COVID-19.

We describe each stage of the prediction pipeline in the following section. We then demonstrate the use of the prediction pipeline for the problem of predicting COVID-19 death with results presented in section III. This work was reviewed by the New England Institutional Review Board (IRB) and was determined to be exempt from board IRB approval, as this research project did not involve human subject research.

2. Methods

OHDSI provides a library of open-source software tools for the process of developing and validating prediction models using observational data. The OHDSI analytics pipeline for patient-level prediction modeling that we demonstrate in this work can be summarized as follows (Fig. 2). As a required initial check before a database is added to the distributed data network and included in a study, *Data Harmonization and Quality Control:* source data

USA (11)	EUROPE (8)	ASIA-PACIFIC (3)
Columbia University (NY – EHR)	CPRD (UK – EHR)	HIRA (South Korea – Administrative Claims)
Department of Veterans Affairs (National – EHR)	DA Germany (Germany – EHR)	DCMC (South Korea – EHR)
HealthVerity (Claims linked to diagnostic testing)	HM Hospitales (Spain – Hospital Billing)	Nanfang Hospital (China – EMR)
IQVIA Open Claims (National – Administrative Claims)	IPCI (Netherlands – EHR)	
Optum EHR (National – EHR)	LPD France (France – EHR)	
Optum SES (National – EHR linked to Socio-economic data)	LPD Italy (Italy – EHR)	
Premier (National – Hospital Billing)	SIDIAP (Spain – EHR)	
Stanford University (CA – EHR)	SIDIAP-H (Spain – EHR Hospital linkage)	
Tufts University (MA – EHR)		
University of Colorado Anschutz Medical Campus (CO – EHR)		
University of Washington Medicine COVID Research Dataset (WA – EHR)		

Together, OHDSI has studied:
- ***>7.4m** patients tested for SARS-CoV-2*
- ***>1.6m** patients diagnosed or tested positive for COVID-19*
- ***>300k** patients hospitalized with COVID-19*

Fig. 1. The OHDSI distributed data network. As of November 2020, it includes 22 sites spread across North America, Europe, and Asia that have COVID-19 patient data mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).

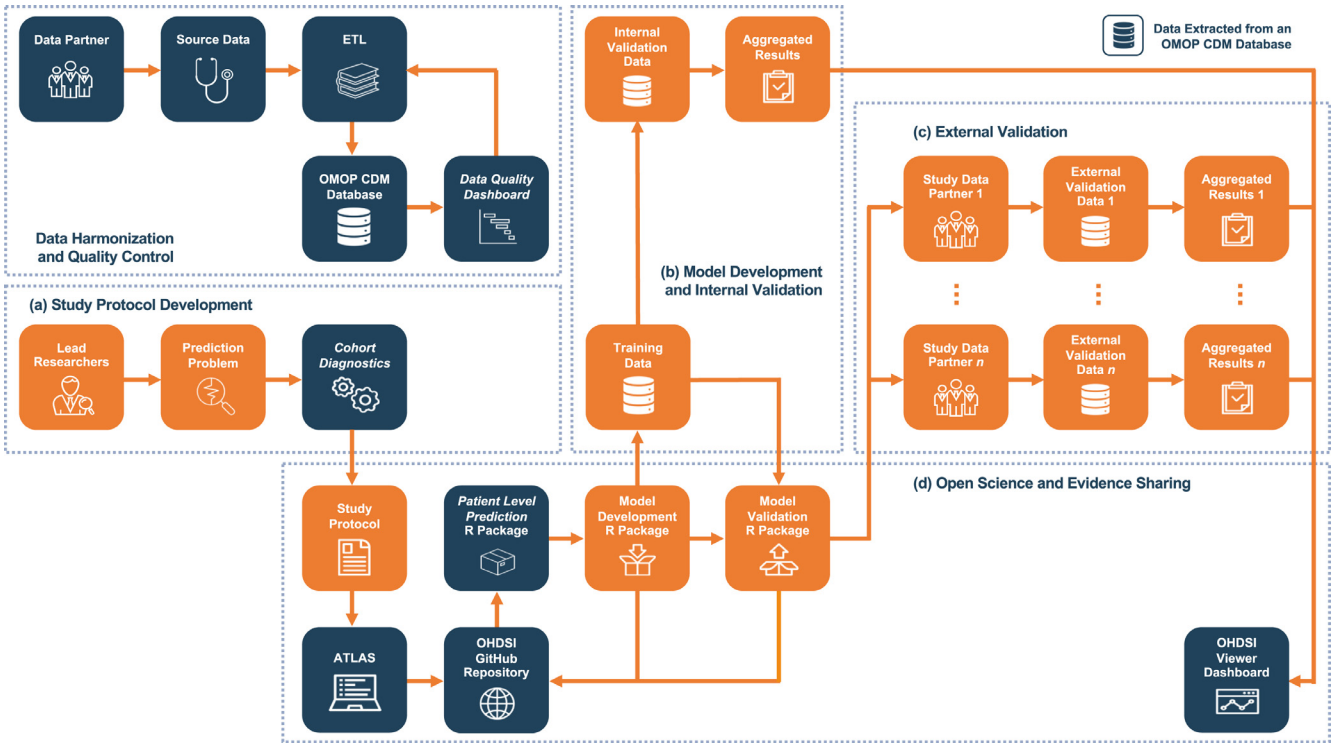

Fig. 2. An overview of the OHDSI analytics pipeline for patient-level prediction modelling. Orange boxes represent study-specific input or output, blue boxes represent non-study-specific input, output, or OHDSI software tools.

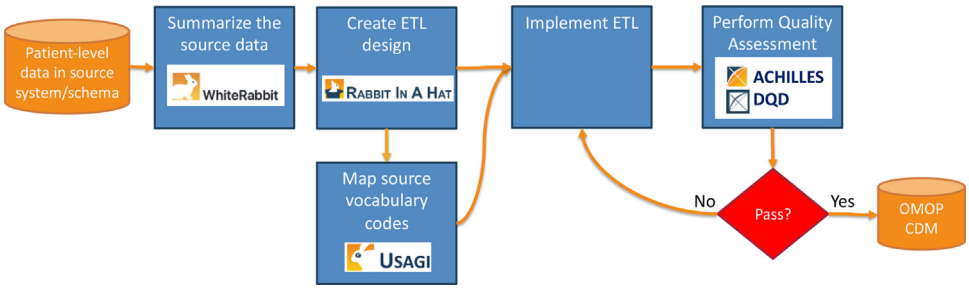

Fig. 3. The step-by-step process for mapping data sources to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), using OHDSI software tools. ETL: Extraction, Transformation and Load; DQD: Data Quality Dashboard.

are harmonized to the Observational Medical Outcomes Partnership (OMOP) CDM structure and coding system using an Extraction, Transformation and Load (ETL) design specification and quality control is performed. To conduct a prediction study, the following steps can be followed. *(a) Study Protocol Development*: we first develop a study protocol by specifying the prediction problem, assessing phenotypes using the OHDSI CohortDiagnostics tool, and specifying machine learning settings; then *(b) Model Development and Internal Validation*: we develop and internally validate the prediction models using the 'Model Development' *R* package, a wrapper of the OHDSI PatientLevelPrediction *R* package, that we generate via a user-friendly website interface called ATLAS; after which *(c) External Validation:* we distribute the automatically generated 'Model Validation' *R* package to participating data partners for external validation of the developed models; finally *(d) Open Science and Evidence Sharing:* we disseminate our collected results on the OHDSI Viewer Dashboard. All documentation including the study protocol, generated *R* packages, and OHDSI software tools, are publicly available on GitHub. In the rest of this section, we describe each stage of the pipeline in detail.

