Health Data Science Group

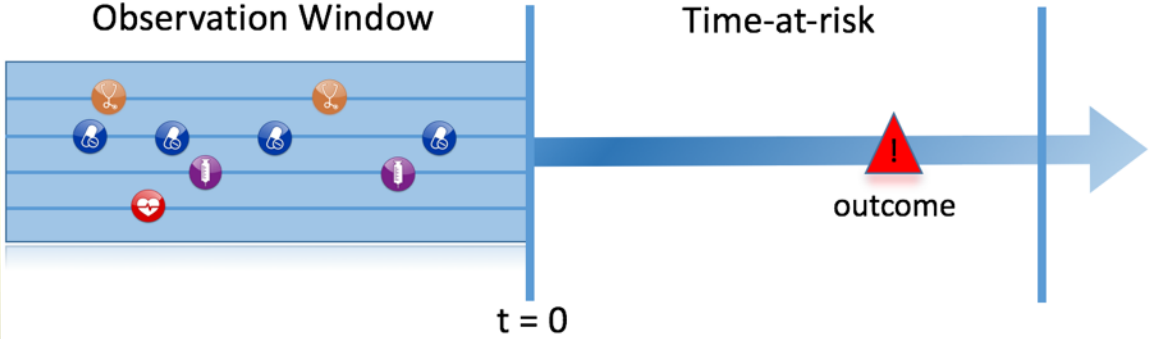
PUTTING MACHINE LEARNING INTO REAL-WORLD PRACTICE:

PATIENT-LEVEL PREDICTION DEVELOPMENT AND VALIDATION IN OBSERVATIONAL DATA

Peter R. Rijnbeek
Associate Professor Health Data Science
Department of Medical Informatics
www.healthdatascience.nl

Erasmus MC
University Medical Center Rotterdam

Problem definition



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

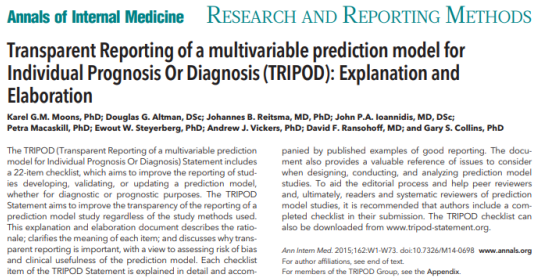
Classified as internal/staff & contractors by the European Medicines Agency

Model Development Pipeline

Problem Definition → Data Extraction → Training → Internal Validation → External Validation → **Dissemination**

Dissemination *of study results should follow the minimum requirements as stated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [1].*

- *Sharing of model development details*
- *Discrimination and Calibration*
- *Internal and external validation*
- *Etc.*



Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

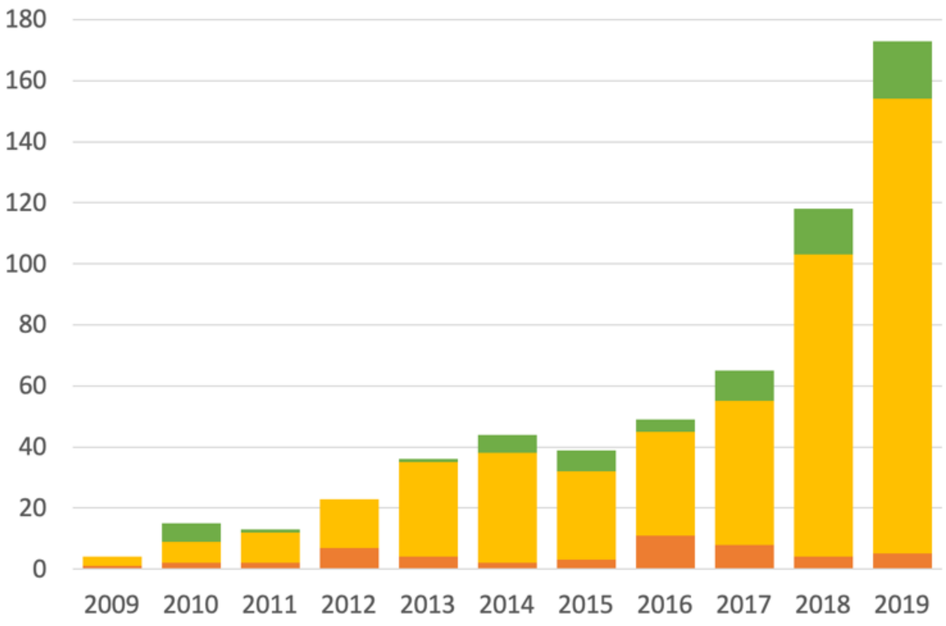
Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 **www.annals.org**

For author affiliations, see end of text.
For members of the TRIPOD Group, see the **Appendix**.

