The Development and Validation of a Predicting Model by means of a Learning Curve

**Authors:**

L. H. John, MSc, Erasmus University Medical Center, Rotterdam, The Netherlands

P. R. Rijnbeek, PhD, Erasmus University Medical Center, Rotterdam, The Netherlands

**Date:** 26 February 2018

**Acknowledgement:** The analysis is based in part on work from the Observational Health Data Sciences and Informatics collaborative. OHDSI (<http://ohdsi.org>) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

The authors declare the following disclosures: P. R. Rijnbeek works for a research group who received unconditional research grants from Boehringer-Ingelheim, Novartis, Pfizer, Yamanouchi , Servier, GSK, Janssen Research and Development. None of these are related to this study.

Table of Contents

[2 List of abbreviations 3](#_Toc504400916)

[3 Abstract 3](#_Toc504400917)

[4 Amendments and Updates 3](#_Toc504400918)

[5 Milestones 4](#_Toc504400919)

[6 Rationale and Background 5](#_Toc504400920)

[7 Research Questions and Objectives 7](#_Toc504400921)

[8 Research methods model development (Part 1) 8](#_Toc504400928)

[8.1 Study Design 8](#_Toc504400930)

[8.1.1 Overview 8](#_Toc504400931)

[8.1.2 Study population 9](#_Toc504400932)

[8.1.3 Additional analysis details 9](#_Toc504400933)

[8.2 Variables 9](#_Toc504400942)

[8.2.1 At risk 9](#_Toc504400943)

[8.2.2 Outcomes 10](#_Toc504400946)

[8.3 Data Sources 10](#_Toc504400947)

[8.4 Quality control 10](#_Toc504400948)

[8.5 Strengths and Limitations of the Research Methods 10](#_Toc504400949)

[9 Research methods validation of existing models for HF in T2DM patients (Part 2) 11](#_Toc504400950)

[10 Protection of Human Subjects 11](#_Toc504400951)

[11 Plans for Disseminating and Communicating Study Results 12](#_Toc504400952)

[12 References 12](#_Toc504400953)

# List of abbreviations

ATC Anatomic Therapeutic Chemical

CYCLOPS Cyclic coordinate descent for logistic, Poisson and survival

OHDSI Observational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

T2DM Type II Diabetes Mellitus

HF Heart Failure

DHF Diastolic Heart Failure

SHF Systolic Heat Failure

# Abstract

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.

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.

This proof of concept study aims to investigate the feasibility of large-scale predictive modelling using observational healthcare data in the OHDSI network. In this study patient-level prediction models will be trained and evaluated for multiple health outcomes in patients with pharmaceutically treated depression.

The models will be developed using the Patient-Level Prediction Package developed in OHDSI. The discriminative performance and calibration scores of the models will be assessed in the OHDSI data network. This study will develop prediction models for 22 outcomes across the OHDSI collaborative network datasets.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author(s) | Comments |
| 0.1 | 22 Jan 2018 | Peter Rijnbeek | Initial Draft |
|  |  |  |  |
|  |  |  |  |

# Milestones

|  |  |
| --- | --- |
| Milestone | Planned / Estimated Date |
| Start of analysis |  |
| End of analysis |  |
| Posting of results |  |
| Submission of manuscript |  |

# Rationale and Background

Type 2 Diabetes Miletus (T2DM) and Heart Failure (HF) are two major and growing contributors to the global burden of disease both in developed and developing countries.

As an example for the European situation, the recent study in the UK [Sharma M et al.] demonstrated that the incidence of T2DM rose from 3.69 per 1000 person-years at risk (PYAR) (95% CI 3.58 to 3.81) in 2000 to 3.99 per 1000 PYAR (95% CI 3.90 to 4.08) in 2013 among men; and from 3.06 per 1000 PYAR (95% CI 2.95 to 3.17) to 3.73 per 1000 PYAR (95% CI 3.65 to 3.82) among women Prevalence of T2DM more than doubled from 2.39% (95% CI 2.37 to 2.41) in 2000 to 5.32% (95% CI 5.30 to 5.34) in 2013. This has a high impact on the National Health Service budget in the UK. In the U.S. the 2017 National Diabetes Statistics Report ([link](https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf)) created by the Center of Disease Control (CDC) showed that an estimated 30.3 million people of all ages, i.e. 9.4% of the U.D. population had diabetes in 2015. The percentage of adults with diabetes increased with age, reaching a high of 25% among those aged 65 years or older.

