



# EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK


806968 – EHDEN

European Health Data & Evidence Network

WP3 – Personalized Medicine

## D3.6 Third Report on the Implementation of the Analytical Pipeline for Personalized Medicine


Lead contributor	Peter Rijnbeek (1 – EMC)
Lead contributor email	p.rijnbeek@erasmusmc.nl
Other contributors	Alexandros Rekkas, Ross Williams, Luis H John, Aniek Markus, Cynthia Yang, Tom Seinen, Egill Fridgeirsson, Jan Kors, Maria de Ridder (1 – EMC) Daniel Prieto Alhambra (3 – UOXF) Sulev Reisberg (4 - UTARTU) Mark Kruger (16 -SANOFI)
Due date	31/10/2021
Delivery date	22/12/2021
Deliverable type	R
Dissemination level	PU
DoA - Version	V2
Date	15/10/2021

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>		<b>Version: v1.1 – Final</b>
	<b>Author(s): Peter Rijnbeek et al.</b>		<b>Security: PU</b>   2/27



## Table of Contents


<b>Document History</b> .....	<b>3</b>
<b>Definitions</b> .....	<b>4</b>
<b>Publishable Summary</b> .....	<b>5</b>
<b>1. Introduction</b> .....	<b>6</b>
<b>2. Methods Research</b> .....	<b>7</b>
2.1 Exhaustive search for optimal decision rules.....	7
2.2 Association Rule and Frequent Pattern Mining .....	9
2.4 Predictive analytics using unstructured data.....	12
2.5 Deep Learning – Attention based models.....	13
2.6 Iterative Pairwise External Validation.....	15
2.7 Continuous risk-based assessment of treatment effect heterogeneity.....	17
2.8 Disease Trajectories .....	19
2.9 Systematic review on clinical prediction modelling .....	21
2.10 Prediction Model Library.....	22
<b>3. Next Steps</b> .....	<b>25</b>
<b>References</b> .....	<b>26</b>

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>		<b>Version:</b> v1.1 – Final
	<b>Author(s):</b> Peter Rijnbeek et al.		<b>Security:</b> PU



## DOCUMENT HISTORY

Version	Date	Description
V1.0	28 November 2021	Final Draft for internal review
V1.0	6 December 2021	Advanced draft for formal review and consortium review
V1.1	22 December 2021	Final version

 <b>EHDEN</b> <small>EUROPEAN HEALTH DATA &amp; EVIDENCE NETWORK</small>	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>		<b>Version:</b> v1.1 – Final
	<b>Author(s):</b> Peter Rijnbeek et al.		<b>Security:</b> PU




## DEFINITIONS

Participants of the EHDEN Consortium are referred to herein according to the following codes:

<b>EMC</b>	Erasmus Universitair Medisch Centrum Rotterdam- The Netherlands <b>(Project Coordinator)</b>
<b>Synapse</b>	Synapse Research Management Partners S.L. - Spain
<b>UOXF</b>	The Chancellor, Masters and Scholars of the University of Oxford - United Kingdom
<b>UTARTU</b>	Tartu Ulikool - Estonia
<b>UAVR</b>	Universidade de Aveiro – Portugal
<b>The Hyve</b>	The Hyve BV – the Netherlands
<b>Odysseus</b>	Odysseus Data Services SRO – Czech Republic
<b>EPF</b>	Forum Europeen des Patients (FPE) - Belgium
<b>NICE</b>	National Institute for Health and Care Excellence – United Kingdom
<b>UMC</b>	Stiftelsen WHO Collaborating Centre for International Drug Monitoring - Sweden
<b>ICHOM</b>	International Consortium for Health Outcomes measurement LTD - United Kingdom
<b>Janssen</b>	Janssen Pharmaceutica NV - Belgium <b>(Project Lead)</b>
<b>Pfizer</b>	Pfizer Limited – United Kingdom
<b>Abbvie</b>	AbbVie Inc - United States
<b>IRIS</b>	Institut De Recherches Internationales Servier - France
<b>SARD</b>	Sanofi Aventis Recherche & Developpement - France
<b>Bayer</b>	Bayer Aktiengesellschaft - Germany
<b>Lilly</b>	Eli Lilly and Company Limited – United Kingdom
<b>AZ</b>	AstraZeneca AB - Sweden
<b>Novartis</b>	Novartis Pharma AG - Switzerland
<b>UCB</b>	UCB Biopharma SPRL - Belgium
<b>Celgene / BMS</b>	Celgene Management SARL – Switzerland
<b>BI</b>	Boehringer Ingelheim International GmbH - Germany

<b>Grant agreement</b>	The agreement signed between the beneficiaries and the IMI JU for the undertaking of the EHDEN project (806968).
<b>Project</b>	The sum of all activities carried out in the framework of the Grant Agreement.
<b>Consortium</b>	The EHDEN Consortium, comprising the above-mentioned legal entities.
<b>Consortium agreement</b>	Agreement concluded amongst EHDEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s):</b> Peter Rijnbeek et al.	<b>Security:</b> PU	5/27



## PUBLISHABLE SUMMARY


The goal of WP3 “Personalized Medicine” is to establish a standardized process to enable personalized decision-making that can be utilized for multiple outcomes of interest and can be applied to observational healthcare data from any patient subpopulation.

In the first report (D3.2) on the implementation of the analytical pipeline for personalized medicine, we introduced the analytical pipelines for Patient-Level Prediction and Population-Level Effect Estimation. Furthermore, we discussed our initial work for the development of a pipeline for Risk Stratified Effect Estimation to assess heterogeneity of treatment effect.

In the second report (D3.4) we provided an update on the work done in the second year, including heterogeneity of treatment effect (HTE), and disease trajectory analyses. That deliverable included an overview of use cases in which the analytical pipelines have been applied and described the advances made in methodological research, the start of a natural language processing pipeline, and work done to develop a pipeline for disease trajectories.

In this third report we introduce new additional analytical tools and provide a further update on the methodological research.

