Background.
Observational Health Data Sciences and Informatics (OHDSI) is an open science collaborative that aims to improve health by empowering community collaborations that generate reliable evidence from observational data to promote better health decisions and better care. OHDSI is a multi-stakeholder and multi-disciplinary community of researchers and has established an international data network to provide opportunities for large-scale observational research. OHDSI has four strategic pillars: data standardization, to harmonize the structure and content of patient-level healthcare data for the purposes of observational analysis; methodological research, to establish and evaluate scientific best practice for generating reliable evidence from the patient-level data; open-source analytics development, to codify best practices into tools that can be reproducibly implemented across the community; and clinical use cases, where the scientific best practices can be consistently applied through the community’s open-source tools to specific clinical problems raised by the community. OHDSI focuses on three types of clinical use cases: clinical characterization – descriptive statistics to summarize real-world practice, such as disease natural history and treatment utilization; population-level effect estimation – causal inference for safety surveillance and comparative effectiveness; and patient-level prediction – inference for precision medicine and disease interception.

Methods.
The OHDSI Patient-level Prediction workgroup has established a framework for designing prediction studies, and has implemented a set of open-source tools to support application of this framework to different clinical questions. The framework involves specification of two cohorts, a time-at-risk, and a model specification. A cohort is defined as a set of patients satisfying one or more inclusion criteria for a duration of time. The target cohort T represents the set of patients for which we are interested and who are at-risk for the given outcome during the defined time-at-risk, and the outcome cohort O represents the set of patients who experienced the outcome of interest. We seek to estimate the probability of a person in T also belongs to O within the defined time-at-risk. We standardize the feature extraction process for constructing baseline covariates using observations from the longitudinal medical history prior to T cohort entry, including demographics, conditions, drug exposures, procedures, and measurements. We apply multiple machine learning algorithms, including regularized regression, gradient boosting, random forest, multilayer perception, k-nearest neighbors, ada boost, convolutional neural networks, recurrent neural networks and deep neural networks, to learn models and identify the approaches that yield the highest discrimination. We perform internal validation to evaluate discrimination and calibration within a given database, but also utilize the data network to do external validation by allowing models to be transported to other patient-level databases who similarly adopt the OHDSI data standards and tool stack.

Under this framework, OHDSI has addressed various types of clinical questions, including: disease onset and progression: amongst patients who are newly diagnosed with <a disease>, which patients will go on to have <another disease or related complication> within <time horizon from diagnosis>?: treatment choice: amongst patients with <indicated disease> who are treated with either <treatment 1> or <treatment 2>, which patients were treated with <treatment 1>?: treatment response: amongst patients who are new users of <a drug>, which patients will <have a desired effect> in <time window>?: treatment safety - amongst patients who are new users of <a drug>, which patients will experience <an adverse event> within <time horizon following exposure start>?: and, treatment adherence - amongst patients who are new users of <a drug>, which patients will achieve <adherence metric threshold> at <time horizon>?

Conclusion.
The combination of the front-end ATLAS web application (http://www.ohdsi.org/web/atlas) and the back-end PatientLevelPrediction R package (https://github.com/OHDSI/patientlevelprediction) provides clinical researchers with the ability to design a prediction study, including specification of target and outcome cohorts, selection of relevant features, and evaluation of multiple machine learning algorithms, generate a R script to reproducibly generate analysis results, and a R Shiny application to interactively explore the results. The OHDSI stack enables clinical researchers to go end-to-end from the point of having patient-level data through the entire workflow of study design, model development and evaluation, and application of model to a target population. In this presentation, we will demonstrate the OHDSI toolkit for patient-level prediction using the clinical use case of predicting suicidal ideation or behavior amongst patients with major depressive disorders initiating antidepressant therapy.