HF incidence and prevalence is estimated…[To ADD]

Diabetes and HF frequently co-exists and research in their bi-directional relationship is currently given a lot of attention []. Multiple studies have shown HF is a major contributor to cardiovascular morbidity and mortality in patients with diabetes, but also conversely those with heart failure are at higher risk of developing diabetes []. There is an increased recognition that diabetic patients develop heart failure independent of the presence of coronary artery disease or its associated risk factors []. Patients with diabetes are much more likely to develop congestive heart failure than patients without diabetes (incidence rate 30.9 vs. 12.4 cases per 1,000 person-years) [Nichols GA et al.] Heart failure and peripheral arterial disease are the most common initial manifestations of cardiovascular diseases in patient with T2DM [Shah AD et al.]

As described in detail in the recent review of Tousoulis et al [], the pathophysiological connection between both disease and their frequent adverse interactions should impact treatment choice. Interestingly, current guidelines for heart failure therapy from the European [Ponikowski P] as well as the American cardiology societies [Yancy C] do not recommend specific therapeutic approaches in patients with diabetes compared with subjects without diabetes. The current Diabetes Guidelines [[link](http://care.diabetesjournals.org/content/diacare/suppl/2016/12/15/40.Supplement_1.DC1/DC_40_S1_final.pdf)] do mention that Metformin is contraindicated in patients with significant HF, Thiazolidinediones should be used very cautiously in those with or at risk for CHF. Patients with diabetes and heart failure may benefit most from glucose-lowering therapies with SGLT2 inhibition [Lehrke et al]. In contrast, no improvement in heart failure, or potential detrimental effects, have been reported for glucose-lowering strategies that directly or indirectly increase the availability of insulin. Meta-analyses showed that intensive glucose lowering is not associated with any significant reduction in cardiovascular risk but results in a significant increase in heart failure risk []. The importance of HF risk assessment in the T2DM population was well demonstrated by the withdrawal of rosiglitazone, a thiazolidinedione, from the EU market due to evidence of increased risk of hospitalizations for HF[].

Patient-level prediction of the risk of HF in T2DM can potentially help to define the optimal therapeutic strategy and can improve healthcare outcome. For this, the detection of both high-risk patients as well as low risk patients is of importance. Unfortunately, the number of prediction models for HF in T2DM patients is still very limited []. Research in prediction of cardiovascular risk has mainly focused on the general population [ref Collins, paper on forgotten comorbidity].

To develop a prediction model for HF in T2DM we have to follow some best-practices. First, to obtain accurate and well calibrated models we need large and rich datasets. The number of persons in the cohort at risk (T2DM) and the number of patients that develop HF should be high enough. Furthermore, the inclusion of potential predictors should ideally not be expert based but data driven to utilize the full medical history of the patients. Second, the model needs to be externally validated on a large set of data sources. This very important step is often lacking in predictive modelling studies because data is not readily accessible. Third, model selection is in our view an empirical question that needs to be answered by systematically assessing a broad set of machine learning algorithms and design choices. Fourth, we should enable external validation by others by using a fully transparent and reproducible pipeline, including sharing of all model details. Many prediction models in literature fail on one or more of these best practices which limits their clinical uptake considerably [Murphy paper].

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 to develop patient-level prediction models demands a standardized pipeline for both model development and evaluation.