2.1. Data harmonization and quality control

OHDSI uses the OMOP CDM which transforms source data into a common format using a set of common terminologies, vocabularies, and coding schemes [26]. To support the ETL of the source data to the CDM, the OHDSI community has developed several open-source software tools (Fig. 3).

First, the WhiteRabbit tool produces a scan that summarizes every table, column, and value in a given source dataset [27]. This profiling step is important to understand the complexity of the source data.

Second, the Rabbit-in-a-hat tool is an application for interactive design of an ETL to the OMOP CDM [28]. It reads the WhiteRabbit scan report and displays a graphical user interface containing all the source and OMOP CDM tables which need to be connected by the user. The final product is a design specification document that is then used to guide the implementation.

Third, the Usagi tool supports the mapping of source vocabularies to the OMOP standardized vocabularies [29]. Based on the ETL design specification that is created in Rabbit-in-a-hat the

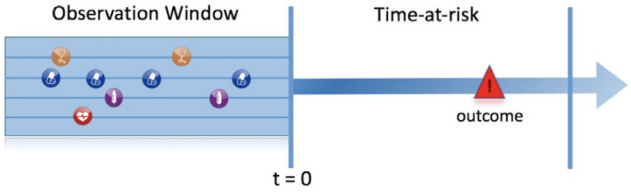

Fig. 4. Prediction problem specification in OHDSI.

code is written to transform the source data into the OMOP CDM format.

Finally, the Data Quality Dashboard (DQD) and the ACHILLES characterization tool are used to interrogate the quality of the resulting OMOP CDM-mapped dataset [30]. The DQD uses a systematic approach to run and evaluate over 3300 data quality checks. It assesses how well a dataset conforms with OMOP standards and how well source concepts are mapped to standard concepts. ACHILLES computes over 170 visualizations of the data and displays them in an open-source application designed to allow exploration and identification of potential anomalies and data outliers [31].

2.2. Stages of the study

(a) Study Protocol Development.

Any research team from anywhere around the world can propose a study on the OHDSI Forum (https://forums.ohdsi.org). Interested investigators co-design a study protocol. The study protocol must transparently specify the prediction problem of interest and study design choices such as sensitivity analysis, Model Development methods, and evaluation techniques. Next, the collaborators determine the feasibility of the study across the data network and the validity of the specified study design choices using various OHDSI software tools.

(1) *Specifying the prediction problem.*

The OHDSI community has standardized the prediction problem specification into three components [32], which are shown in Fig. 4:

- *The target population* This is the set of patients for whom we wish to predict the individual risk. The index date ($t = 0$) is the reference point in time for each patient in the target population. Only information from a specified observation window preceding the index date is used for engineering the candidate predictors.
- *The outcome* This is the medical condition or event we wish to predict.
- *The time-at-risk* This is a time interval on or after the index date, within which we wish to predict the outcome occurrence.

For example, if we wanted to develop a model to predict death within 30 days in patients hospitalized with COVID-19, then a suitable target population could be patients with a hospital stay who have COVID-19. The index date would be the first day of the hospital stay. The outcome would be a patient's death, and the time-at-risk would be the period between index and 30 days after index.

(2) *Generating and assessing phenotypes.*

Improving the syntactic and semantic interoperability of the data through the CDM and the standardized vocabularies does not solve all interoperability issues. For instance, data may originate from different clinical settings and have different levels of granularity. As a consequence, identifying the target population and outcome in the data can still be a challenge, even in observational data that are mapped to the OMOP CDM.

Defining an algorithm to identify patients within a database who have a certain condition or medical event is known as 'phenotyping'. In general, a phenotype can be defined as an index rule followed by inclusion/exclusion rules. The rules can use sets of OMOP-standardized concept IDs to identify certain conditions or events. For example, our target population phenotype could be defined as follows.

Patients with an inpatient visit (concept ID 9201 or 262) satisfying the following inclusion criteria:

- COVID-19 positive test (concept ID 37310282) OR COVID-19 diagnosis (concept ID 439676, 37311061, 4100065 or 37311060) during the visit,
- $\geq$ 365 days of prior observation at index.

The index date is the date of the qualifying inpatient visit.

When developing a prediction model, it is important that the target population and outcome phenotypes correctly identify the desired individuals. A suitable phenotype is one that is highly sensitive (the majority of the patients in the database with the condition or event are correctly identified) and has a high positive predictive value (the majority of the patients identified by the phenotype have the condition or event). Mis-specifying the target population and/or outcome phenotype definitions is likely to lead to poor performance when implementing the prediction model in clinical practice. A significant amount of work is required to develop suitable phenotypes and expert knowledge of a specific database is required to guide this process.

The OHDSI community has developed a process to generate and assess suitable phenotypes. This process starts with a literature review of existing phenotype definitions for the target population and outcome. Commonly used phenotypes identified by the literature review are then identified as candidate phenotypes. If no phenotypes exist in the published literature, then a clinician and data expert collaborate to propose new candidate phenotypes. The candidate phenotypes need to be assessed to determine whether they are capturing the correct patients.

The OHDSI CohortDiagnostics tool creates descriptive results for each candidate phenotype such as the characteristics of the patients identified by the phenotype, the validity of the concept ID sets, and the number of patients identified by the phenotype across calendar time [33]. This is repeated across the OHDSI network of databases. The results are then inspected by a panel of clinicians to compare the characteristics of the patients identified by the phenotype, the temporal trend of the phenotype and the number of patients identified by the phenotype across numerous databases. The specific aspects of a phenotype that are inspected are:

- *Generalizability* Do the patients captured by the phenotype appear to represent the real-world patients with the medical condition? This requires inspecting the characteristics of the patients identified by the phenotype in view of the literature and expert consensus.
- *Consistency across the network* Is the phenotype identifying patients consistently across the network or does it seem to fail for one or more database? This may indicate an issue with the transportability of the definition. Common issues include unsuitable data or incorrect concept ID sets.
- *Correctness of concept ID sets* Do we have the correct concept IDs for identifying inclusion/exclusion criteria used by the phenotype? OHDSI uses string similarity and associations to identify potential missing concept IDs.

If issues are identified with a candidate phenotype, then revisions to the phenotype are made and the process is repeated until no issues are observed.

(3) *Assessing suitability of source databases.*

Once the phenotypes are defined and validated, the next step is identifying whether each OHDSI observational database is suitable for Model Development and/or validation. This involves qualitative and quantitative assessment. If issues are identified, then other databases should be considered instead.

Initial feasibility assessment Consulting with a person who has expert knowledge of the database is important to determine any issues in the way the data are captured that may impact the phenotypes. For example, some databases lack older or younger patients or may not capture complete lab results or medication.