This work falls under Task 3.2. “Development of an integrated patient-level prediction pipeline” (M6-M60), Task 3.3 “Development of an integrated risk-effect estimation pipeline” (M6-M60), and Task 3.4 “Development of a pipeline for disease trajectory analysis” (M12-M36).

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>		<b>Version:</b> v1.1 – Final
	<b>Author(s):</b> Peter Rijnbeek et al.		<b>Security:</b> PU 6/27

## 1. INTRODUCTION

The goal of WP3 is to build analytical pipelines that can utilize all the data in the OMOP Common Data Model (CDM) for patient-level prediction (PLP), population-level effect estimation (PLE), heterogeneity of treatment effect (HTE), and disease trajectory analyses. In this short deliverable we provide an update on the analytical pipeline development and methodological research.


In this reporting period many additional agile development steps have been made on the R packages that have been described in more detail in previous deliverables (see Table 1). This includes the addition of functionality, documentation updates, code restructuring, etc. The packages are currently hosted on GitHub within EHDEN, OHDSI, and the Erasmus MC GitHub organization but these will in the near future be located in the OHDSI repository. We refer to these repositories for more details on the release updates etc. For example, in the PatientLevelPrediction R package functionality was added to perform recalibration of predictive models as discussed as next step in deliverable D3.4 and logged in the [package news](#).

Table 1. Overview of R packages

Package	Description
<a href="#">OHDSI/PatientLevelPrediction</a>	An R package for performing patient level prediction in an observational database in the OMOP <a href="#">CDM</a> .
<a href="#">OHDSI/CohortMethod</a>	An R package for performing new-user cohort studies in an observational database in the OMOP CDM.
<a href="#">OHDSI/RiskStratifiedEstimation</a>	An R package for performing effect estimation in risk strata to assess treatment effect heterogeneity
<a href="#">EHDEN/Trajectories</a>	An R package for detecting and visualizing statistically significant event sequences in OMOP CDM data.
<a href="#">mi-erasmusmc/Triton</a>	An R package for creating covariates from unstructured text in OMOP CDM data.
<a href="#">mi-erasmusmc/Explore</a>	An R package for finding a short and accurate decision rule in disjunctive normal form by exhaustive search
<a href="#">mi-erasmusmc/AssociationRuleMining</a>	An R package that implements association rule mining and frequent pattern mining against the OMOP-CDM

Table 1 contains two new analytical tools, which we will describe in the methods research section below: one for searching exhaustively for optimal decision rules, and one for performing association-rule mining and frequent-pattern analysis.

For most of our work articles have already been submitted to peer-reviewed journals and made available as preprint for community feedback. Where applicable we have added links to these preprints, and other material such as posters, videos, and demos.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s):</b> Peter Rijnbeek et al.	<b>Security:</b> PU	7/27



## 2. METHODS RESEARCH

### 2.1 Exhaustive Search for Optimal Decision Rules

Many PLP models are developed and published in literature, but only few are used in clinical practice [1]. EXPLORE (Exhaustive Procedure for LOGic-Rule Extraction) is an exhaustive search algorithm designed to find optimal decision rules [2]. This algorithm has several features that make it attractive for PLP models. First, the resulting prediction model is a (short) decision rule and can thus be considered interpretable, which can contribute to create trustworthy AI [3]. Second, the exhaustive search nature of the algorithm allows users to specify additional constraints on the model (e.g. restricting the rule length and forcing certain features to be included in the model) and/or performance (e.g. minimum specificity and sensitivity). The aim of this work in EHDEN is to investigate the potential of EXPLORE for PLP models by comparing its performance with two more frequently used methods for PLP.

EXPLORE [2] generates decision rules of pre-specified length in disjunctive normal form (DNF). A formula in DNF is a disjunction of terms (OR,  $\vee$ ), where the terms are conjunctions (AND,  $\wedge$ ) of literals, and the literals are feature-operator-value triples ( $A > a$ ). An example of a DNF formula is  $(A > a \wedge B = b) \vee C \geq c$ ; the resulting decision rule has the form: if (DNF formula) then class = 1 else class = 0. To find the best decision rule, EXPLORE performs an exhaustive search of all possible rules of pre-specified length using smart techniques to reduce the search space. For example, by reducing the number of values that need to be checked for each feature (subsumption pruning) and disregarding subspaces that cannot contain the optimal decision rule (branch-and-bound). The exhaustive search approach guarantees that we find an optimal decision rule and allows users to specify additional constraints while optimizing over a chosen performance metric. For more details on the EXPLORE algorithm we refer to the original publication [2].

The R package that implements the EXPLORE algorithm, is currently under development and can be downloaded from GitHub: <https://github.com/mi-erasmusmc/explore/>.

Results on standard University of California Irvine (UCI) datasets show that EXPLORE can achieve similar performance for prediction problems compared to LASSO logistic regression and RandomForest (see Figure 1), while the model size is substantially smaller. Moreover, the results show EXPLORE's capability to learn under different types of constraints. Another benefit that should be further explored in future research is the possibility to add mandatory features in the model based on domain knowledge, to enhance the face value and generalizability of the resulting model.

The current results are limited in the sense that the studied prediction problems are simpler than real-world settings. Routinely collected health care data is typically more complex; it contains a much larger number of observations, thousands of features, and possibly more variation in the values of features. These are all factors that will influence the computational feasibility of EXPLORE. In the upcoming months we plan to investigate this further by adding EXPLORE to the PatientLevelPrediction package and evaluating the feasibility and performance on real-world clinical prediction problems using the Integrated Primary Care Information (IPCI) database mapped to OMOP CDM.

