Large-Scale Predictive Modelling using Electronic Health Records

Peter R. Rijnbeek, PhD
Department of Medical Informatics
Erasmus MC Rotterdam,
The Netherlands
Dr. Rijnbeek works for a research group who received unconditional research grants from Boehringer-Ingelheim, Novartis, Pfizer, Yamanouchi, Servier, GSK, Johnson & Johnson.
Massive numbers of electronic health records (EHR) are currently being collected globally in observational databases, including structured data in the form of diagnoses, medications, laboratory test results, and unstructured data contained in clinical narratives. This opens unprecedented possibilities for research and ultimately patient care.

Challenges

Observational databases differ in both purpose and design. Each has different logical organizations and physical formats, and the terminologies used to describe the medicinal products and clinical conditions vary from source to source.

We need to standardize
Translation to a common data model and standard vocabularies

Any common data model aims to achieve both syntactic and semantic operability.

**syntactic operability:**
common underlying data structure
(standard grammar)

**semantic operability:**
common understanding required to interchange information
(standard vocabulary)
The OMOP CDM and OHDSI

Observational Health Data Sciences and Informatics (OHDSI) has been established as a multi-stakeholder, interdisciplinary collaborative to create open-source solutions for large-scale analytics using the OMOP CDM. [http://ohdsi.org](http://ohdsi.org)

OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

Deep information model
OMOP CDM v5.0.1

Standardized clinical data
- Person
  - Observation_period
  - Specimen
  - Death
  - Visit_occurrence
  - Procedure_occurrence
  - Drug_exposure
  - Device_exposure
  - Condition_occurrence
  - Measurement
  - Note
  - Observation
  - Fact_relationship

Standardized derived elements
- Cohort
- Cohort_attribute
- Condition_era
- Drug_era
- Dose_era

Standardized health system data
- Location
- Care_site
- Provider
- Payer_plan_period
- Cost

Standardized economics

Standardized meta-data
- CDM_source

Standardized vocabularies
- Concept
- Vocabulary
- Domain
- Concept_class
- Concept_relationship
- Relationship
- Concept_synonym
- Concept_ancestor
- Source_to_concept_map
- Drug_strength
- Cohort_definition
- Attribute_definition

Standardized health system data
- Device_exposure
- Procedure_occurrence
- Drug_exposure
- Condition_occurrence
- Measurement
- Note
- Observation
- Fact_relationship
OHDSI Collaborators:

- >140 researchers in academia, industry, government, health systems
- >20 countries
- Multi-disciplinary expertise: epidemiology, statistics, medical informatics, computer science, machine learning, clinical sciences

Databases converted to OMOP CDM within OHDSI Community:

- >50 databases
- >660 million patients
Clinicians are confronted with prediction questions on a daily basis. What options do they have?

Deny ability to predict at the individual patient level

Quote an overall average to all patients

Utilize knowledge and personal experience

Provide a personalized prediction based on an advanced clinical prediction model
Problem definition

Among a population at risk we aim to predict which patients at a defined moment in time (t=0) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.
Growing interest in prediction modelling
Current Stroke Guidelines


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 $\geq$ 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>
Current status of prediction modelling

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 (EHRs) offer unique opportunities and challenges for developing prediction models with clinical outcomes. We conducted a systematic review of clinical risk prediction studies using EHR data.

Methods: We searched PubMed and Ovid Medline and Scopus for studies published from 2009 to 2014. Articles were included if they reported on clinical risk prediction models developed using EHR data. Studies were excluded if they were not in English or if they did not report on clinical outcomes.

Results: We identified 107 articles from 15 different countries. Studies were generally very large (median sample size = 26100) and utilized a diverse array of predictor variables. 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
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.
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)

<table>
<thead>
<tr>
<th>Train set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-01-15</td>
<td></td>
</tr>
</tbody>
</table>
1. Which models?

2. How to evaluate the model?
Models

Model training is an empirical process in which multiple models are compared.

Regularized Logistic Regression

\[ x_1 \rightarrow b_1 \rightarrow + \rightarrow \frac{1}{\sigma + \exp(b_1 + b_2 + \ldots + b_n)} \text{ (for } y = 1) \]

Random Forest

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

Gradient Boosting Machines

Many other models for example:

- K-nearest neighbors
- Naïve Bayes
- Support Vector Machines
- Deep Learning
- Etc.
Model Validation

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

etc.
External validation is performed using data from multiple populations not used for training.
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

Large-scale patient-level prediction

A case study: prediction in patients with Pharmaceutically Treated Depression
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
Model Discrimination Stroke

Gradient Boosting
Random Forest
Regularized Regression

AUC
1.00 0.90 0.80 0.70 0.60 0.50
Model Discrimination

Outcomes

Gradient Boosting
Random Forest
Regularized Regression

Low performance on MDCR

AUC

1.00
0.90
0.80
0.70
0.60
0.50

OPTUM
MDCD
MDCR
CCAЕ
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

<table>
<thead>
<tr>
<th>Model Type</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting</td>
<td>1.00</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.90</td>
</tr>
<tr>
<td>Regularized Regression</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Discrimination of different algorithms is comparable.
## Model Discrimination

Outcomes

<table>
<thead>
<tr>
<th>Method</th>
<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>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regularized Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But not always! For open-angle glaucoma Gradient Boosting is better
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
• Deep Learning
• Learning Curves
• ...

Clinical impact for the patient
• How to assess?
We need contributions from many disciplines: clinicians, statisticians, machine learning experts, data custodians etc.


p.rijnbeek@erasmusmc.nl
jreps@its.jnj.com
First Annual
EUROPEAN OHDSI
SYMPOSIUM
March 23th 2018
Tutorials March 24th

Bridging Europe

Erasmus MC Rotterdam The Netherlands
When and where

Department of Medical Informatics
Erasmus MC Rotterdam, The Netherlands

• March 23th 2018 OHDSI Symposium
• March 24th 2018 Tutorials
• Max 250 participants
• Poster sessions
• www.ohdsi-europe.org
• email: info@ohdsi-europe.org

Registration for the Symposium is open.

Registration for tutorials will be announced through the website and OHDSI communication channels.