Databases that pass the initial feasibility assessment are then reviewed using the CohortDiagnostics tool. The results can be inspected to identify datasets that satisfy:

- *Adequate size* Is the number of patients identified by the target population phenotype in a given database sufficient for developing a prediction model? Our recent but as yet unpublished study on generating learning curves to empirically assess the sample size at which convergence towards maximum achievable performance starts shows that, typically, more than 1000 patients with the outcome are needed for Model Development [34]. For accurate Model Validation, at least 100 patients with the outcome is recommended [35].
- *Continuous observation time* Are the patients identified by the target population phenotype in a given database observed long enough to have a sufficient lookback period to capture predictors, and enough follow-up time to cover the time-at-risk? The incidence rate should be inspected for sufficient outcomes during the time-at-risk.

If there is no suitable database across the network then it may be worth exploring alternative prediction specifications (e.g., use a proxy for the target population).

(4) *Model Development settings.*

The study protocol must specify the settings used for Model Development including [32]:

- The candidate predictors to be included in the model, *e.g.*, drugs (at various ingredient and drug levels), diagnoses, procedures, measurements, as well as diagnostics and summary scores.
- The train/test split design - by default a 25% test set and 75% train set is used, where k-fold cross-validation is applied on the train set to select optimal hyper-parameters [36]. The user can choose to divide patients into the train and test sets randomly (stratified by outcome) or based on time.
- The set of classifiers to be used, including gradient boosting machine, random forest, regularized logistic regression, decision tree, AdaBoost, and multi-layer perceptron neural network.
- The hyper-parameter search per selected classifier - if using a grid search the user can specify the values to investigate [36].
- Sensitivity analysis options - whether to include patients lost to follow-up or patients who had the outcome prior to index.

(b) *Model Development and internal validation.*

All model development settings can be specified *via* a user-friendly website interface called ATLAS [7]. This includes the prediction problem components (the target population and outcome phenotypes and the time-at-risk) in addition to the above model-specific settings (*e.g.*, candidate predictor settings and model development settings).

(i) *Developing the 'Model Development' R package.*

Once the settings have been specified, ATLAS generates a study-specific open-source *R* package called the 'Model Development' *R* package. This is a wrapper of the PatientLevelPrediction *R* package [37] and can be run on any OMOP CDM database to develop and internally validate the models specified in the study protocol.

(ii) *Executing the 'Model Development' R package.*

The 'Model Development' *R* package can be implemented by providing the connection details to the CDM, the CDM database name, a database schema with read/write access that is used to create temporary tables, and the location where the log/data/model will be saved to. The output is a directory containing the extracted data, the developed models, all the settings required to replicate the study and summary information about the internal validation of the models. In addition, an *R* Shiny app is generated that displays the results interactively to the user.

(c) *External validation*

After developing the models, the 'Model Development' *R* package can automatically generate a 'Model Validation' *R* package for externally validating the models. This package contains the data extraction source code for the various settings (phenotypes, time-at-risk, and predictors) and the developed models that need to be validated. The 'Model Validation' *R* package is another wrapper of the PatientLevelPrediction *R* package that uses the stored settings to call functions to extract the data, apply the models and assess the performance using the standard evaluation metrics. Therefore, the automatically generated 'Model Validation' *R* package is able to fully replicate the data extraction process used to develop the models across any database mapped to the OMOP CDM and then applies and validates the models. Once installed, the user just points the 'Model Validation' *R* package to their OMOP CDM data, and the *R* package will execute the external validation. The output is a collection of .csv files containing evaluation metrics such as information on the discrimination and calibration of each model.

(d) *Open Science and Evidence Sharing*

All study documentation, including the study protocol and automatically generated *R* packages, are shared publicly. The 'Model Development' and 'Model Validation' *R* packages can be uploaded to the ohdsi-studies GitHub (https://github.com/ohdsi-studies) to enable any researcher to run the model development and external validation analysis on their data mapped to the OMOP CDM. Results for each of the databases participating in the study can be combined in an *R* Shiny application and then uploaded to the publicly available OHDSI Viewer Dashboard.

The open-source OHDSI software tools involved in the prediction pipeline are regularly updated and revised versions are maintained on GitHub. This allows researchers in the field to implement additional settings or methods into our proposed pipeline. The PatientLevelPrediction *R* package has a flexible model integration, making it easy for researchers to add custom machine learning models. The OHDSI Forum is open for all to join, to contribute to the development and use of software tools, and to co-create scientific questions.

2.3. COVID-19 demonstration

In this section, we demonstrate using the prediction pipeline to develop and validate a COVID-19 prediction model. We were interested in predicting a patient's risk of death within 30 days from the point they are hospitalized with COVID-19. We demonstrate how this was done using the stages of the prediction pipeline described in Section 2.2.

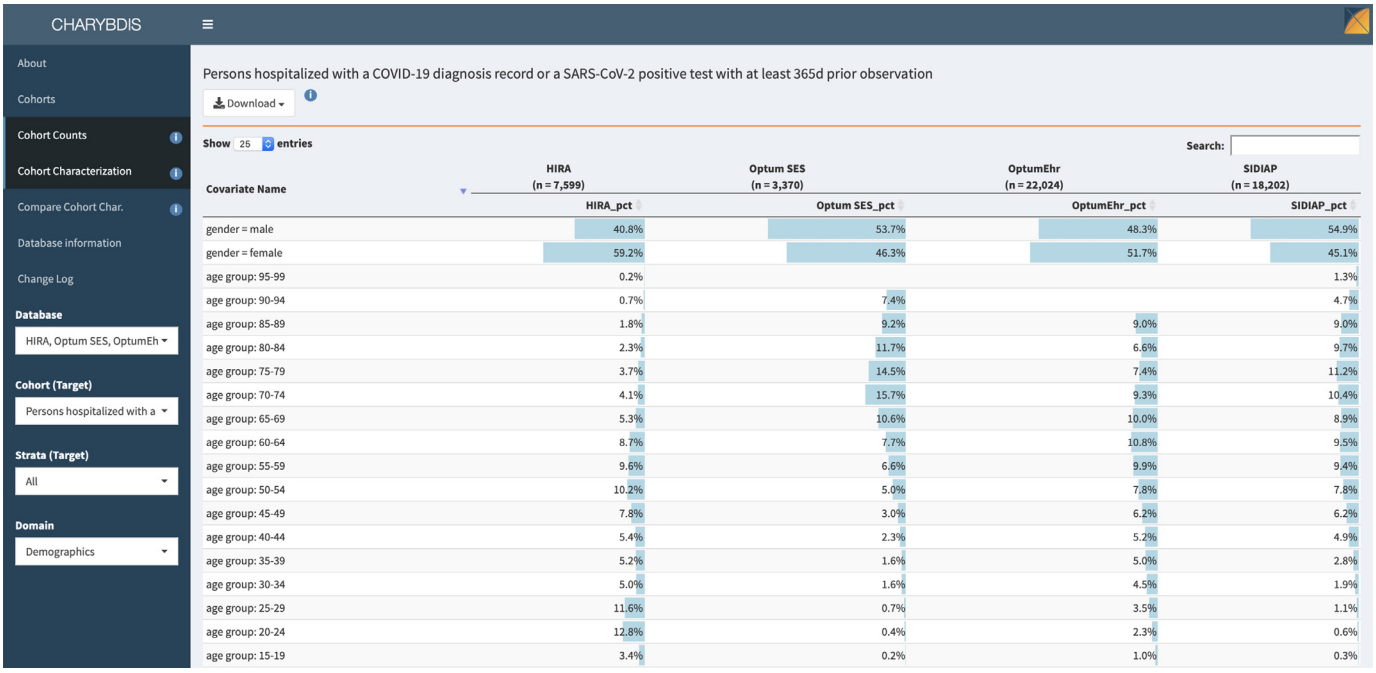

Fig. 5. A snapshot of the CohortDiagnostics tool for assessing phenotypes. Here, Optum SES refers to the Optum Claims database.