[1] *Moons, KG et al. Ann Intern Med. 2015;162(1):W1-73*

Erasmus MC

Current status of predictive modelling:



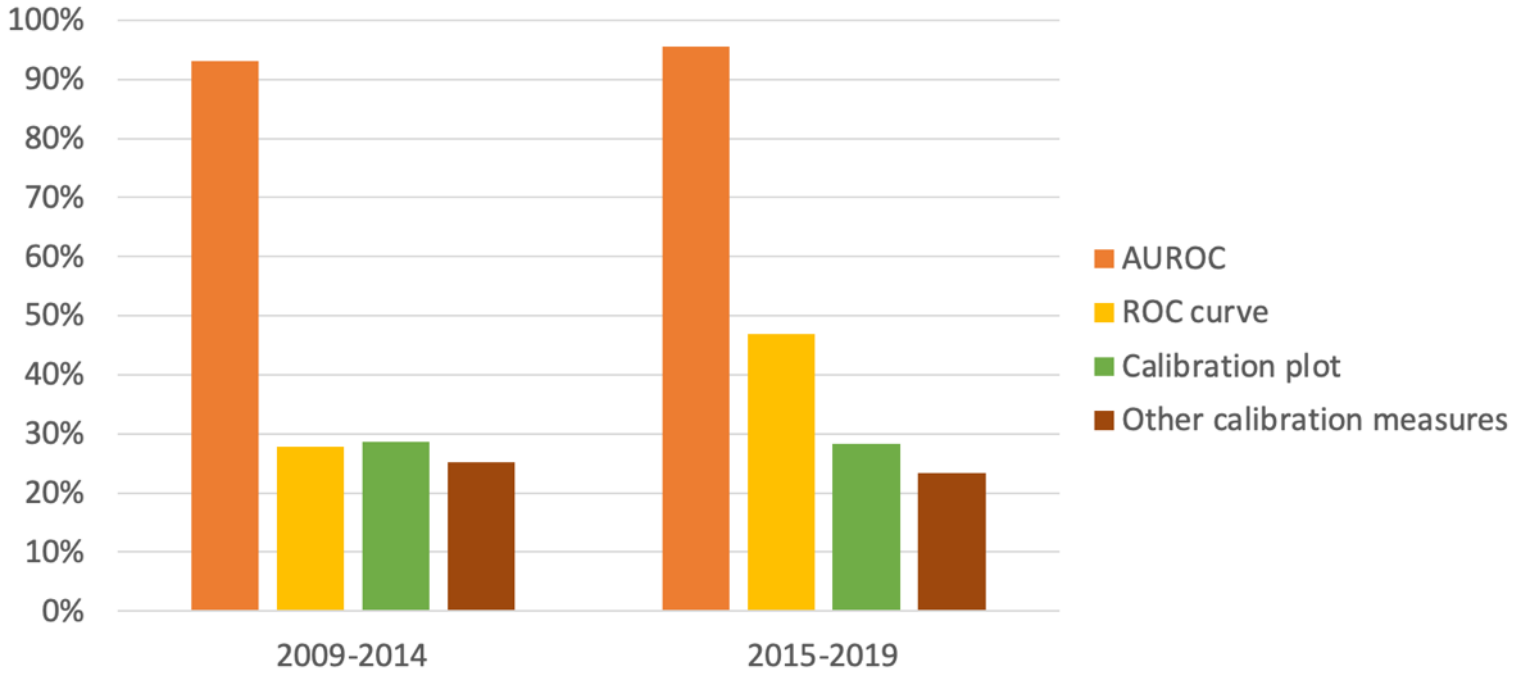
Review of 422 papers with 579 models

- **External validation**
- **Internal validation**
- **No validation**

Preliminary results of review led by Cynthia Yang of Erasmus MC

Classified as internal/staff & contractors by the European Medicines Agency

Current status of predictive modelling:



Preliminary results of review led by Cynthia Yang of Erasmus MC

Classified as Utrecht University Use Only - Erasmus Medical Center

What is needed?

Full transparency and reproducibility

1. Standardised Health Data with respect to structure and terminology
2. Standardised Analytical Pipelines that enforce best modelling practices
3. Share models and allow extensive external validation across many databases
4. Disseminate **all** performance results

Classified as internal/staff & contractors by the European Medicines Agency

Our mission for Patient-Level Prediction

The Observational Health Data Sciences and Informatics (OHDSI) and the European Health Data and Evidence Network (EHDEN) developed a systematic process to learn and evaluate large-scale patient-level prediction models using observational health data in a large data network

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
www.ohdsi.org

EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK
www.ehden.eu

Erasmus MC

Classified as internal/staff & contractors by the European Medicines Agency

Patient-Level Prediction Framework

Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, Peter R Rijnbeek

Journal of the American Medical Informatics Association, Volume 25, Issue 8, August 2018, Pages 969–975, https://doi.org/10.1093/jamia/ocy032
Published: 27 April 2018 **Article history** ▾

PDF | Split View | Cite | Permissions | Share ▾

Abstract

Objective

To develop a conceptual prediction model framework containing standardized steps and describe the corresponding open-source software developed to consistently implement the framework across computational environments and observational healthcare databases to enable model sharing and reproducibility.