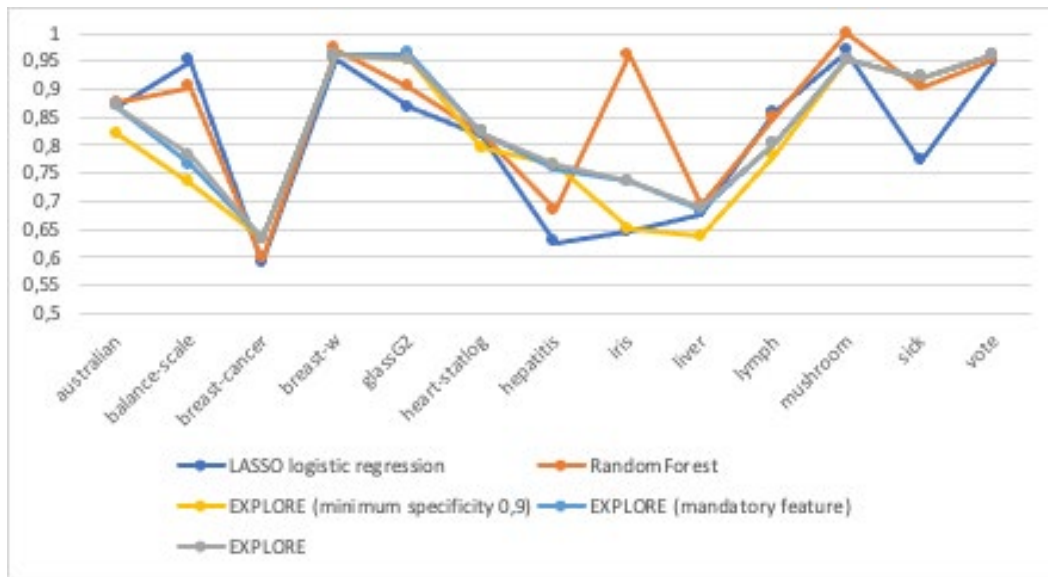



Figure 1: Comparison performance of AUC on standard UCI datasets between LASSO logistic regression, RandomForest, and EXPLORE (maximum rule length 3).



	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s): Peter Rijnbeek et al.</b>	<b>Security: PU</b>	9/27



## 2.2 Association-Rule and Frequent-Pattern Mining

Data-mining tasks aim to extract and analyze information to support decision making [4]. This includes pattern mining, dating back to the mid-1990s. Initially stated as a problem in the ‘market basket’ domain, pattern mining aims to discover structure and correlations in databases. Applying such methods to observational health data seems promising to reveal interesting and sometimes unexpected patterns [5]. For example, association-rule mining aims to answer the question, ‘Given a cohort of patients, which concepts are most likely to occur together?’ It could be used to measure the association between two or more concepts from any domain in the CDM, such as conditions, drugs, procedures, etc. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

*Table 2. Association rule and frequent pattern mining parameters*

Parameter	Description
minimum support	Threshold for the minimum number of patients who have the concept set in their medical history, e.g., (obesity, diabetes)
minimum confidence	Threshold for determining how often the left-hand side of a rule occurs together with the right-hand side, e.g., (obesity, diabetes) -> (heart failure)

Another example are frequent-pattern mining methods that take into account the chronological ordering of concepts. These methods can be used to answer the question, “What are the most frequent sequences of concepts observed in a cohort of patients?” Frequent patterns are required to satisfy a minimum support.

We are developing an open source R package called [AssociationRuleMining](#), which acts as a framework to perform association-rule and frequent-pattern mining analysis using data in the OMOP CDM. The framework provides an opportunity to assess the temporal structures of the medical history of patients, which can be used to characterize patients or can be used in PLP.

The AssociationRuleMining R package makes use of the open-source association rules implementation such as “Apriori” [6], “Eclat” [7], and “FP-Growth” [8] for mining highly associated sets of concepts, and “SPADE” [9], for mining frequent patterns. The AssociationRuleMining Package is fully integrated in the OHDSI framework using [DatabaseConnector](#) and [FeatureExtraction](#), and runs on any defined cohort. The resultant frequent patterns can be automatically added as custom covariates for use with other OHDSI packages, such as PatientLevelPrediction [10].

After execution, the generated association rules or frequent patterns can be explored in R. Depending on the size of the cohort and the settings, the number of extracted patterns can be very large. We are therefore working on methods to visualize the results interactively. We are exploring interactive networks (Figure 2) to visualize rules and interactive Sankey diagrams to visualize frequent patterns (Figure 3).

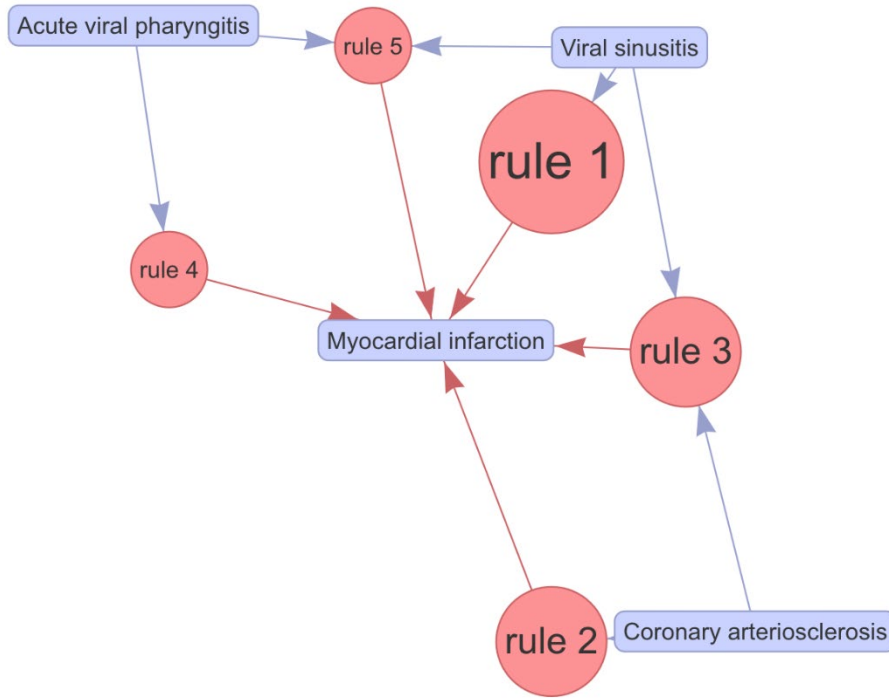


Figure 2: Top-5 association rules with highest support in a myocardial infarction cohort. A rule can contain multiple concepts (blue arrows) that are associated with an outcome (red arrow). The size of the red circles indicates the level of support (the bigger the circle the higher the support). For example, Coronary arteriosclerosis and Viral sinusitis were observed to be associated to Myocardial infarction with high support.