Table 1
Prediction problem specification.

Component	Specification
Target	Patients hospitalized with COVID-19
Outcome	Death
Time-at-risk	0 days to 30 days after the hospital visit

Table 2
Phenotype definitions.

Component	Phenotype definition
Target	Patients with an inpatient visit on or after December 2019 with a COVID-19 positive test or COVID-19 diagnosis within 21 days before the visit or during the visit and >=365 days prior observation.
Outcome	Death record in database

(a) Study Protocol Development.

(1) *Specifying the prediction problem.*

We studied the following prediction problem: "Within patients hospitalized with COVID-19, predict the risk of death on the hospitalization date and up to 30 days after using data recorded up to 1 day prior to hospitalization", defined in Table 1.

(2) *Generating and assessing phenotypes.*

The phenotypes used to identify 'patients hospitalized with COVID-19' and 'death' are defined in Table 2. The phenotype for 'death' was defined as any record of death in the database. The phenotype for 'patients hospitalized with COVID-19' was previously developed in a large-scale COVID-19 characteristic study detailed in https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis. The CohortDiagnostics results are available at https://data.ohdsi.org/Covid19CharacterizationCharybdis/ for the cohort 'Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365 d prior observation'. This phenotype was investigated across 16 OMOP CDM databases to ensure transportability (Fig. 5).

(3) *Assessing suitability of source databases.*

Across the OHDSI network, four OMOP CDM databases capturing death and containing inpatient visit data were identified as suitable, as per the database suitability checks (described in Section 2.2) performed using the CohortDiagnostics tool. Table 3 describes the four databases. The largest one, Optum claims, was used to develop the models and Optum EHR, HIRA-COVID, and SIDIAP were used for external validation.

(4) *Model Development settings.*

Two sets of candidate predictors were used.

- Age and gender: this set included gender and binary indicators of age in 5-year groups (40–45, 45–50, ..., 95 +). We used this set of candidate predictors to create a benchmark model.
- All: the second set included 57,627 candidate predictors including binary ones indicating the occurrences of various conditions, drugs, observations procedures or measurements, that were recorded any time prior, as well as in the year prior, to the index visit (not including day of the visit date), in addition to age and gender.

We chose a random (stratified by outcome) 75/25 train/test split with 3-fold cross-validation on the train set to select the optimal hyper-parameter settings per classifier. We trained an L1-regularized logistic regression model as the reference model, using cross-validation to select the strength of regularization. As a sensitivity analysis, we also trained Gradient Boosting Machine, decision tree, random forest, multi-layer perceptron (MLP) neural network, and AdaBoost models. The pipeline supports binary classification and survival analysis. In this demonstration we chose to use binary classification instead of survival analysis due to the short 30-day period. When predicting outcomes over longer periods of time we recommend using the Cox regression (or alternative survival models) rather than binary classification.

(b) *Model Development and internal validation.*

We developed the 'Model Development' *R* package in ATLAS (http://atlas-covid19.ohdsi.org/#/prediction/39). Within ATLAS, the

Table 3
The databases used in this research.

Database full name	Database short name	Country	Data type	Time period
Optum© De-Identified Clinformatics® Data Mart Database	Optum Claims	USA	Claims	January 2020 - June 2020
Optum© De-identified Electronic Health Record Dataset	Optum EHR	USA	EHR	January 2020 – October 2020
The Information System for Research in Primary Care	SIDIAP	Spain	Primary care EHR linked to hospital admissions	January 2020 – May 2020
Health Insurance and Review Assessment – COVID-19 database	HIRA-COVID	South Korea	Claims	January 2020 – May 2020

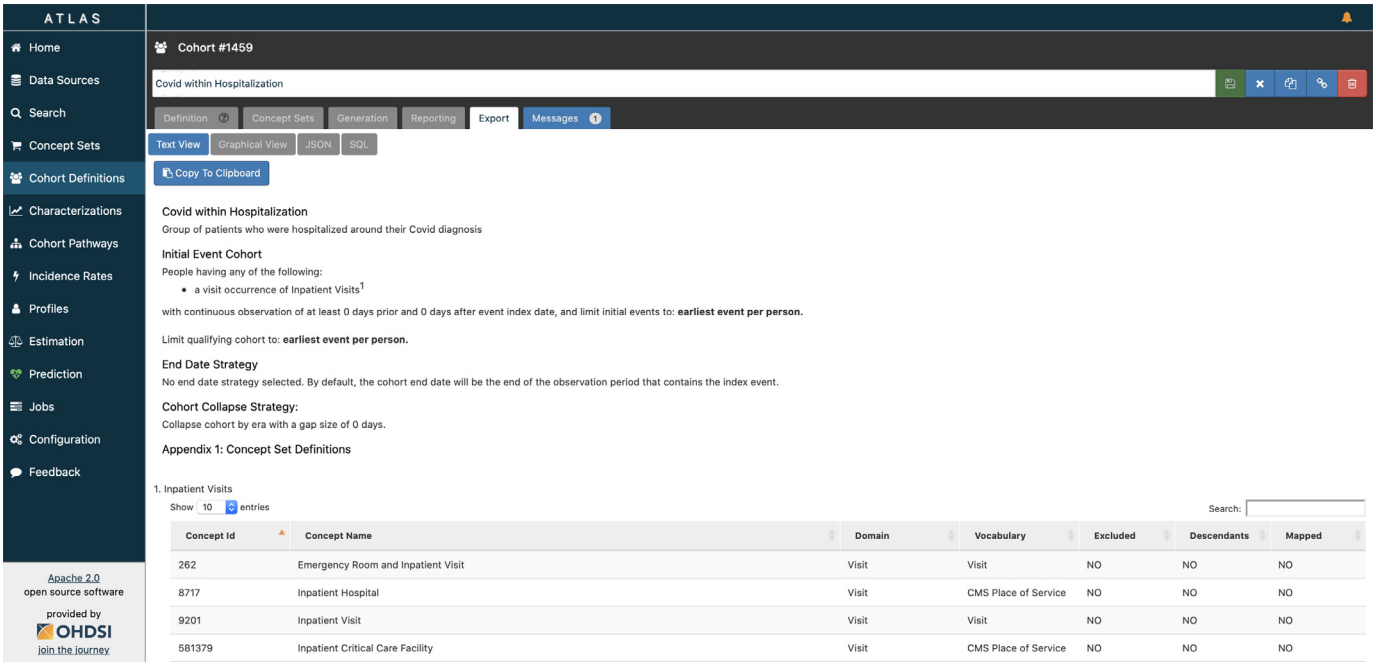

Fig. 6. A snapshot of the ATLAS tool for prediction model development.

phenotype definitions specified in Table 2 were created using the "Cohort Definitions" tab (Fig. 6). Next, the model settings were created using the "Prediction" tab. Once the prediction study was designed, the 'Model Development' *R* package was automatically generated by clicking on "Download Study Package". The 'Model Development' *R* package contains all the functionality to develop and internally validate the prediction model using OMOP CDM data.

(c) *External validation.*

The 'Model Validation' *R* package was automatically generated using the 'Model Development' *R* package.