A patient-level prediction model problem is defined by an ‘at risk’ cohort (the group of people we wish to do the prediction for), the ‘outcome’ cohort (the outcome we wish to predict) and the ‘at-risk’ period (time window relative to the start of the at risk cohort index date). In this study we focus on a single ‘at-risk’ cohort (Type II Diabetes) and a single ‘outcome’ cohort HF and we use a ‘at-risk’ period of 1 year. The study will be implemented across the OHDSI collaborator network to externally validate the models and evaluate their transportability across the world.

A secondary goal in our study is to validate known models. We like to create insight in how reproducible these models are based on the available information in the papers.

[Overview of literature on HF prediction in T2DM -> Ross can you add some detail]

# Research Questions and Objectives

Our study consists of two parts:

1. **Development and validation of a new prediction model for HF in T2DM patients.**

In this part of the study we will use all available data sources to train and validate a prediction model. We will use all available data in the electronic health record prior to the incident T2DM event. In this part of our study we have the following research questions:

* 1. Are we able to develop models to predict HF in T2DM patients with high discriminative abilities that are well calibrated?
  2. Are these models transportable to other databases not used for training (external validation)?

1. **Validation of existing models for HF in T2DM patients**

We want to investigate the performance of various patient level prediction models to predict HF in patients with T2DM and compare these with our models. In this part we have the following research questions:

* 1. Are the models currently available in literature externally valid?
  2. Is the model developed use our data-driven approach outperforming the currently available models?

# Research methods model development (Part 1)

## Study Design

### Overview

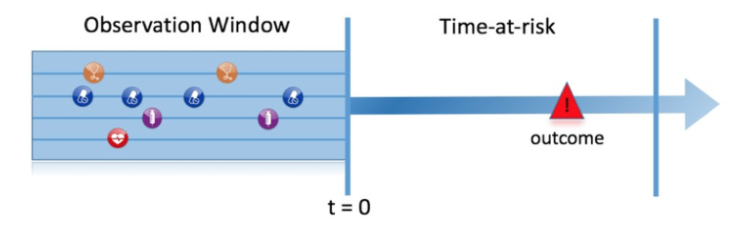


Figure 2. The prediction problem

For a single ‘at risk’ cohort consisting of people with T2DM we will develop prediction models that predict the occurrence of HF within a year of the start of the initial T2DM diagnosis. We will assess five different classifiers to also investigate database and method differences for each prediction problem. This will be distributed across the OHDSI network observational databases.

The prediction problem will leverage the full medical history at the start of the T2DM diagnosis, i.e. demographics, conditions and drugs. The classification outcome labels will be constructed by setting all the ‘at risk’ people who develop the outcome as class 1 and all the ‘at risk’ people who do not develop the outcome as class 0. The classifiers will then use the labeled data (features paired with the label) to learn patterns that predict the class of a person based on their features up to the start of the depression episode (the index date). We will use 75% of the labelled data to train the models and the remaining 25% of the data will be used to perform an internal validation of the prediction model. This 25% of the data will not be seen by the classifier when it is being trained. We will perform 3-fold cross validation when training the classifiers to identify the optimal hyper-parameters for each classifier.

The following algorithms will be used in this study: regularized logistic regression, random forest, gradient boosting machines. Details on these models can found elsewhere.

To internally evaluate the models on the test set, the area under the receiver operating characteristic curve (AUC) is determined to evaluate the discriminative performance of the models. Additionally, model calibration is assessed by plotting the predicted risk against the observed fraction and the brier score for general prediction performance is calculated. The models are externally validated by applying each trained model to the other participating databases.

### Study population

All subjects in the database will be included who meet the following criteria: (note: the index date is the start of the T2DM era)

* Recorded condition of T2DM
* At least 365 days of observation time prior to the index date
* No diagnose of HF preceding the index date

### Additional analysis details

The lasso logistic regression risk models will be fitted using a regularized regression with a LaPlace prior. The regularization hyperparameter will be selected by optimizing the likelihood in a 3-fold cross-validation.