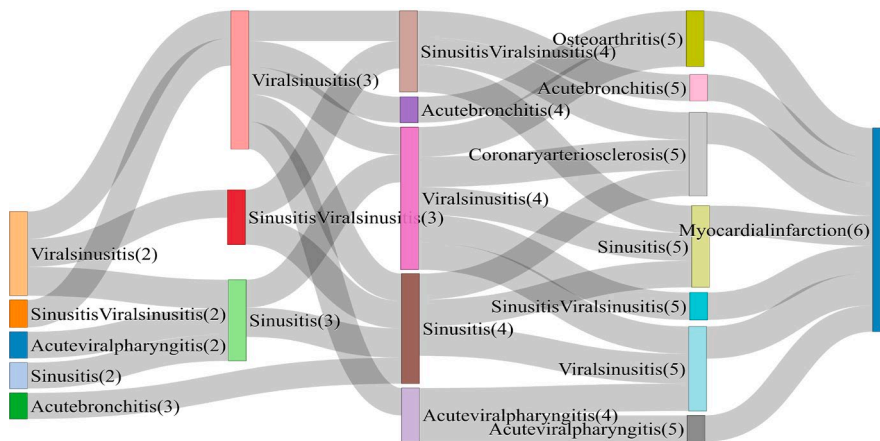



Figure 3: Sankey diagram showing frequent patterns in a myocardial infarction cohort. Connections between nodes indicate succession of events, in this case diagnoses.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s): Peter Rijnbeek et al.</b>	<b>Security: PU</b>	11/27



Our ultimate aim is to assess the value of different association-rule and frequent-pattern methods in prediction problems (see next section) and for characterization of patient populations (future work).

### **Predictive analytics using frequent patterns**

Electronic Health Records (EHRs) constitute a rich source of longitudinal information describing patients' encounters with the healthcare system. This longitudinal dimension, usually referred to as temporality, can be described as the timing and spacing of occasions of measurements over a period of time [11]. It can consist of a variety of information including and not limited to diagnoses, symptoms, drug prescriptions, and laboratory data, as well as unstructured clinical notes and/or radiological images. Temporal information therefore offers a unique opportunity to study the progress of health in patients; it is however a largely unexplored field.


Temporal pattern mining of EHR data has the potential to uncover previously unknown relationships among comorbidities (conditions occurring together and in a temporal order) and treatment pathways (drugs prescribed in a temporal order), which can complement clinical knowledge and traditional medical research methods [12]. PLP models that make use of standard machine learning algorithms usually do not incorporate any form of temporality: concepts are used as features that contain a single numeric or categorical value [13]. Previous work has attempted to introduce temporality as an additional attribute through temporal abstraction of consecutive events to be further included as features in prediction models with substantial improvement over their atemporal counterparts [14-16]. Further, various deep learning algorithms have been developed that are able to incorporate the temporal dimension [17, 18]. However, these models are suffering from issues of interpretability hindering further their adoption in clinical practice.

In this work we aim to assess the predictive value of frequent patterns as potential candidate predictors in clinical prediction models, using standard machine learning algorithms. Specifically, we aim to test several feature sets containing frequent patterns extracted from databases mapped to the OMOP CDM to evaluate their predictive value, using standard algorithms such as LASSO and xgboost. The candidate frequent patterns will be generated using the AssociationRuleMining package and model development and validation will be implemented using the [PatientLevelPrediction](#) package. We have fully integrated the two packages to execute this study.

We aim to generate and test the predictive value of several sets of candidate features. First, three different sets of frequent patterns consisting of the full, closed and maximal sets will be generated that represent frequent patterns as binary features. Second, frequent patterns will be represented as continuous features indicating the frequency of each pattern for each patient. Third, frequent patterns will be clustered and included as categorical variables representing cluster membership.

Additionally, we will consider further pre-processing steps. Initially, removing repeated concepts from the extracted frequent patterns can substantially remove redundancy and reduce the final set of candidate features. This will be achieved in two different ways. In the first, we will consider only the first occurrence of a concept from the medical history of patients. In the second, we will concatenate consecutive concepts from the resulting frequent patterns. Further, for sequential rules (rules with chronological ordered events) we make use of a minimum support and minimum confidence criterium. In addition, a distinction will be made for patterns that indicate higher frequency in one of the two outcome classes, producing in this way highly discriminative patterns.

The primary source of data will be the OMOP CDM format of the Dutch Integrative Primary Care Information (IPCI) database. We will use information for conditions, drug prescriptions and procedures, and further consider the use of continuous information such as measurements and laboratory tests. We aim to make use

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s): Peter Rijnbeek et al.</b>	<b>Security: PU</b>	12/27



of higher levels of abstraction from the OMOP CDM such as the condition and drug eras and explore the feasibility of the other generalisations, such as single phenotype use from the OHDSI phenotype library.


## 2.4 Predictive Analytics Using Unstructured Data

The Text Represented In Terms Of Numeric-features (TRITON) package discussed in D3.4, was updated with several new features. We developed TRITON as a standardized Natural Language Processing (NLP) pipeline tool, within the OHDSI framework, for extracting textual features in a data-driven and language-independent manner. This tool extends the *FeatureExtraction* framework in the form of a custom covariate builder and constructs a set of text-based covariates. TRITON is publicly available on GitHub at [github.com/mi-erasmusmc/Triton](https://github.com/mi-erasmusmc/Triton).

Additional to the ability to create sparse bag-of-word text representations, it is now also possible to create dense text representations such as topic models and word or document embeddings. Furthermore, TRITON now has the ability of creating features based on information in the CDM `note_nlp` table. The `note_nlp` table is populated with clinical concepts extracted from the text and can include contextual information (e.g., negation, experiencer, temporal aspects, severity). In the upcoming period we will further develop the NLP pipeline and apply it to multiple use cases.