(d) *Open science and evidence sharing.*

The protocol is available at https://github.com/ohdsi-studies/CovidDeath/blob/master/inst/doc/protocol.docx.

The 'Model Development' *R* package is available at https://github.com/ohdsi-studies/CovidDeath/tree/master/CovidDeathDev and the 'Model Validation' *R* package is available at https://github.com/ohdsi-studies/CovidDeath.

3. Results

Table 4 presents the discriminative performance of the models. Using L1-regularized logistic regression, the model including the set of all variables resulted in an internal validation AUC of 0.74 (0.72–0.76) for Optum Claims and external validation AUCs of 0.76 (0.75–0.78) for Optum EHR, 0.78 (0.77–0.78) for SIDIAP, and 0.90 (0.87–0.93) for HIRA-COVID. In comparison, the L1-regularized logistic regression model including only age and gender predictors resulted in an internal validation AUC of 0.70 (0.69–0.72) for Optum Claims, and external validation AUCs of 0.75 (0.74–0.77) for Optum EHR, 0.79 (0.78–0.80) for SIDIAP, and 0.93 (0.91–0.94) for HIRA-COVID. For both sets of candidate predictors (age and gender only, and all variables), AdaBoost, random forest, gradient boosting machine, and decision tree yielded similar or lower internal and external validation AUCs compared to L1-regularized logistic regression, whereas the MLP neural network consistently resulted in lower AUCs.

The models were well calibrated with respect to age and gender (Figs. 7 and 8).

The internal and external validation results were made publicly available in the OHDSI Viewer Dashboard at: https://data.ohdsi.org/

Table 4
Discriminative performance (measured using the area under the receiver operating characteristic curve (AUC) with a 95% confidence interval (CI)) of the different classifiers in predicting 30-day death outcome in patients hospitalized with COVID-19 (Optum claims is the internal discrimination estimated using the test set; the other databases are the external validation discrimination estimates).

Development data sample size (Outcome %)	Candidate predictors	Validation Database	Validation data sample size (Outcome %)	AUC (95% CI) L1-regularized logistic regression	AUC (95% CI) AdaBoost	AUC (95% CI) Random Forest	AUC (95% CI) Gradient Boosting Machine	AUC (95% CI) MLP Neural Network	AUC (95% CI) Decision Tree
16,991 (15.6%)	Age and gender	HIRA-COVID	6,445 (2.1)	**0.93 (0.91–0.94)**	**0.93 (0.91–0.94)**	0.80 (0.75–0.84)	0.91 (0.88–0.93)	0.57 (0.52–0.62)	0.85 (0.81–0.89)
		Optum Claims	5,663 (15.6)	**0.70 (0.69–0.72)**	**0.70 (0.69–0.72)**	0.68 (0.66–0.70)	**0.70 (0.69–0.72)**	0.55 (0.53–0.57)	**0.70 (0.68-0.71)**
		Optum EHR	22,023 (4.3)	**0.75 (0.74–0.77)**	**0.75 (0.74–0.77)**	0.68 (0.66–0.70)	0.74 (0.73–0.76)	0.52 (0.50–0.54)	0.72 (0.70–0.73)
		SIDIAP	18,201 (12.3)	**0.79 (0.78–0.80)**	**0.79 (0.78–0.80)**	0.74 (0.73–0.75)	**0.79 (0.78–0.80)**	0.57 (0.56–0.58)	0.78 (0.77–0.79)
	All	HIRA-COVID	6445 (2.1)	**0.90 (0.87–0.93)**	0.87 (0.83–0.91)	0.88 (0.85–0.91)	0.76 (0.72–0.80)	0.59 (0.54–0.64)	0.82 (0.78–0.86)
		Optum Claims	5663 (15.6)	**0.74 (0.72–0.76)**	0.73 (0.72–0.75)	0.72 (0.71–0.74)	0.69 (0.67–0.71)	0.69 (0.67–0.70)	0.70 (0.68–0.72)
		Optum EHR	22,023 (4.3)	**0.76 (0.75–0.78)**	0.74 (0.72–0.75)	0.72 (0.70–0.74)	0.65 (0.63–0.67)	0.65 (0.63–0.67)	0.69 (0.67–0.70)
		SIDIAP	18,201 (12.3)	**0.78 (0.77–0.78)**	0.77 (0.76–0.78)	0.77 (0.76–0.78)	0.68 (0.67–0.69)	0.55 (0.54–0.57)	0.73 (0.72–0.74)

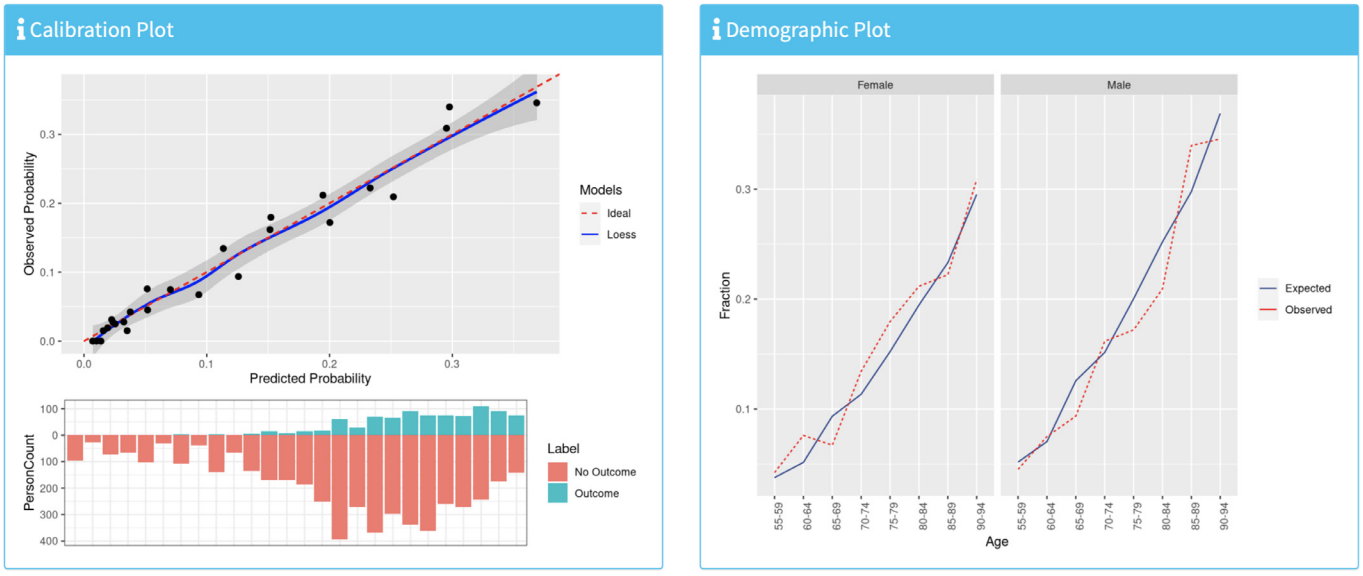

Fig. 7. Calibration performance for internal validation of the L1-regularized logistic regression model for predicting 30-day death outcome in patients hospitalized with COVID-19 on Optum Claims data, overall (left panels) and by age and gender (right panels).

