***[Note: All submissions must be in PDF format. Failure adhere to the format requirements may result in rejection of your submission without review]***

| | |
|---|---|
| Name: | Peter R. Rijnbeek, PhD |
| Affiliation: | Erasmus MC, Rotterdam, The Netherlands |
| Email: | p.rijnbeek@erasmusmc.nl |
| Presentation type (select one): | Poster |

# Best Practices for Patient-Level Prediction in OHDSI

**Peter R. Rijnbeek, PhD[1], Patrick Ryan, PhD[2], Hamed Abedtash, PharmD[3], Rich D. Boyce, PhD[4], David Dorr, MD, MS[5], George Hripcsak, MD, MS[6], Gill S. Mandev, PhD[7], Kenney Ng, PhD[8], Narges Razavian, PhD[9], David Sontag, PhD[9], James Weaver, MPH, MS[2], Andrew E. Williams, PhD[10], Johan van der Lei, PhD[1], Martijn Schuemie, PhD[2], Jenna Reps, PhD[2]**
**[1]Erasmus MC, Rotterdam, The Netherlands; [2]Janssen Research and Development, Raritan, NJ [3]Indiana University School of Informatics and Computing, Indianapolis, IN; [4]University of Pittsburgh, Pittsburgh, PA; [5]Oregon Health & Science University, Portland, OR; [6]Columbia University Medical Center, New York, NY; [7]Columbia University, New York, NY; [8]IBM Research, Cambridge MA, USA; [9]New York University, New York, NY; [10]Maine Medical Center Research Institute, Portland, Maine**

## Abstract

*Observational Health Data Sciences and Informatics (OHDSI) holds the promise of making massive-scale, patient-specific predictive modeling a reality. The OHDSI network contains longitudinal data on over 600 million patients observed for multiple years and comprising over 5 billion clinical observations. The data is stored in a common data model (CDM), enabling uniform and transparent analysis. These large standardized populations contain rich data to build highly predictive large-scale models and also provide immediate opportunity to serve large communities of patients who are in most need of improved quality of care. Effective exploitation of these massive dataset demands novel methodology and an interdisciplinary approach. The focus of the Patient-Level Prediction workgroup is to build a standardized, fully transparent workflow on top of the OMOP CDM. One of the first steps of the group was to establish best practices for patient-level predictive modelling. This work describes the consensus of the team which will provide a solid foundation for our future research and tool development in this promising field.*

## Introduction

Clinical decision making is a complicated task in which the clinician has to infer a diagnosis or treatment pathway based on the available medical history of the patient and the current clinical guidelines. Clinical prediction models have been developed to support this decision making process and are used in clinical practice in a wide spectrum of specialties. These models predict a diagnostic or prognostic outcome based on a combination of patient characteristics, e.g. demographic information, disease history, treatment history [ref recent review].

Surprisingly, most of the currently used models are estimated using small datasets and contain a limited set of patient characteristics. This low sample size, and thus low statistical power, forces the data analyst to make stronger modelling assumptions. The selection of the often limited set of patient characteristics is strongly guided by the expert knowledge at hand. This contrasts sharply with the reality of modern medicine wherein patients generate a rich digital trail, which is well beyond the power of any medical practitioner to fully assimilate. Presently, health care is generating a large amount of patient-specific information contained in the Electronic Health Record (EHR). This includes structured data in the form of diagnoses, medications, laboratory test results, and unstructured data contained in clinical narratives. Currently, it is unknown how much predictive accuracy can be gained by leveraging the large amount of data originating from the complete EHR of a patient.

The overarching goal of the patient-level prediction workgroup is to establish a standardized process for developing accurate and well-calibrated patient-centered predictive models that can be used to make predictions for multiple outcomes that are of interest to patients and can be applied to observational healthcare data from any patient subpopulation of interest. This should enable large-scale comparisons of methods and modelling techniques, for large sets of outcomes, utilizing the full EHR. A recent review showed that transparency and proper validation is often lacking[1]. Therefore, we believe that an important prerequisite for predictive modelling in OHDSI is to establish best-practices that prescribe full transparency and standardization of the modelling steps. Recently, this need has been addressed by the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis)[2], and the guidelines from the PROGRESS group initiative[3,4]. Based on the recommendations in these papers and the prior experience of the team members, an initial list of best practices for predictive modelling in OHDSI has been established.

## Results

First a number of general principles have been defined:

1. **Transparency**. Others should be able to reproduce a study in every detail using the provided information. All analysis code should be made available as open source on the OHDSI Github.
2. **Problem pre-specification.** A study protocol should unambiguously pre-specify the planned analyses.
3. **Code validation**: Unit tests, code review, or double coding steps are required to validate the developed code base. It is recommended to test the code on benchmark datasets.

Secondly, best-practices were created by consensus for each of the following modelling aspects:

1. **Data characterization and cleaning** is required before modelling, for example by investigating the covariate prevalence in the cohort. Tools are being developed in the community to facilitate this. A data cleaning step is recommended, e.g. to remove outliers in lab values.
2. **Data transformation** steps should be explained in detail to support reproducibility and scalability of the prediction model. It is recommended to document the process of recoding data values or formats to enable replication of the model development and implementation.
3. **Handling of missing values** should be declared. The workgroup believes handling of missing data in patient-level prediction is an interesting area of future research and is needed to better establish best practices.
4. **Feature construction and selection** should be completely transparent using a standardized approach to enable replication and to enable application of the model on unseen data.
5. **Inclusion and exclusion criteria** should be made explicit. It is recommended to do sensitivity analyses on the choices made. Visualization tools could help and this will be further explored.
6. **Model development** is done using a split-sample approach. The percentage used for training could depend on the number of cases, but as a rule of thumb 80/20 split is recommended. Hyper-parameter training should only be done on the training set possibly using cross-validation methodologies.
7. **Internal validation** should only be done once on the holdout set. The following minimum set of performance measures are required:
    a. Overall performance: Brier score (unscaled/scaled)
    b. Discrimination: Area under the ROC curve (AUC)
    c. Calibration: Intercept + Gradient of the line fit on the observed vs predicted probabilities

    Additionally, box plots of the predicted probabilities for the outcome vs non-outcome people, the ROC plot and a scatter plot of the observed vs predicted probabilities with a line fit and the line x=y as reference.
8. **Model Sharing** should be made possible for all studies. Several scenarios are being explored to enable central sharing of models. The goal is support replication of results, re-calibration of models, but also sharing final models with a wider community outside OHDSI.

## Conclusion

Taking into account the team's combination of backgrounds, the unfettered access to a unique data resource, and the substantial collaborative track record of OHDSI, we believe the patient-level prediction workgroup is well positioned to advance the field. The first set of best practices for patient-level predictive modelling in the OHDSI context have been established and will form the basis of the challenging but extremely interesting road ahead.

## References

1. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2016.
2. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, 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.
3. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
4. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-90.