R-package

www.github.com/OHDSI/PatientLevelPrediction

Book-of-OHDSI
https://book.ohdsi.org/

OHDSI
技术指南
实用真实世界临床数据分析工具

EHDEN Academy *https://academy.ehden.eu*

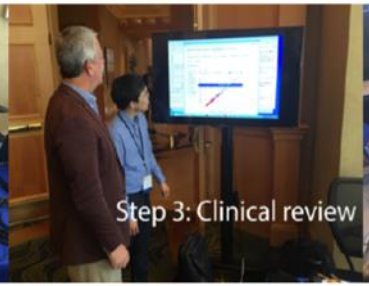
Study Results
https://data.ohdsi.org

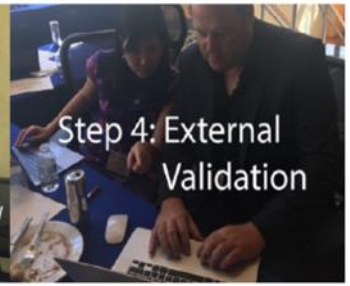
Prediction team in action

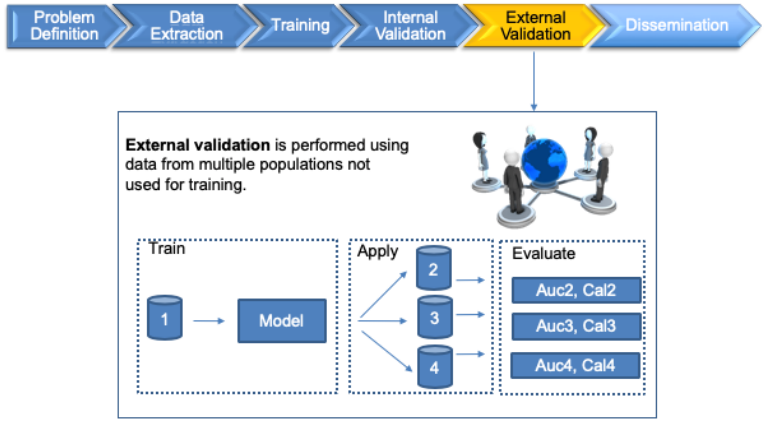
OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Patient-Level Prediction Team Work







Erasmus MC

Classified as internal/staff & contractors by the European Medicines Agency

Model Specification

ATLAS

- Home
- Data Sources
- Search
- Concept Sets
- Cohort Definitions
- Characterizations
- Cohort Pathways
- Incidence Rates
- Profiles
- Estimation
- Prediction
- Jobs
- Configuration
- Feedback

Apache 2.0 open source software

provided by OHDSI

VIEW: All | Prediction Problem Settings | Analysis Settings | Execution Settings | Training Settings

Prediction Problem Settings

Target Cohorts + Add Target Cohort

Show 10 entries

Remove	Name
✖	[EHDEN RA] Female new users of methoxtrexate monotherapy used for PLP
✖	[EHDEN RA] New users of methoxtrexate monotherapy used for PLP

Showing 1 to 2 of 2 entries

Outcome Cohorts

Show 10 entries

Remove	Name
✖	[EHDEN RA] Leukopenia events using diagnoses and measurements
✖	[EHDEN RA] Pancytopenia events using diagnoses and measurements
✖	[EHDEN RA] Pancytopenia or leukopenia events using diagnoses and measurements
✖	[EHDEN RA] Stroke (ischemic or hemorrhagic) events (any visit) (1)
✖	[EHDEN RA] Opportunistic Infections (2)
✖	[EHDEN RA] Serious Infection events (2)
✖	[EHDEN RA] Serious Infection, opportunistic infections and other infections of interes

Model Settings + Add Model Set

Show 10 entries Filter:

Remove	Model	Options
✖	LassoLogisticRegressionSettings	{"variance":0.01,"seed":null}

Showing 1 to 1 of 1 entries Previous

Covariate Settings + Add Covariate S

Column visibility | Copy | CSV Show 10 entries Filter:

Remove	Options
✖	DemographicsGender, DemographicsAge, DemographicsAgeGroup, DemographicsIndexMonth, ConditionOccurrenceAnyTimePrior, DrugGroupEraLongTerm (+4 more covariate settings)
✖	DemographicsGender, DemographicsAgeGroup, DemographicsIndexMonth, ConditionOccurrenceAnyTimePrior, CharlsonIndex, Dcsi (+1 more covariate settings)

Showing 1 to 2 of 2 entries Previous

Population Settings + Add Population S

Column visibility | Copy | CSV Show 10 entries Filter:

Remove	Risk Window Start	Risk Window End	Washout Period	Include All Outcomes	Remove Subjects With Prior Outcome	Minimum Time At Risk
✖	1d from cohort start date	730d from cohort start date	365d	true	true	1d
✖	1d from cohort start date	90d from cohort start date	365d	true	true	1d
✖	365d from cohort start date	1826d from cohort start date	365d	true	true	1d

Previous

Classified as internal/staff & contractors by the European Medicines Agency

Generate R-Package and share with the world

Review & Download

Review Full Study Specification

Please review the full study specification below and scroll down the page to download the study package.