CovidDeathPrediction/, which shows the *model summary* (including model coefficients or variable importance for non-generalized linear models), *model performance* (including discrimination (AUC, F1 Score, Precision (also known as Positive predictive value), Recall (also known as Sensitivity), and more) and calibration (observed vs predicted risk, overall and by age and gender)), and *all model settings* (Fig. 9). For instance, the hyperparameter values for all models used in this study are available in the "Settings" tab and Fig. 10 shows the intercept term and coefficients for the final age and gender L1-regularized logistic regression model in the "Model" tab. The complete model can also be downloaded from this tab.

4. Discussion

As an open science initiative, researchers from anywhere in the world can join the OHDSI collaborative, and any data custodian can become a data partner by mapping their data to the OMOP CDM. The fast-growing OHDSI distributed data network enables performance assessment at a scale that can be highly valuable to develop prediction models that may impact patient care and outcome. The proposed pipeline is an expansion of existing machine learning software. It includes software tools and methods for extracting suitable data for a given prediction problem from big observational healthcare data and provides an easy process for sharing prediction models. The pipeline is transparent, and the software tools are open source. Due to the constant progress in the machine learning community, the machine learning part of the pipeline was developed to be flexible and different machine learning software can be readily integrated. For example, the caret *R* package [38] or new state-of-the-art machine learning methods [39] could be readily integrated into the pipeline. In addition, the types of prediction problems supported by the pipeline is expanding, with current support available for both binary classification and survival analysis.

Once a prediction problem is specified, having suitable phenotypes and data is essential. Our prediction pipeline includes the important step of validating phenotypes across a network of databases prior to implementing any model development. This

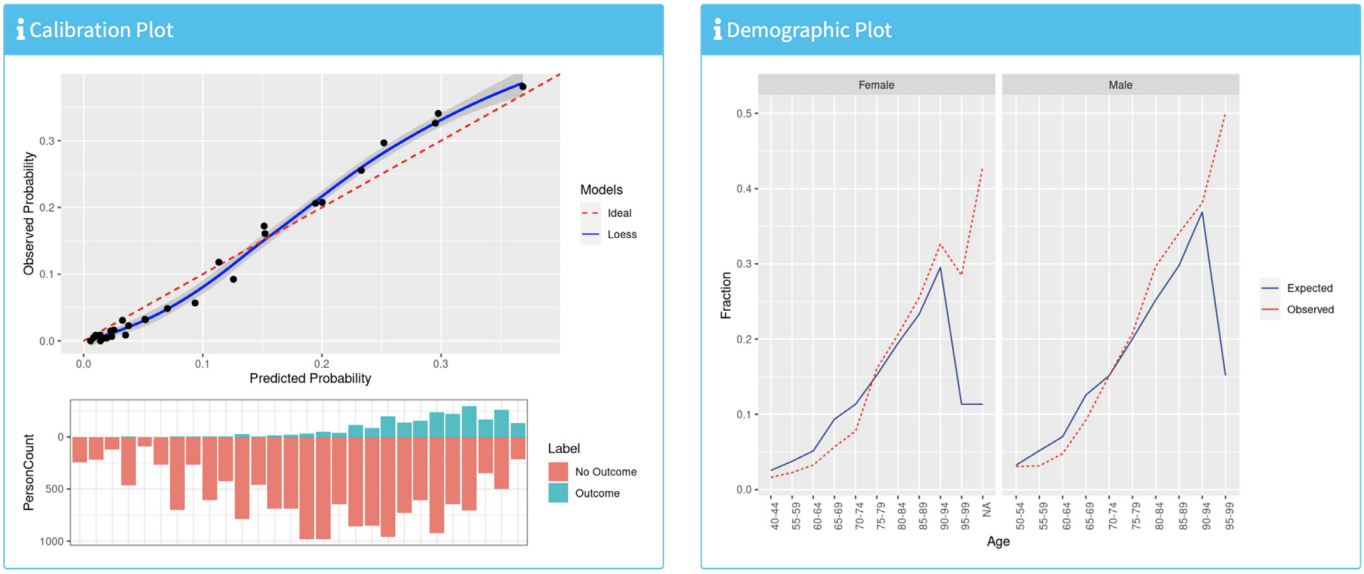

Fig. 8. Calibration performance for external validation of the L1-regularized logistic regression model for predicting 30-day death outcome in patients hospitalized with COVID-19 on SIDIAP data, overall (left panels) and by age and gender (right panels).

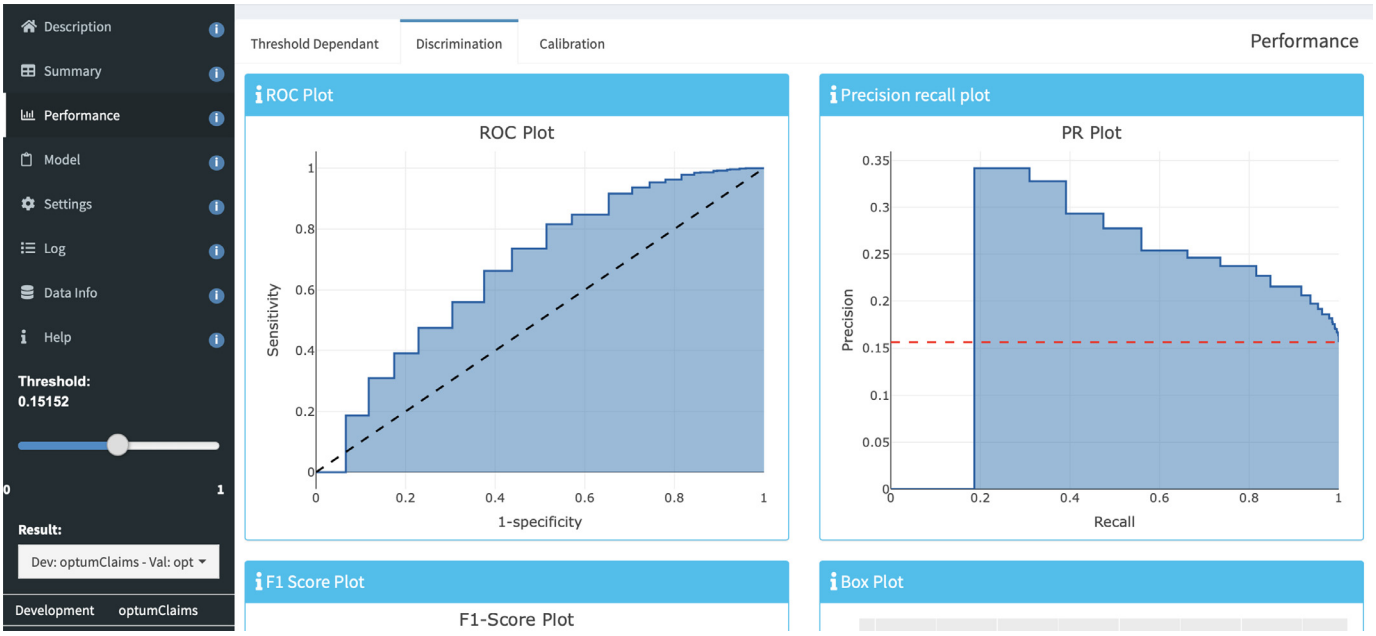

Fig. 9. A snapshot of the viewer dashboard. It contains the model summary, model performance, and all model settings.

step aims to ensure phenotypes are transportable and aims to improve the reliability of the model. External validation across diverse datasets is made possible due to the OHDSI standardizations and collaborative network [37]. This is a key strength of our prediction pipeline and in this paper, we demonstrated how it was possible to perform external validation of prediction models across multiple countries. The majority of published COVID-19 prediction models were unable to provide such an extensive set of external validation results. Finally, our prediction pipeline enforces best practices for transparent reporting of prediction models as described in the TRIPOD statement [2].

