Journey toward Patient-Level Prediction

Peter R. Rijnbeek, PhD
Department of Medical Informatics
Erasmus MC, Rotterdam, The Netherlands

Jenna Reps, PhD
Janssen Research and Development
The Journey toward Patient-Level Prediction

Peter R. Rijnbeek, PhD
Department of Medical Informatics
Erasmus MC, Rotterdam, The Netherlands
Prediction is difficult, especially about the future!
Among a population at risk (Depression), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (Stroke) during a time-at-risk (1 year). Prediction is done using only information about the patients in an observation window prior to that moment in time.
Growing interest in prediction modelling
Patient-level prediction models are already in clinical practice

CHADS2 for patients with atrial fibrillation:
+1 Congestive heart failure
+1 Hypertension
+1 Age >= 75
+1 Diabetes mellitus
+2 History of transient ischemic attack

JAMA, 2001; 285: 2864-2870
Validation of the CHADS\textsubscript{2} clinical prediction rule to predict ischaemic stroke

A systematic review and meta-analysis

Claire Keogh; Emma Wallace; Clara Dillon; Borislav D. Dimitrov; Tom Fahey
Royal College of Surgeons, Dublin, Ireland

Summary
The CHADS\textsubscript{2} predicts annual risk of ischaemic stroke in non-valvular atrial fibrillation. This systematic review and meta-analysis aims to determine the predictive value of CHADS\textsubscript{2}. The literature was systematically searched from 2001 to October 2010. Data was pooled and analysed using discrimination and calibration statistical measures, using a random effects model. Eight data sets (n=2815) were included. The diagnostic accuracy suggested a cut-point of $\geq 1$ has higher sensitivity (92\%) than specificity (12\%) and a cut-point of $\geq 4$ has higher specificity (96\%) than sensitivity (33\%). Lower summary estimates were observed for cut-points $\geq 2$ (sensitivity 79\%, specificity 42\%) and $\geq 3$ (specificity 77\%, sensitivity 50\%). There was insufficient data to analyse cut-points $\geq 5$ or $\geq 6$. Moderate pooled $c$ statistic values were identified for the classic (0.63, 95\% CI 0.52–0.75) and revised (0.60, 95\% CI 0.43–0.72) view of stratification of the CHADS\textsubscript{2}. Calibration analysis indicated no significant difference between the predicted and observed strokes across the three risk strata for the classic or revised view. All results were associated with high heterogeneity, and conclusions should be made cautiously. In conclusion, the pooled $c$ statistic and calibration analysis suggests minimal clinical utility of both the classic and revised view of the CHADS\textsubscript{2} in predicting ischaemic stroke across all risk strata. Due to high heterogeneity across studies and low event rates across all risk strata, the results should be interpreted cautiously. Further validation of CHADS\textsubscript{2} should perhaps be undertaken, given the methodological differences between many of the available validation studies and the original CHADS\textsubscript{2} derivation study.

Keywords
Atrial fibrillation, cerebral infarct, risk factors, risk prediction, CHADS\textsubscript{2}

Thromb Haemost 2011; 106: 528–538
Recommendation:

In patients with **nonvalvular atrial fibrillation**, the CHA$_2$DS$_2$-VASc score is recommended for assessment of stroke risk

<table>
<thead>
<tr>
<th>CHA$_2$DS$_2$-VASc Risk</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHF or LVEF $\leq$ 40%</td>
<td>1</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1</td>
</tr>
<tr>
<td>Age $&gt; 75$</td>
<td>2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>Stroke/TIA/Thromboembolism</td>
<td>2</td>
</tr>
<tr>
<td>Vascular Disease</td>
<td>1</td>
</tr>
<tr>
<td>Age 65 - 74</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
</tr>
</tbody>
</table>
Reviews of published prediction models

- 800 models in individuals with CVD (Sessler 2015)
- 396 models for predicting cardiovascular disease (Damen 2016)
- 111 models for prostate cancer (Shariat 2008)
- 102 models for TBI (Perel 2006)
- 83 models for stroke (Counsell 2001)
- 54 models for breast cancer (Altman 2009)
- 43 models for type 2 diabetes (Collins 2011; van Dieren 2012)
  - 30+ more models have since been published!
- 31 models for osteoporotic fracture (Steurer 2011)
- 29 models in reproductive medicine (Leushuis 2009)
- 26 models for hospital readmission (Kansagara 2011)

Courtesy of Gary Collins
Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review

Benjamin A Goldstein¹,², Ann Marie Navar²,³, Michael J Pencina¹,², John PA Ioannidis⁴,⁵

ABSTRACT

Objective Electronic health records (EHR) are rich in clinical and biomedical information. Our objective was to examine the opportunities and challenges of using EHR data for developing risk prediction models.