Full Analysis List 132 | Prediction Problem Settings 22 | Analysis Settings 6

Column visibility | Copy | CSV | Show 10 entries | Filter:

Target Cohorts
[EHDEN RA] New users of methoxtrexate monotherapy used for PLP (66)
[EHDEN RA] Female new users of

Outcome Cohorts
[EHDEN RA] Leukopenia events using diagnoses and measurements (12)
[EHDEN RA] Opportunistic

Model Name
LassoLogisticRegressionSettings (132)

Risk Window
1-90 (44)

Target Cohort Name	Outcome Cohort Name	Model Name	Model Settings	Covariate Settings	Risk Window Start	Risk Window End
[EHDEN RA] New users of methoxtrexate monotherapy used for PLP	[EHDEN RA] Serious Infection, opportunistic infections and other infections of interest event (1)	LassoLogisticRegressionSettings	{"variance":0.01,"seed":null}	"attr_class":"covariateSettin...	1	730
[EHDEN RA] New users of methoxtrexate monotherapy used for PLP	[EHDEN RA] Serious Infection, opportunistic infections and other infections of interest event (1)	LassoLogisticRegressionSettings	{"variance":0.01,"seed":null}	"attr_class":"covariateSettin...	1	90
[EHDEN RA] New users of methoxtrexate monotherapy used for PLP	[EHDEN RA] Serious Infection, opportunistic infections and other infections of interest event (1)	LassoLogisticRegressionSettings	{"variance":0.01,"seed":null}	"attr_class":"covariateSettin...	365	1826
[EHDEN RA] New users	[EHDEN RA] Serious Infection,					

Classified as internal/staff & contractors by the European Medicines Agency

Share model performance

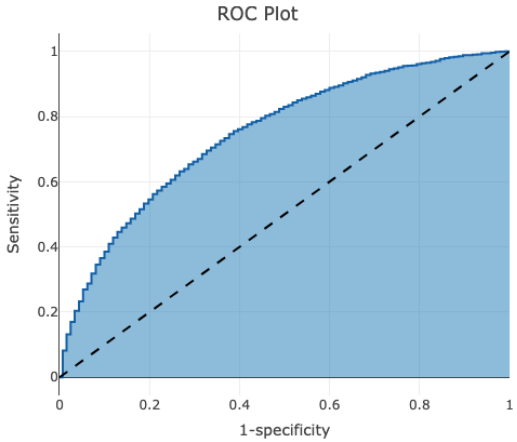
Prediction Viewer

About | Internal Validation | External Validation

Evaluation Summary | Characterization | ROC | Calibration | Demographics | Preference | Box Plot | Settings

Plot | Table

Train



Test

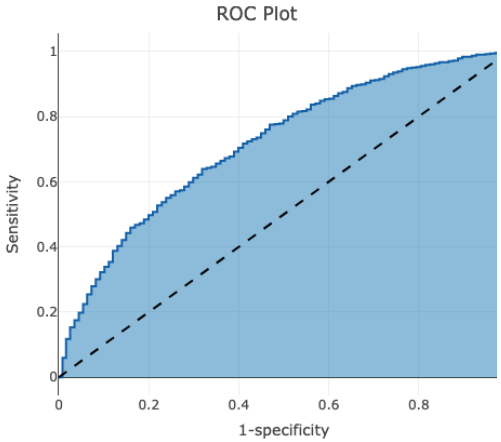


Figure 2. The Receiver Operating Characteristics (ROC) curve shows the ability of the model to discriminate between people with and without the outcome during the time at risk. It is a plot of sensitivity vs 1-specificity at every probability threshold. The higher the area under the ROC plot the higher the discriminative performance of the model. The diagonal refers to a model assigning a class at random (area under de ROC = 0.5).

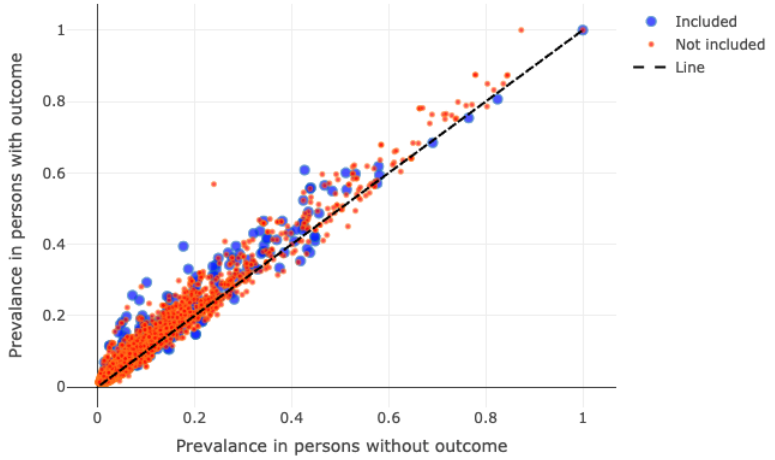


Figure 1. The variable scatter plot shows the mean covariate value for the people with the outcome against the mean covariate value for the people without the outcome. The meaning of the size and color of the dots depends on the settings on the left of the figure.