The aim of the EH DEN use cases is to determine the added value of textual data in EHRs for improving PLP models. This is currently done in two ways. First, we are performing a systematic review on the use of unstructured text data in clinical prediction models. We will summarize the prediction problems and methodological landscape and determine whether the information extracted from text data has value supplementing that of structured data. This systematic review is in its last phase and nearly ready for publication.

Second, we will investigate a variety of methods to generate text-based features and determine whether predictive performance improves if these features are used on their own and in addition to the features based on coded information. Different feature sets will be generated using the TRITON framework and tested for their predictive value. Apart from the bag-of-words, topic model, and embedding text representations that will be considered, we will also perform clinical concept extraction by detecting SNOMED CT and UMLS concepts and their contextual information in the text. The publicly available MedSpacy python library [19] provides a modular pipeline for concept extraction applicable to multiple languages.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s): Peter Rijnbeek et al.</b>	<b>Security: PU</b>	13/27



## 2.5 Deep Learning – Attention-Based Models

The OHDSI community has developed a framework for developing and validating predictions models on observational data standardized to the OMOP CDM [10]. Although the framework supports deep learning models, the focus of research using the PLP framework has mostly been on traditional machine learning models. Recently, there have been rapid advances in the deep learning field which might work well on the type of observational data that are in the OMOP CDM.

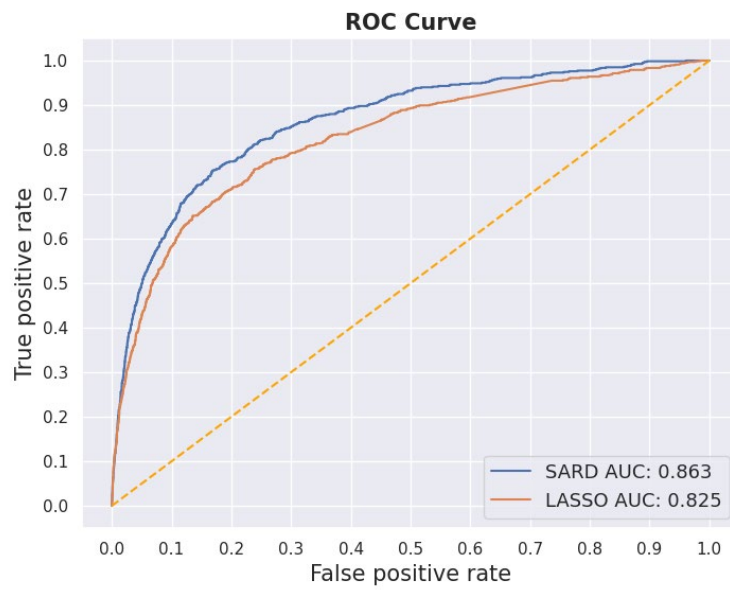
A relevant advance in the field of deep learning is that of attention-based models [20]. Attention is a mechanism where the relations between the input features of a sequence are learnt. These relations are then used to build representations which are used for the task of interest. These types of models have improved the state-of-the art in diverse fields such as natural language processing and computer vision. Recently there has been work done in translating these models to EHR data [21, 22]. While it has been hard to improve upon a strong linear baseline on EHR data, one approach in particular shows promising results where a model learns, using reverse distillation, from a strong linear baseline and then subsequently outperforms it.

We implemented a recent attention-based model and integrated it into the PLP framework. The model is the Self-Attention with Reverse Distillation (SARD) model [22] and we used a windowed L1 regularized linear regression model (LASSO) as baseline. We predicted mortality in the following year after a first occurrence of a general practitioner's visit after the patient reached the age of 60. We used condition, drug and procedure codes as features from 3 years before the index date. The data used was the IPCI database from the Netherlands. We split the data into 60% training set, 20% validation set and 20% test set with a stratified split. Loss on validation set was used for early stopping and to select the best hyperparameters. Results are reported for the test set. For LASSO a full grid search was used to select hyperparameters. C is the parameter controlling the sparsity, we searched over a grid of C from 0.0001 to 10,000 spaced evenly on a log scale with 20 values. For SARD a randomized search of hyperparameters was used with 100 samples, then with the resulting model a full grid search was done over the values of  $\alpha = [0, 0.05, 0.1, 0.15, 0.2]$ . Alpha is the parameter that controls the influence of distillation loss vs cross entropy loss during fine-tuning of the SARD model.

For our task, the population at risk includes 150,277 patients with 3329 outcomes. The hyperparameters selected for the linear model are a C of 0.01 and windows of [30, 90, 180, 365 and 1095] days before index. For SARD the selected hyperparameters were 2 attention heads with embedding dimension of 32 per head. The number of attention layers was 6 with no dropout and alpha was 0.15. For SARD all concepts in visits over 365 days before index were considered to be part of the same visit.

The linear baseline performs well with an area under the curve (AUC) of 82.5% while SARD reached 86.3% AUC (Figure 4). These results are in line with those in [22].

Attention-based deep learning models are promising, although a simple linear baseline is still competitive. When fully integrated into the PLP framework these kinds of models can be used in a fast and straightforward way on various OMOP CDM databases. Future work will explore other classes of attention-based models, look at their performance with external validation as well as explore the interpretability of the attention weights.



*Figure 4: Receiver-operating curves of the two models*



## 2.6 Iterative Pairwise External Validation

External validation of prediction models is increasingly being seen as a minimum requirement for acceptance in clinical practice [23-25]. The lack of interoperability of healthcare databases, however, has been the biggest barrier to this occurring at a large scale. Recent improvements in database interoperability enable a standardized analytical framework for model development and external validation [26, 27]. External validation of a model in a new database lacks context, whereby the external validation can be compared to a benchmark in this database [28-30]. Iterative pairwise external validation (IPEV) is a framework which uses a rotating model development and validation approach to contextualize the assessment of performance across a network of databases. As a use case we developed and validated models to predict 1-year risk of heart failure in patients initializing a second pharmacological intervention for type 2 diabetes.