Methods We searched PubMed for relevant articles published between 2009 and 2014. A systematic review of clinical and biobehavioral research in electronic health records was performed.

Results We identified 107 articles from 15 different countries. Studies were generally very large (median sample size = 26100) and utilized a diverse array of predictors. Most used validation techniques (n = 94 of 107) and reported model coefficients for reproducibility (n = 83). However, studies did not fully leverage the breadth of EHR data, as they uncommonly used longitudinal information (n = 37) and employed relatively few predictor variables (median = 27 variables). Less than half of the studies were multicenter (n = 50) and only 26 performed validation across sites. Many studies did not fully address biases of EHR data such as missing data or loss to follow-up. Average c-statistics for different outcomes were: mortality (0.84), clinical prediction (0.83), hospitalization (0.71), and service utilization (0.71).

Conclusions EHR data present both opportunities and challenges for clinical risk prediction. There is room for improvement in designing such studies.

- Median of 27 predictor variables
- Median sample size 26100
- 26/107 external validation
- Longitudinal information is not used
Current status of prediction modelling

- Inadequate internal validation
- Small sets of features
- Incomplete dissemination of model and results
- No transportability assessment
- Impact on clinical decision making unknown

Relatively few prediction models are used in clinical practice
Mission for Patient-Level Prediction

OHDSI aims to develop a systematic process to learn and evaluate large-scale patient-level prediction models using observational health data in a data network.

- Evidence Generation
- Evidence Evaluation
- Evidence Dissemination
**Problem pre-specification.** A study protocol should unambiguously pre-specify the planned analyses.

**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.
Data is extracted from the OMOP CDM using the Feature Extraction R-Package.

Data characterization is required before modelling. Tools are being developed in the community to facilitate this.

A data cleaning step is recommended, e.g. to remove outliers in lab values.
Model training and Internal validation is done using a train test split:

1. Person split: examples are assigned randomly to the train or test set, or 

2. Time split: a split is made at a moment in time (temporal validation)

Train set                      Test set  
2014-01-15
Model Training

1. Which models?

2. How to evaluate the model?
Model training is an empirical process in which multiple models are compared.

### Regularized Logistic Regression

\[ P(y) = \frac{1}{1 + e^{-(x_1 b_1 + x_2 b_2 + \ldots + x_n b_n)}} \]

### Random Forest

- Random subset of patients and features per tree
- Forest Majority Vote

### Gradient Boosting Machines

- Reweight
- Reweight
- Combine

Many other models for example:

- K-nearest neighbors
- Naïve Bayes
- Support Vector Machines
- Etc.
Patient-Level Prediction Roadmap

Evidence Generation
Protocol Sharing
CDM Extractions
Code Sharing
Train / Test split

Evidence Evaluation

Evidence Dissemination
What makes a good model?

**Discrimination**: differentiates between those with and without the event, i.e. predicts higher probabilities for those with the event compared to those who don’t experience the event

**Calibration**: estimated probabilities are close to the observed frequency
How to assess discrimination?

Suppose our classifier is simply BMI > x.

Both classes (blue = 0, red = 1) have their own probability distribution of BMI.

The choice of X then determines how sensitive or specific our algorithm is.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

True Positive Rate (TPR) = TP / (TP + FN)
False Positive Rate (FPR) = FP / (FP + TN)
The Receiver Operator Curve (ROC) is developed during World War II for the analysis of radar images. Radar operators had to decide whether a blip on the screen represented an enemy target, a friendly ship, or just noise.
Calibration Assessment

How close is the average predicted probability to the observed fraction with the outcome?

- Underestimation
- Overestimation
External validation is performed using data from multiple populations not used for training.
Patient-Level Prediction Roadmap

Evidence Generation
- Protocol Sharing
- CDM Extractions
- Code Sharing
- Train / Test split

Evidence Evaluation
- Standardized Process
- Discrimination
- Calibration
- External Validation

Evidence 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.

- Internal and external validation
- Sharing of full model details
- Sharing of all analyses code to allow full reproducibility

Website to share protocol, code, models and results for all databases

Patient-Level Prediction Roadmap

Evidence Generation
- Protocol Sharing
- CDM Extractions
- Code Sharing
- Train / Test split

Evidence Evaluation
- Standardization
- Discrimination
- Calibration
- External Validation

Evidence Dissemination
- Publications (TRIPOD)
- Model sharing
- Full transparency
Large-scale patient-level prediction

A case study: prediction in patients with Pharmaceutically Treated Depression
Objectives

• Assess the feasibility of large-scale predictive model development

• Investigate the performance of different classifiers across the outcomes and databases