Age and gender were found to be the main predictors of death within 30-days of hospitalization with COVID-19, which suggests our findings are consistent with the literature [5]. Adding more variables improved the model performance in Optum Claims and Optum EHR, with the best performing model being L1-regularized logistic regression. Interestingly, the L1-regularized logistic regression, AdaBoost, and gradient boosting machine models that only used age and gender predictors performed the best in the SIDIAP data.

In HIRA-COVID, we also found that the L1-regularized logistic regression models that only used age and gender predictors had the highest AUC of all models. This suggests that this model may be more transportable across countries and healthcare settings. However, although the Korean COVID-19 patient population itself is young, almost all deaths were in elderly patients over 65 years of age [8], and age being a dominant predictor is a possible reason for the better performance of the models using only age and gender. Further, a single measure cannot fully evaluate the model's performance and other measures may provide a different interpre-

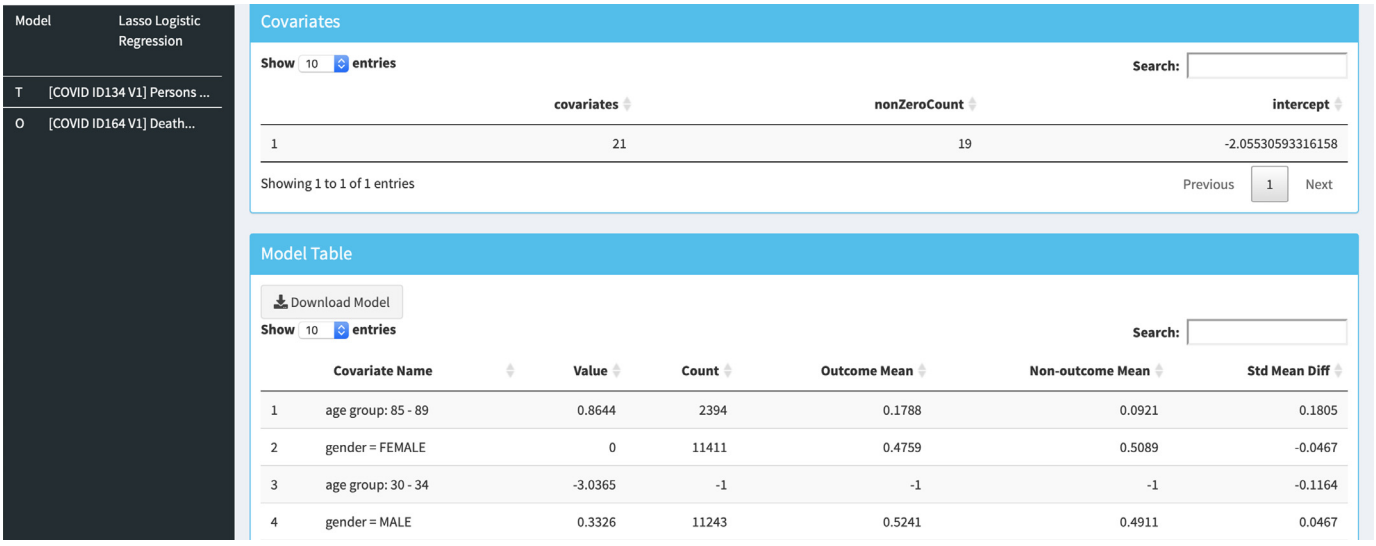

Fig. 10. A snapshot of a Model Table in the Viewer Dashboard. It contains the complete model specification including intercept term and coefficient values for each covariate included in the final model.

tation. For instance, the AUPRC scores (in the Viewer Dashboard) for HIRA are lower than for the other databases, possibly due to a relatively low death rate in South Korea.

As with studies based on distributed data networks, a limitation of the approach presented in this paper is that it relies on data partners to map their data to the OMOP CDM. This initial mapping can be time-consuming. However, once done, a database can rapidly be integrated into a network study. Despite including data from three different continents, there are many regions of the world that are not represented in this paper. As more databases actively join the OHDSI data network (including from South Asia and Latin America), we can rapidly extend the external validation to them in the near future.

Using four databases from across the world, we developed and externally validated prediction models for 30-day risk of death in patients hospitalized with COVID-19. This study demonstrating the proposed pipeline, focusing on COVID-19 mortality, was initiated on November 1, 2020, and the results were publicly shared in an *R* Shiny app on December 15, 2020, which means it took only weeks to complete the study. The speed of the study did not compromise the quality of the study due to using the proposed reliable pipeline. We demonstrated the quality of the developed models *via* extensive external validation of phenotypes and prediction models. The model performances were generally consistent across diverse datasets, with AUCs ranging from 0.75 to 0.93, suggesting there was minimal bias in model development. The complete analytical source code used for the study is publicly shared for transparency and reproducibility. We hence demonstrated how the OHDSI analytics pipeline for patient-level prediction modeling offers a standardized approach for rapid yet reliable development and validation of prediction models, one that allows researchers to address various sources of bias. This work is a step towards obtaining prediction models that can provide reliable evidence-based guidance for use in clinical practice.

Funding

This work was supported by the European Health Data & Evidence Network (EHDEN). EHDEN has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under Grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This work was also supported by the Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol). The IDIAPJGol received funding from the Health Department from the Generalitat de Catalunya with a grant for research projects on SARS-CoV-2 and COVID-19 disease organized by the Direcció General de Recerca i Innovació en Salut. This work was also supported by the Bio Industrial Strategic Technology Development Program (20003883) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [Grant number: HI16C0992]. The investigators acknowledge the philanthropic support of the donors to the University of Oxford's COVID-19 Research Response Fund. Study sponsors had no involvement in the study design, in the collection, analysis, and interpretation of data, in the writing of the manuscript, nor in the decision to submit the manuscript for publication.

Declaration of Competing Interest

CB, MJS, AGS, JMR are employees of Janssen Research & Development and shareholders of Johnson & Johnson.

Acknowledgments

Statements of ethical approval

All databases obtained IRB approval or used deidentified data that was considered exempt from IRB approval. Informed consent was not necessary at any site.

References

[1] World Health Organization, COVID-19 weekly epidemiological update, edition 45, 22 June 2021., World Health Organization [Online], 2021 [Online]. Available: https://apps.who.int/iris/handle/10665/342009.

[2] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, Circulation 131 (2) (2015) 211–219, doi:10.1186/s12916-014-0241-z.

[3] H. Al-Najjar, N. Al-Rousan, A classifier prediction model to predict the status of coronavirus COVID-19 patients in South Korea, Eur. Rev. Med. Pharmacol. Sci. 24 (6) (2020) 3400–3403, doi:10.26355/eurrev_202003_20709.

[4] Y. Shi, X. Yu, H. Zhao, H. Wang, R. Zhao, J. Sheng, Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan, Crit. Care 24 (1) (2020) 108 Mar 18, doi:10.1186/s13054-020-2833-7.