The method follows a 2-step process involving 1) development of baseline and data-driven models in each database according to best practices; 2) validation of these models across the remaining databases. We introduce a heatmap visualization that supports the assessment of the internal and external model performance in all available databases. We leveraged the power of the OMOP CDM to create an open-source software package to increase the consistency, speed and transparency of this process.

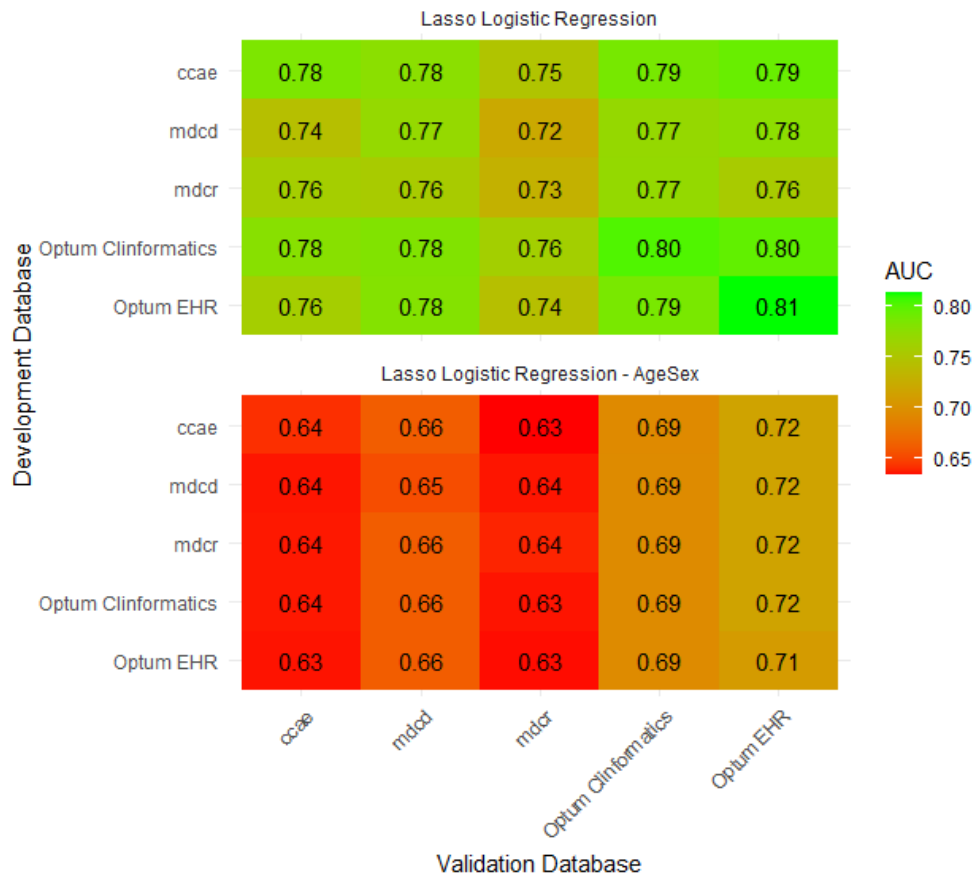



Figure 5. A heatmap of the AUC values across internal validation (values on the lead diagonal) and external validations of the developed prediction models. The colour scale runs from red (low discriminative ability) to green (high discriminative ability). The upper section details the performances for the data-driven model. The lower half details the same but then for the Age and Sex model. Abbreviations: ccae: Commercial Claims and Encounters, mdcd: Medicaid, mdcr: Medicare, optum EHR: optum electronic health records.

A total of 403,187 patients were included in the study from 5 databases. We developed 5 models which when assessed internally had a discriminative performance ranging from 0.73 to 0.81 AUC with acceptable




	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s):</b> Peter Rijnbeek et al.	<b>Security:</b> PU	16/27



calibration. When externally validating these models in a new database, three models achieved consistent performance and in context often performed similarly to models developed in the database itself. The visualization, in Figure 5, of IPEV provided valuable insights and allows for the direct comparison between the models across the databases and across the models within a database. From this the model developed in the CCAE (Commercial Claims and Encounters) database is identified as the best performing model overall.

Using IPEV lends weight to the model development process. The rotation of development through multiple databases provides context to model assessment leading to improved understanding of transportability and generalizability. The inclusion of a baseline model in all modelling steps provides further context to the performance gains of increasing model complexity. The CCAE model was identified as a candidate for clinical use. The use case demonstrates that IPEV provides a huge opportunity in a new era of standardized data and analytics to improve insights and trust in prediction models at an unprecedented scale. This paper is currently under submission for publication.



	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s): Peter Rijnbeek et al.</b>	<b>Security: PU</b>	<b>17/27</b>



## 2.7 Continuous Risk-Based Assessment of Treatment Effect Heterogeneity

Predictive approaches to treatment effect heterogeneity aim at the development of models predicting either individualized effects or which of two (or more) treatments is better for an individual. Earlier work presented here focused on the implementation of a risk-based approach to the assessment of treatment effect heterogeneity in the observational setting. This approach focuses on evaluating treatment effects within risk subgroups and has been shown to increase power for the detection of treatment effect heterogeneity. However, it does not provide individualized benefit predictions which are very important in the substrate of personalized medicine. Our aim in this case was to summarize and compare different easy-to-implement risk-based methods for deriving patient-level predictions of absolute benefit of a specific treatment relative to a comparator in a simulation study. We started from the randomised control trial setting, where the implementation of these methods is more straightforward. We intend, however, to extend our methods to the observational setting in the future.

For the simulations we simulated data using diverse assumptions for a baseline prognostic index of risk and the shape of its interaction with treatment (none, linear or quadratic). In each sample we predicted absolute benefit using: models with the Prognostic Index (PI) and a constant relative treatment effect; models including a interaction of treatment with the PI; stratification in quarters of the PI; nonlinear transformations of the PI (restricted cubic splines with 3, 4 and 5 knots); an adaptive approach using Akaike's Information Criterion. We evaluated predictive performance using root mean squared error and measures of discrimination and calibration for benefit. Starting from a base case scenario (sample size 4,250, treatment odds ratio 0.8, AUC of the PI 0.75), we varied the sample size, the treatment effect strength, the PI's discriminative ability, and the size of constant treatment-related harms on the absolute scale.