• Initiate an assessment across the OHDSI data network
Problem definition

Among patients in 4 different databases, we aim to develop prediction models to predict which patients at a defined moment in time (First Pharmaceutically Treated Depression Event) will experience one out of 22 different outcomes during a time-at-risk (1 year). Prediction is done using all demographics, conditions, and drug use data prior to that moment in time.
At Risk Cohort Definition

Patients are included in the cohort of interest at the date of the first occurrence of Pharmaceutically Treated Depression if the following inclusion criteria apply:

1. At least 365 days of history

2. At least 365 days of follow-up or the occurrence of the outcome of interest

3. No occurrence of the event prior to the index date
Setting

### Databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Depression</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAE</td>
<td>659402</td>
<td>1351</td>
</tr>
<tr>
<td>MDCD</td>
<td>79818</td>
<td>356</td>
</tr>
<tr>
<td>MDCR</td>
<td>57839</td>
<td>874</td>
</tr>
<tr>
<td>OPTUM</td>
<td>363051</td>
<td>1183</td>
</tr>
</tbody>
</table>

### Data extraction

- All demographics, conditions, drugs
- All 22 outcome cohorts

### Training and testing

- Time split for training and testing
- Transportability for Stroke

### Models

- Gradient Boosting
- Random Forest
- Regularized Regression

### Outcomes

- Acute liver injury
- Acute myocardial infarction
- Alopecia
- Constipation
- Decreased libido
- Delirium
- Diarrhea
- Fracture
- Gastrointestinal hemorrhage
- Hyperprolactinemia
- Hyponatremia
- Hypotension
- Hypothyroidism
- Insomnia
- Nausea
- Open-angle glaucoma
- Seizure
- Suicide and suicidal ideation
- Tinnitus
- Ventricular arrhythmia and sudden cardiac death
- Vertigo
Regularized Regression on CCAE

Receiver Operator Curve

Threshold: 0.01
Sensitivity: 0.25
Specificity: 0.97

AUC = 0.797

Calibration plot

Slope = 0.783
So what IS the model?

Reminder:

CHA$_2$DS$_2$-VASc is a model in clinical practices, but it was designed and tested for patients with Atrial Fibrillation to predict stroke, not for patients with depression and not for incident strokes....

The variables in this score were:
Age, Gender, Congestive Heart Failure, Hypertension, Diabetes, Vascular disease

Did our model pick those variables automatically from the data?
CHA$_2$DS$_2$-VASc variables

Prevalence in patients without the outcome

More prevalent in patients with the outcome

More prevalent in patients without the outcome
All variables explored in a large-scale model

The OHDSI approach lets the model choose from all conditions and drugs

247 variables out of 16900 including:
1. all the CHADS2 markers
2. plus some other variables that make clinical sense (ex: brain cancer, smoking)
3. plus some other variables that warrant further exploration (ex: antiepileptic, COPD)

Size: value  
Red: positive  
Green: negative
Model Discrimination Stroke

- Gradient Boosting
- Random Forest
- Regularized Regression

AUC
- CCAE
- MDCD
- MDCR
- OPTUM
Model Discrimination

Outcomes

Gradient Boosting
Random Forest
Regularized Regression

Low performance on MDCR
Some outcomes we can predict very well, some we cannot.
Outcomes with AUC > 0.75

Best performing is Regularized Regression on CCAE for Acute Myocardial Infarction
AUC = 86.32
Model Discrimination

Outcomes

Discrimination of different algorithms is comparable
Model Discrimination

Outcomes

- Gradient Boosting
- Random Forest
- Regularized Regression

But not always! For open-angle glaucoma, Gradient Boosting is better.
Transportability Assessment

How well do the models perform on other databases?
<table>
<thead>
<tr>
<th>CCAE</th>
<th>MDCD</th>
<th>MDCR</th>
<th>OPTUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regularized Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transportability to MDCR is low
Transportability between CCAE and OPTUM is very good.
What did we achieve so far?

We showed it is feasible to develop large-scale predictive models for all databases converted to the OMOP CDM. This can now be done for any cohort at risk, outcome, and time at risk.
Continuation of the PLP Journey

Scale up
• Increase the number of database
• Increase the number of cohorts at risk
• Increase the number of outcomes

Method Research
• Performance
• Speed
• Transportability
• Temporal information
• Textual information
• ...

Clinical impact for the patient
• How to assess?
We need you!

• We need contributions from many disciplines: clinicians, statisticians, machine learning experts, data custodians etc.

• Join the large-scale patient prediction study.


p.rijnbeek@erasmusmc.nl
jreps@its.jnj.com
Posters and Demo

• In the afternoon visit the demo of the Patient-Level Prediction R-package

• Visit our posters:
  1. **Best Practices for Patient-Level Prediction in OHDSI**
  2. **Utilizing the OHDSI collaborative network for large-scale prognostic model validation**
Join the journey!

The Journey toward Patient-Level Prediction