Data generated: 2018-10-01 01:11:16

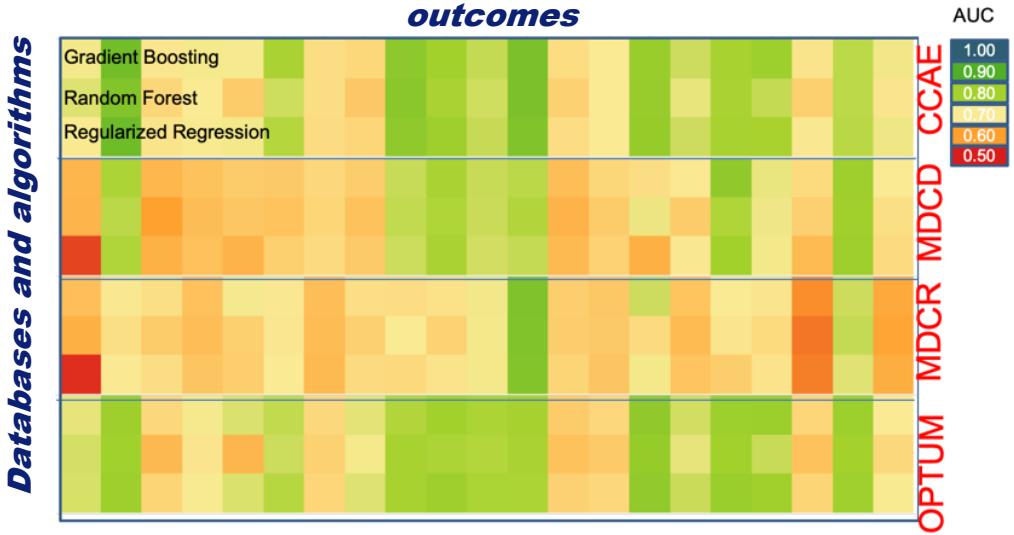
Erasmus MC

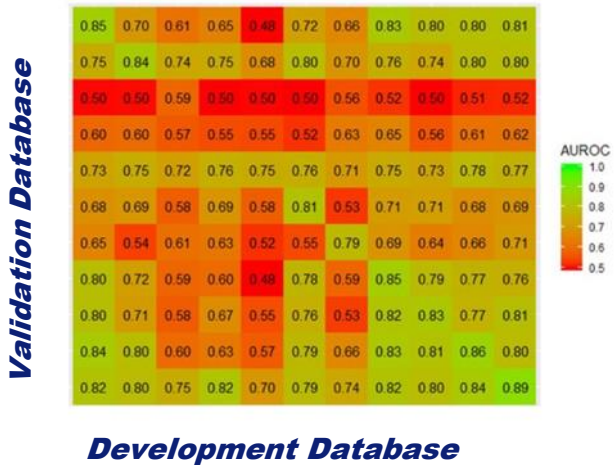
Classified as internal/staff & contractors by the European Medicines Agency

Large scale validation and dissemination

The tool auto generates a word document containing all the model specifications, internal and external validation results, model details etc. etc. which serves as a kickstart for result dissemination.

Multiple interesting visualisation can be created:

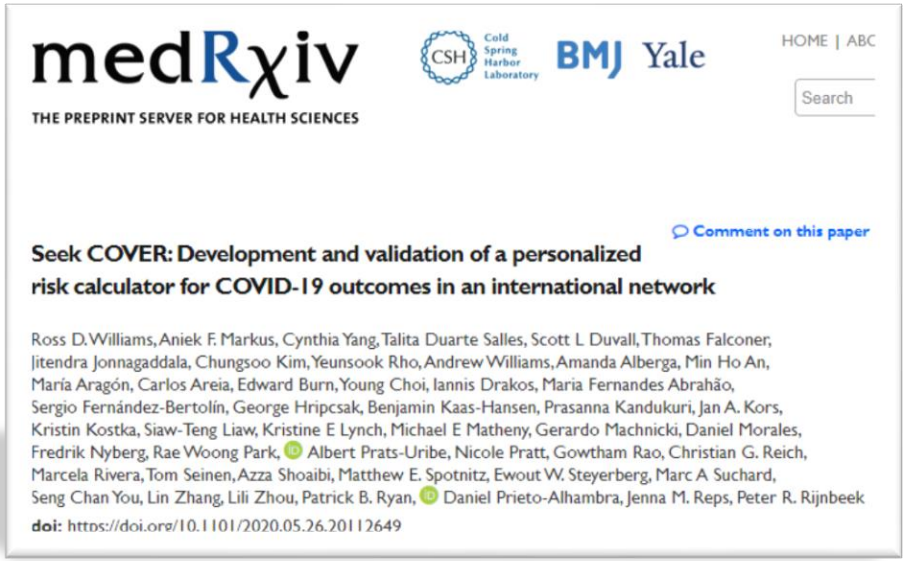




Classified as internal/staff & contractors by the European Medicines Agency

Seek COVER: COVID risk prediction

Objective: develop and externally validate **COV**ID-19 **E**stimated **R**isk scores that quantify a patient's risk of hospital admission, hospitalization requiring intensive services or fatality.