Our analyses simulated data in 648 scenarios. Models including a PI by treatment interaction performed best with smaller sample sizes and/or lower AUC of the PI. Otherwise, Restricted Cubic Splines (RCS) (3 knots) models proved more robust for the majority of the deviation settings. The adaptive approach was unstable



with smaller sample sizes. Therefore, we conclude that, depending on the setting, the linear interaction or the RCS (3 knots) model should be preferred.

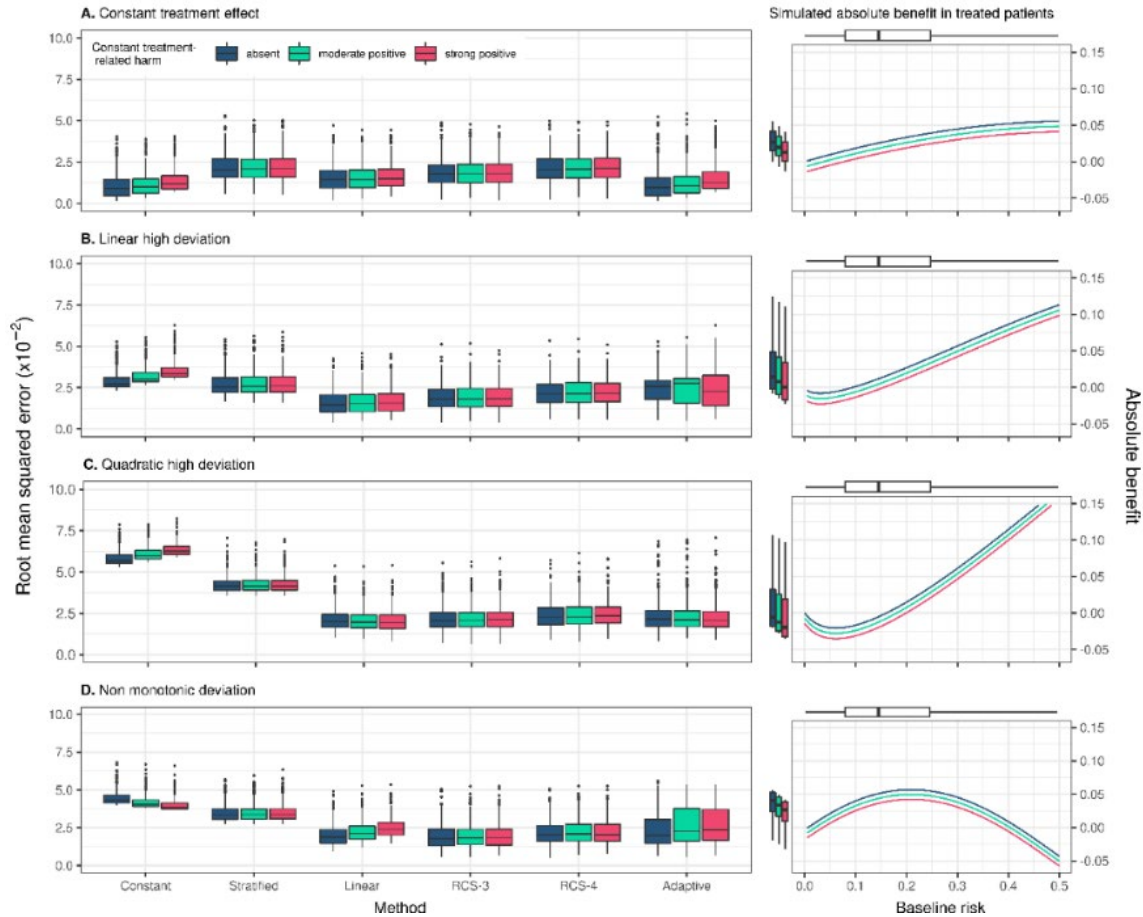


Figure 6: On the right-hand side we present the root mean squared error of the different methods in scenarios based on the base case ( $OR=0.8$ , true prediction  $AUC=0.75$ ,  $N=4250$ ) and introduce different sizes and shapes of deviations (linear, quadratic and non-monotonic) and moderate or strong treatment-related harms applied to the treated patients. On the left-hand side we present the true evolution of absolute benefits in each scenario



## 2.8 Disease Trajectories

Identifying temporal disease sequences (trajectories) where one event leads to another provides a great interest for researchers around the globe. This allows describing progressions and treatment patterns within the dataset or population and predicting future illnesses from the existing ones. Therefore, the number of trajectory studies has been slowly increasing over time. However, the full potential of these kinds of studies has not been revealed, and we believe it is primarily because of two reasons. First, the lack of syntactic and semantic interoperability of health data makes network studies a challenge. Second, there has not been a standardized open-source implementation of an analytical framework for performing this type of analysis.

Using the OMOP CDM is an effective method for tackling the first issue. For the second issue, we have developed a four-step framework based on significant temporal event pair detection and implemented it as an open-source R package. The proposed framework for detecting temporal health event trajectories consists of the following steps: 1. Define a study cohort; 2. Specify study parameters; 3. Identify temporal clinical event pairs; and 4. Count trajectories that consist of temporal clinical event pairs.