[5] L. Wynants, et al., Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal, BMJ 369 (2020) m1328 Apr., doi:10.1136/bmj.m1328.
[6] M. Yuan, W. Yin, Z. Tao, W. Tan, Y. Hu, Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China, PLoS One 15 (3) (2020) e0230548, doi:10.1371/journal.pone.0230548.
[7] Observational Health Data Sciences and Informatics, The Book of OHDSI, Observational Health Data Sciences and Informatics [Online], 2020 Available: https://ohdsi.github.io/TheBookOfOhdsi/ (Accessed date: 7 December 2020).
[8] E. Burn, et al., Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study, Nat. Commun. 11 (1) (2020) 5009 Oct., doi:10.1038/s41467-020-18849-z.
[9] T. Duarte-Salles, D. Vizcaya, A. Pistillo, 30-day outcomes of children and adolescents with COVID-19: an international experience, Pediatrics (2021).
[10] COVID-19: Reminder of Risk of Serious Side Effects with Chloroquine and Hydroxychloroquine, European Medicines Agency, Amsterdam, The Netherlands, 2020 Apr. 23, [Online]. Available: https://www.ema.europa.eu/en/news/covid-19-reminder-risk-serious-side-effects-chloroquine-hydroxychloroquine .
[11] A. Golozar, et al., Baseline phenotype and 30-day outcomes of people tested for COVID-19: an international network cohort including >3.32 million people tested with real-time PCR and >219,000 tested positive for SARS-CoV-2 in South Korea, Spain and the United States, medRxiv [Preprint] (2020), doi:10.1101/2020.10.25.20218875.
[12] L.Y.H. Lai, et al., Clinical characteristics, symptoms, management and health outcomes in 8598 pregnant women diagnosed with COVID-19 compared to 27,510 with seasonal influenza in France, Spain and the US: a network cohort analysis, medRxiv [Preprint] (2020), doi:10.1101/2020.10.13.20211821.
[13] J.C.E. Lane et al., Risk of depression, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multinational network cohort study. Rheumatology. 60 (7) (2021) 3222-3234, doi:10.1093/rheumatology/keaa771.
[14] J.C.E. Lane, et al., Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study, Lancet Rheumatol. 2 (11) (2020) e698–e711 Nov., doi:10.1016/S2665-9913(20)30276-9.
[15] X. Li, et al., Characterising the background incidence rates of adverse events of special interest for COVID-19 vaccines in eight countries: multinational network cohort study, BMJ 373 (2021) n1435, doi:10.1136/bmj.n1435.
[16] D.R. Morales, et al., Renin-angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis, Lancet Digit. Health (2020) Dec., doi:10.1016/S2589-7500(20)30289-2.
[17] A. Prats-Uribe, et al., Use of repurposed and adjuvant drugs in hospital patients with COVID-19: multinational network cohort study, BMJ 373 (2021) n1038, doi:10.1136/bmj.n1038.
[18] M. Recalde, et al., Characteristics and outcomes of 627 044 COVID-19 patients living with and without obesity in the United States, Spain, and the United Kingdom, Int. J. Obes. (2021) 1–11, doi:10.1038/s41366-021-00893-4.
[19] J.M. Reps, et al., Implementation of the COVID-19 vulnerability index across an international network of health care data sets: collaborative external validation study, JMIR Med. Inform. 9 (4) (2021) e21547 Apr 5, doi:10.2196/21547.
[20] A. Shoaibi, S.P. Fortin, R. Weinstein, J.A. Berlin, P. Ryan, Comparative effectiveness of famotidine in hospitalized COVID-19 patients, Off. J. Am. Coll. Gastroenterol.| ACG 116 (4) (2021) 692–699.
[21] E.H. Tan, et al., COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries, Rheumatology (2021), doi:10.1093/rheumatology/keab250.
[22] R.D. Williams, et al., Seek COVER: development and validation of a personalized risk calculator for COVID-19 outcomes in an international network, medRxiv [Preprint] (2020), doi:10.1101/2020.05.26.20112649.
[23] R.W. Platt, R. Platt, J.S. Brown, D.A. Henry, O.H. Klungel, S. Suissa, How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias, Pharmacoepidemiol. Drug Saf. (2019) Jan., doi:10.1002/pds.4722.
[24] E. Burn, et al., Opioid use, postoperative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study, Lancet Rheumatol. 1 (4) (2019) e229–e236 Dec., doi:10.1016/S2665-9913(19)30075-X.
[25] E. Burn, et al., Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study., Nature communications 11 (1) (2020) 1–11, doi:10.1038/s41467-020-18849-z.
[26] Observational Health Data Sciences and Informatics, OMOP Common Data Model, GitHub repository [Online] (2020) http://ohdsi.github.io/CommonDataModel/. (Accessed 7 December 2020).
[27] Observational Health Data Sciences and Informatics, WhiteRabbit, GitHub repository [Online] (2020) https://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html. (Accessed 7 December 2020).
[28] Observational Health Data Sciences and Informatics, Rabbit in a Hat, GitHub repository [Online] (2020) https://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html. (Accessed 7 December 2020).
[29] Observational Health Data Sciences and Informatics, Usagi, GitHub repository [Online] (2020). (Accessed 7 December 2020). https://github.com/ohdsi/usagi.
[30] Observational Health Data Sciences and Informatics, DataQualityDashboard, GitHub repository [Online] (2020) https://ohdsi.github.io/DataQualityDashboard/. (Accessed 7 December 2020).
[31] V. Huser, et al., Multisite evaluation of a data quality tool for patient-level clinical data sets, EGEMS 4 (1) (2016) 1239 (Wash DC)Nov., doi:10.13063/2327-9214.1239.
[32] J.M. Reps, M.J. Schuemie, M.A. Suchard, P.B. Ryan, P.R. Rijnbeek, Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data, J. Am. Med. Inform. Assoc. 25 (8) (2018) 969–975, doi:10.1093/jamia/ocy032.
[33] Observational Health Data Sciences and Informatics, CohortDiagnostics, GitHub repository [Online] (2020) https://ohdsi.github.io/CohortDiagnostics/. (Accessed 7 December 2020).
[34] L.H. John, J.A. Kors, J.M. Reps, P.B. Ryan, and P.R. Rijnbeek, How little data do we need for patient-level prediction?, arXiv [Preprint] (2020), doi: arXiv:2008.07361.
[35] G.S. Collins, E.O. Ogundimu, D.G. Altman, Sample size considerations for the external validation of a multivariable prognostic model: a resampling study, Stat. Med. 35 (2) (2016) 214–226 Jan., doi:10.1002/sim.6787.
[36] M.A. Suchard, S.E. Simpson, I. Zorych, P. Ryan, D. Madigan, Massive parallelization of serial inference algorithms for a complex generalized linear model, ACM Trans. Model. Comput. Simul. 23 (1) (2013), doi:10.1145/2414416.2414791.
[37] J.M. Reps, et al., Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation, BMC Med. Res. Methodol. 20 (1) (2020) 102 May, doi:10.1186/s12874-020-00991-3.
[38] M. Kuhn, et al., Building predictive models in R using the caret package, J. Stat. Softw. 28 (1) (2008) 1–26.
[39] D. Patel, N. Zhou, S. Shrivastava, J. Kalagnanam, Doctor for machines: a failure pattern analysis solution for industry 4.0, in: Proceedings of the IEEE International Conference on Big Data, IEEE, 2020, pp. 1614–1623. (Big Data).