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

HOME | ABC

Search

Comment on this paper

Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network

Ross D. Williams, Aniek F. Markus, Cynthia Yang, Talita Duarte Salles, Scott L Duvall, Thomas Falconer, Jitendra Jonnagaddala, Chungsoo Kim, Yeunsook Rho, Andrew Williams, Amanda Alberga, Min Ho An, María Aragón, Carlos Areia, Edward Burn, Young Choi, Iannis Drakos, Maria Fernandes Abrahão, Sergio Fernández-Bertolín, George Hripcsak, Benjamin Kaas-Hansen, Prasanna Kandukuri, Jan A. Kors, Kristin Kostka, Siaw-Teng Liaw, Kristine E Lynch, Michael E Matheny, Gerardo Machnicki, Daniel Morales, Fredrik Nyberg, Rae Woong Park, Albert Prats-Uribe, Nicole Pratt, Gowtham Rao, Christian G. Reich, Marcela Rivera, Tom Seinen, Azza Shoaibi, Matthew E. Spotnitz, Ewout W. Steyerberg, Marc A Suchard, Seng Chan You, Lin Zhang, Lili Zhou, Patrick B. Ryan, Daniel Prieto-Alhambra, Jenna M. Reps, Peter R. Rijnbeek

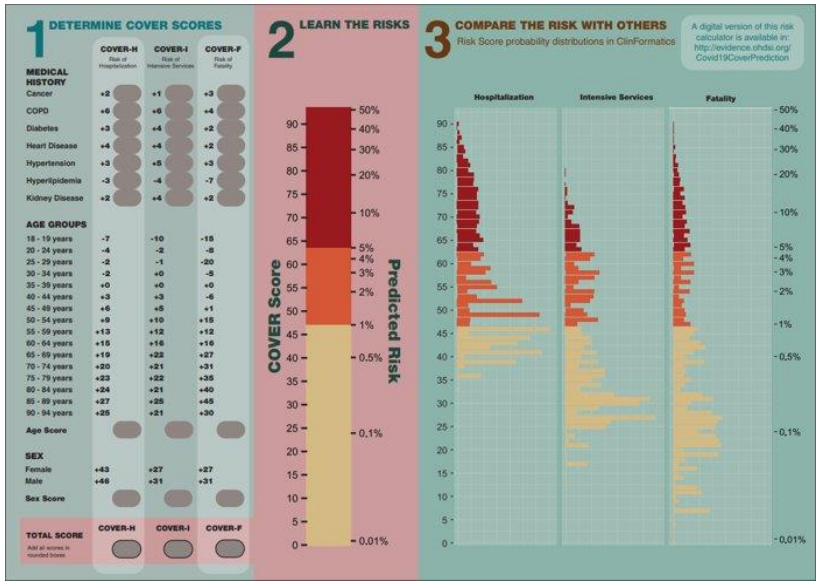
doi: https://doi.org/10.1101/2020.05.26.20112649

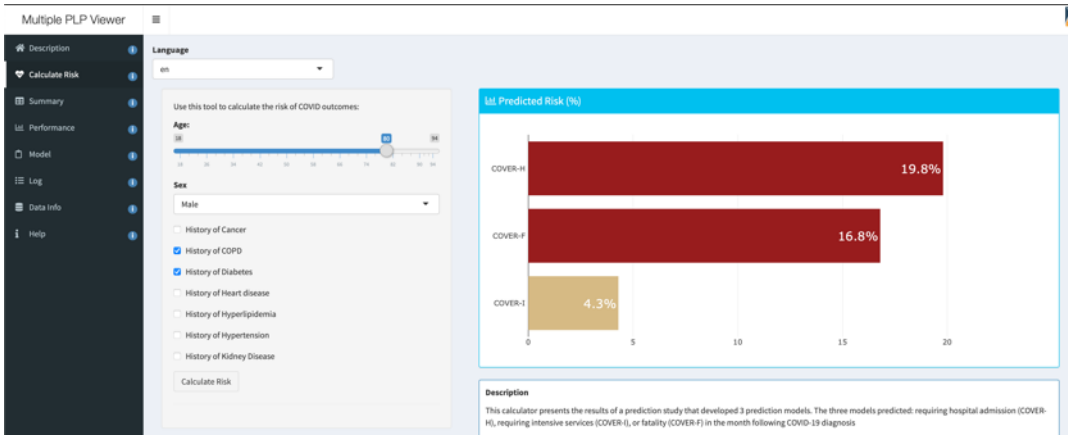
- ***14 data sources from 6 countries***
- ***Externally validated in 44,507 COVID cases from 5 data sources in South Korea, Spain, USA***