The random forest hyper-parameter will be selected by performing a 3-fold cross validation using a grid search on the following parameters space:

* ntrees =50, 500 or 1000
* mtries = sqaure root of number of features, 50 or 500
* max\_depth =4,10 or 17
* varImp =True or False (if True a preliminary step is included where an initial random forest is trained and the features with a feature importance score greater than the median across all features are only included in the main model)

The gradient boosting machine uses decision trees and the hyper-parameter will be selected by performing a 3-fold cross validation using a grid search on the following parameters space:

* ntrees = 50, 100 or 250
* max\_depth = 4, 7 or 16
* min\_rows = 1 or 20

## Variables

### At risk

#### T2DM

Index rule defining the index date:

* First condition record of T2DM

Inclusion rules based on the index date:

* At least 365 days of observation time prior to the index date
* At least 30 days of observation time after the index date
* Must have no condition records of X recorded within any days prior to index date

### Outcomes

#### Heart Failure

Index rule defining the index date:

* First occurrence of a heart failure code in the concept set specified in the Appendix X.

Inclusion rules based on the index date:

* None

## Data Sources

The analyses will be performed across a network of observational healthcare databases. All databases have been transformed into the OMOP Common Data Model version 5. The complete specification for OMOP Common Data Model, version 5 is available at: <https://github.com/OHDSI/CommonDataModel>.

The initial experiment will be performed on CCAE, MDCD, MDCR, OPTUM.

## Quality control

We will investigate the risk model by

* Calculating the calibration and discrimination measures and comparing against any existing model (identified using a literature search) benchmarks
* Comparison of performance between methods to identify any outlier models. Any outliers will be inspected to check for potential errors.
* Determining the robustness of outcome specific models across the OHDSI datasets
* Inspection of the fitted lasso regularized logistic regression outcome model for large coefficients

The PatientLevelPrediction package itself, as well as other OHDSI packages on which PatientLevelPrediction depends, follow OHDSI’s best practice for code development which includes, e.g., use unit testing.

## Strengths and Limitations of the Research Methods

Strength

* The analysis can help us gain insight into the feasibility of the usage of observational data for predicting HF in the T2DM population.
* The analysis can help gain insight into to suitability of each classifier when applied for prediction modelling using observational data.
* The analysis can be readily implemented across the OHDSI network – this will give us insight into how much models vary across datasets and how transportable models are.

Limitations

* The classifiers cannot include censored patients, so any person who does not get observed for the complete follow-up period or experiences the outcome during follow-up is excluded. This can cause a bias in the population used to develop and model and may prevent the model being applicable to the general population.
* Not all medical events are recorded into the observational datasets and some recordings can be incorrect. This results in a noisy dataset with potential outcome misclassification. It is unknown to what extent misclassification of any of the outcomes occurs.
* Some of the classifiers have a large number of hyper-parameters and performing a large search for the optimal hyper-parameter may not be possible due to time limits. This may result in sub-optimal performance for certain classifiers.
* Although the CDM standardizes the vocabularies of the datasets, the concept recording distributions are likely to differ between databases and it is unknown how much this will limit model transportability.

# Research methods validation of existing models for HF in T2DM patients (Part 2)

A literature study on HF prediction in T2DM resulted in the models described in table below.

|  |  |  |  |
| --- | --- | --- | --- |
| First Author | Title | Model | Performance |
| Yang, X et al. | Development and validation of a risk score for hospitalization for heart failure in patients with Type 2 diabetes mellitus |  |  |
|  |  |  |  |

To replicate these models, we need to define the concept sets for each of the predictors used in these models, run these models in the different databases, and compare the predictions with the observed frequency. The concept sets are available in the Appendix of the protocol.

In the Patient-Level Prediction R Package functionality has been added to define the covariates and specify the betas in logistic regression models described in the papers.

# Protection of Human Subjects

The study is using only de-identified data. Confidentiality of patient records will be maintained at all times. All study reports will contain aggregate data only and will not identify individual patients or physicians.

# Plans for Disseminating and Communicating Study Results

The study results will be posted on the OHDSI website after completion of the study. At least one paper describing the study and its results will be written and submitted for publication to a peer-reviewed scientific journal.

# References

To be added!