Under review

Classified as internal/staff & contractors by the European Medicines Agency

Results dissemination

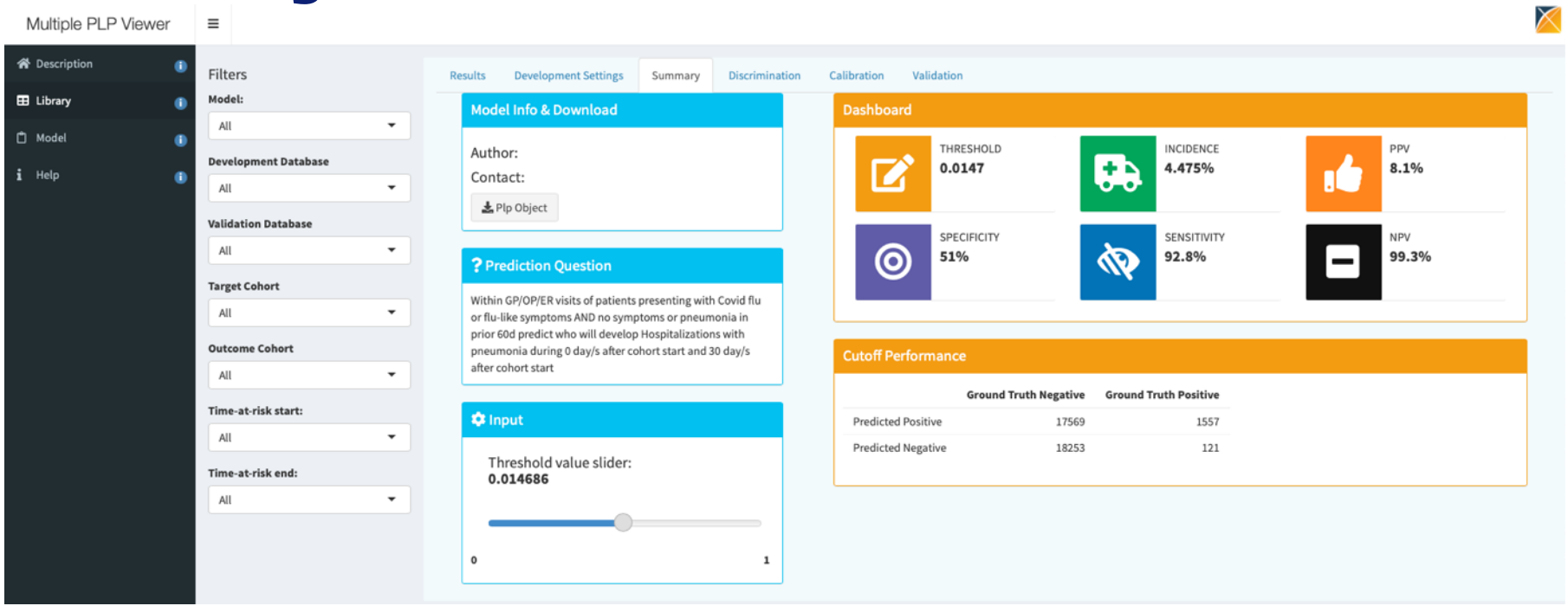




https://data.ohdsi.org/Covid19CoverPrediction/

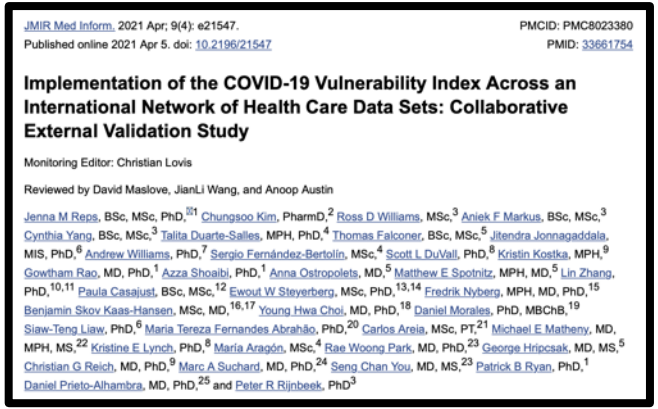
Classified as internal/staff & contractors by the European Medicines Agency

What is next: Prediction Model Library



Classified as internal/staff & contractors by the European Medicines Agency

We need external validation at scale!



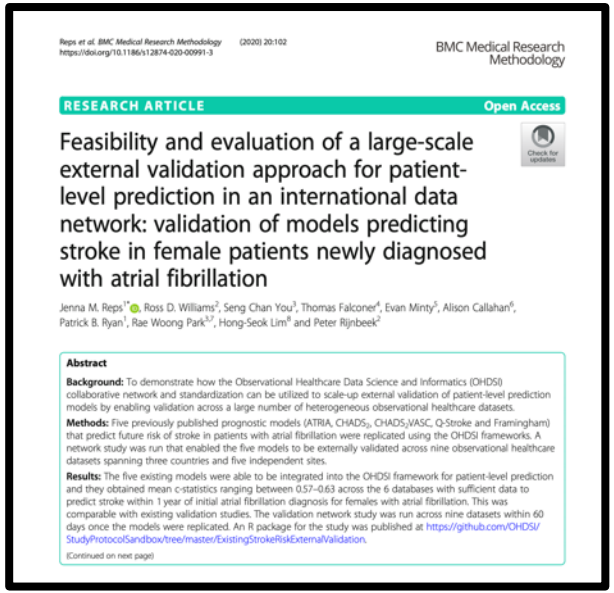
JMIR Med Inform. 2021 Apr; 9(4): e21547. PMCID: PMC8023380
Published online 2021 Apr 5. doi: 10.2196/21547 PMID: 33661754

Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study

Monitoring Editor: Christian Lovis

Reviewed by David Maslove, JianLi Wang, and Anoop Austin

Jenna M Reps, BSc, MSc, PhD,[1][✉1] Chungsoo Kim, PharmD,[2] Ross D Williams, MSc,[3] Aniek F Markus, BSc, MSc,[3] Cynthia Yang, BSc, MSc,[3] Talita Duarte-Salles, MPH, PhD,[4] Thomas Falconer, BSc, MSc,[5] Jitendra Jonnagaddala, MIS, PhD,[6] Andrew Williams, PhD,[7] Sergio Fernández-Bertolín, MSc,[4] Scott L DuVall, PhD,[8] Kristin Kostka, MPH,[9] Gowtham Rao, MD, PhD,[1] Azza Shoaibi, PhD,[1] Anna Ostropolets, MD,[5] Matthew E Spotnitz, MPH, MD,[5] Lin Zhang, PhD,[10,11] Paula Casajust, BSc, MSc,[12] Ewout W Steyerberg, MSc, PhD,[13,14] Fredrik Nyberg, MPH, MD, PhD,[15] Benjamin Skov Kaas-Hansen, MSc, MD,[16,17] Young Hwa Choi, MD, PhD,[18] Daniel Morales, PhD, MBChB,[19] Siaw-Teng Liaw, PhD,[6] Maria Tereza Fernandes Abrahão, PhD,[20] Carlos Areia, MSc, PT,[21] Michael E Matheny, MD, MPH, MS,[22] Kristine E Lynch, PhD,[8] María Aragón, MSc,[4] Rae Woong Park, MD, PhD,[23] George Hripcsak, MD, MS,[5] Christian G Reich, MD, PhD,[9] Marc A Suchard, MD, PhD,[24] Seng Chan You, MD, MS,[23] Patrick B Ryan, PhD,[1] Daniel Prieto-Alhambra, MD, PhD,[25] and Peter R Rijnbeek, PhD[3]



Reps et al. BMC Medical Research Methodology (2020) 20:102
https://doi.org/10.1186/s12874-020-00991-3

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

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

Jenna M. Reps[1*], Ross D. Williams[2], Seng Chan You[3], Thomas Falconer[4], Evan Minty[5], Alison Callahan[6], Patrick B. Ryan[1], Rae Woong Park[3,7], Hong-Seok Lim[8] and Peter Rijnbeek[2]

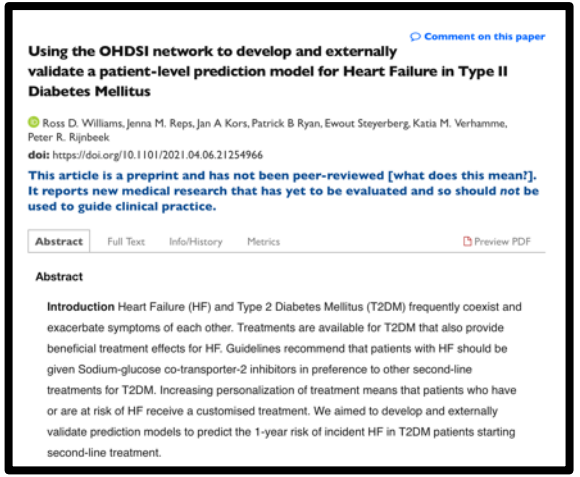
Abstract

Background: To demonstrate how the Observational Healthcare Data Science and Informatics (OHDSI) collaborative network and standardization can be utilized to scale-up external validation of patient-level prediction models by enabling validation across a large number of heterogeneous observational healthcare datasets.

Methods: Five previously published prognostic models (ATRIA, $CHADS_2$, $CHADS_2VASC$, Q-Stroke and Framingham) that predict future risk of stroke in patients with atrial fibrillation were replicated using the OHDSI frameworks. A network study was run that enabled the five models to be externally validated across nine observational healthcare datasets spanning three countries and five independent sites.

Results: The five existing models were able to be integrated into the OHDSI framework for patient-level prediction and they obtained mean c-statistics ranging between 0.57–0.63 across the 6 databases with sufficient data to predict stroke within 1 year of initial atrial fibrillation diagnosis for females with atrial fibrillation. This was comparable with existing validation studies. The validation network study was run across nine datasets within 60 days once the models were replicated. An R package for the study was published at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ExistingStrokeRiskExternalValidation.

(Continued on next page)



Comment on this paper

Using the OHDSI network to develop and externally validate a patient-level prediction model for Heart Failure in Type II Diabetes Mellitus

Ross D. Williams, Jenna M. Reps, Jan A Kors, Patrick B Ryan, Ewout Steyerberg, Katia M. Verhamme, Peter R. Rijnbeek

doi: https://doi.org/10.1101/2021.04.06.21254966

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

Introduction Heart Failure (HF) and Type 2 Diabetes Mellitus (T2DM) frequently coexist and exacerbate symptoms of each other. Treatments are available for T2DM that also provide beneficial treatment effects for HF. Guidelines recommend that patients with HF should be given Sodium-glucose co-transporter-2 inhibitors in preference to other second-line treatments for T2DM. Increasing personalization of treatment means that patients who have or are at risk of HF receive a customised treatment. We aimed to develop and externally validate prediction models to predict the 1-year risk of incident HF in T2DM patients starting second-line treatment.

Join the network!

Classified as internal/staff & contractors by the European Medicines Agency

Acknowledgement

This work would not have been possible without many contributors in the OHDSI community and the EHDEN project

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK

www.ehden.eu

@IMI_EHDEN

IMI_EHDEN

github.com/EHDEN

imi innovative medicines initiative

efpia

This project 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.

Erasmus MC

Classified as internal/staff & contractors by the European Medicines Agency