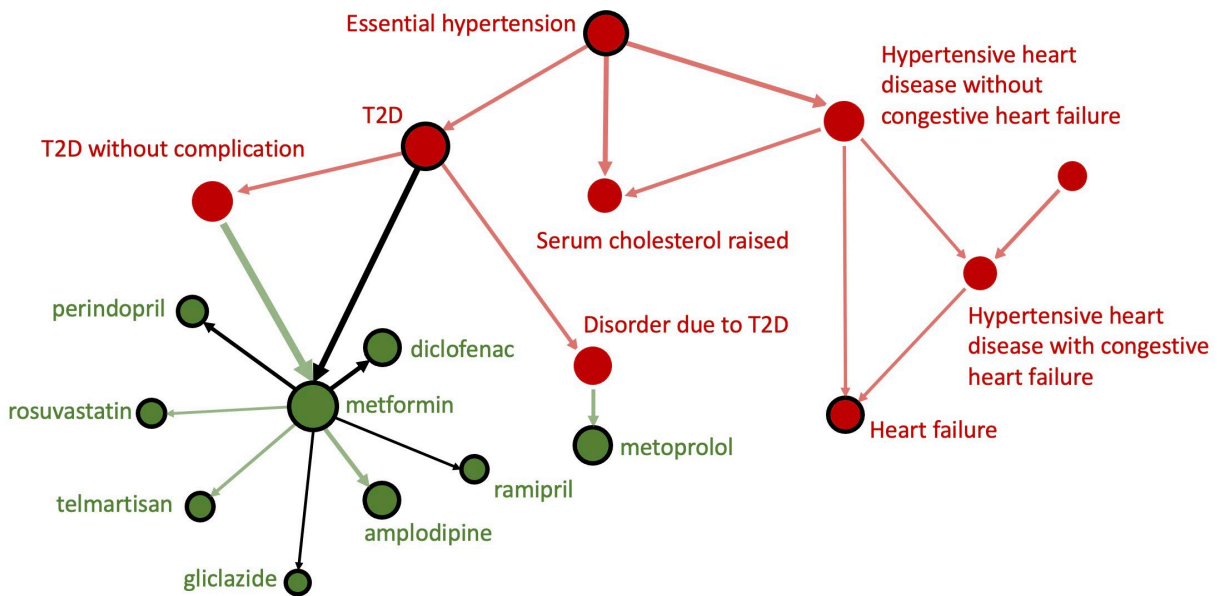



Figure 7. The 20 most prevalent event sequences among Type 2 diabetes mellitus patients having a relative risk >2 in Estonian electronic health records. Five event pairs that passed validation in the IPCI database (Netherlands) are shown with black arrows.


For the first time, there is a complete software package for detecting disease trajectories in health data. We used it on a population-based Estonian dataset to replicate a large Danish population-based study as proof of concept. Out of 40,711 temporal event pairs observed in Denmark and 22,618 in Estonia, the overlap was small (2290 pairs). We also conducted a disease trajectory detection study for Type 2 Diabetes patients in the Estonian and Dutch databases. Out of 943 pairs identified in the first dataset, 117 of them were confirmed in the other. We analysed the causes of the differences and have described them as challenges of these kinds of studies. We have demonstrated the framework on a poster in OHDSI Symposium 2021 (available at <https://www.ohdsi.org/2021-global-symposium-showcase-71/>, figure below), prepared a scientific manuscript “Trajectories: a framework for detecting temporal clinical event sequences from health data

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s):</b> Peter Rijnbeek et al.	<b>Security:</b> PU	20/27



standardised to the OMOP Common Data Model”, and submitted it to a peer-reviewed journal. The manuscript is also available in pre-print:

<https://www.medrxiv.org/content/10.1101/2021.11.18.21266518v1>

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s):</b> Peter Rijnbeek et al.	<b>Security:</b> PU	21/27



## 2.9 Systematic Review on Clinical Prediction Modelling

Before implementing a prediction model in clinical practice, it is important to ensure that its prediction performance is generalizable and robust by externally validating the model across various databases. This systematic review aims to provide further insights into the conduct and reporting of clinical prediction model development and validation over time. We focus on assessing the reporting of information necessary to enable external validation by other investigators. In particular, the prediction problem definition needs to be clearly reported and the final model needs to be completely presented.

We searched Embase, Medline, Web-of-Science, Cochrane Library and Google Scholar to identify studies that developed one or more multivariable prognostic prediction models using EHR data published in the period 2009-2019. The search was performed on November 15, 2019. We extracted data on data origin, data characteristics, data handling, modelling method, prediction problem definition, final model presentation, and model validation. To investigate the trends in the period 2009-2019, we assessed the extracted data for the periods 2009-2014 and 2015-2019 separately.

Our initial search resulted in a total of 9,942 papers. After duplicates were removed, 6,235 titles and abstracts were screened. From this, 1,075 potentially eligible papers were identified. Upon full-text inspection, 422 studies were eventually included for data extraction. In total, we extracted data for 579 clinical prediction models (with 1-6 models per study). We observed a steep increase over the years in the number of developed models, with 135 models in 101 studies in the period 2009-2014 and 444 models in 321 studies in the period 2015-2019. The percentage of models externally validated in the same paper remained stable at around 10%. Throughout 2009-2019, for both the target population and the outcome definitions, code lists were provided for less than 20% of the models. For about half of the models that were developed using regression analysis, the final model was not completely presented.

Overall, we observed limited improvement over time in the conduct and reporting of clinical prediction model development and validation. We found that the prediction problem definition was often not clearly reported, and the final model was often not completely presented, with little to no improvement over time. Thus, improvement in the reporting of information necessary to enable external validation by other investigators is still urgently needed to increase clinical adoption of developed models.

Link to preprint: <https://doi.org/10.1101/2021.10.22.21265374>.



## 2.10 Prediction Model Library

The OHDSI community is busy developing prediction models, but these have yet to make an impact on clinical practice. A reason for this is that the dissemination and understanding of prediction modelling in healthcare can still be improved. The current practice leaves the different models developed by different researchers disconnected from each other. A centralized repository, or library, will collate this information together and provide improved access and usability for a range of users including regulators, clinicians and prediction researchers.

We created a relational database to store results from multiple OHDSI PLP studies. This database models the structure of the results objects generated from model development or validation. An entity relationship diagram is available in figure 8. This database can be accessed through a dedicated application, which allows for the exploration of the results of multiple studies. It will also provide the ability to select models to download, this will then create a package including the required settings and cohort definitions as well as the model.

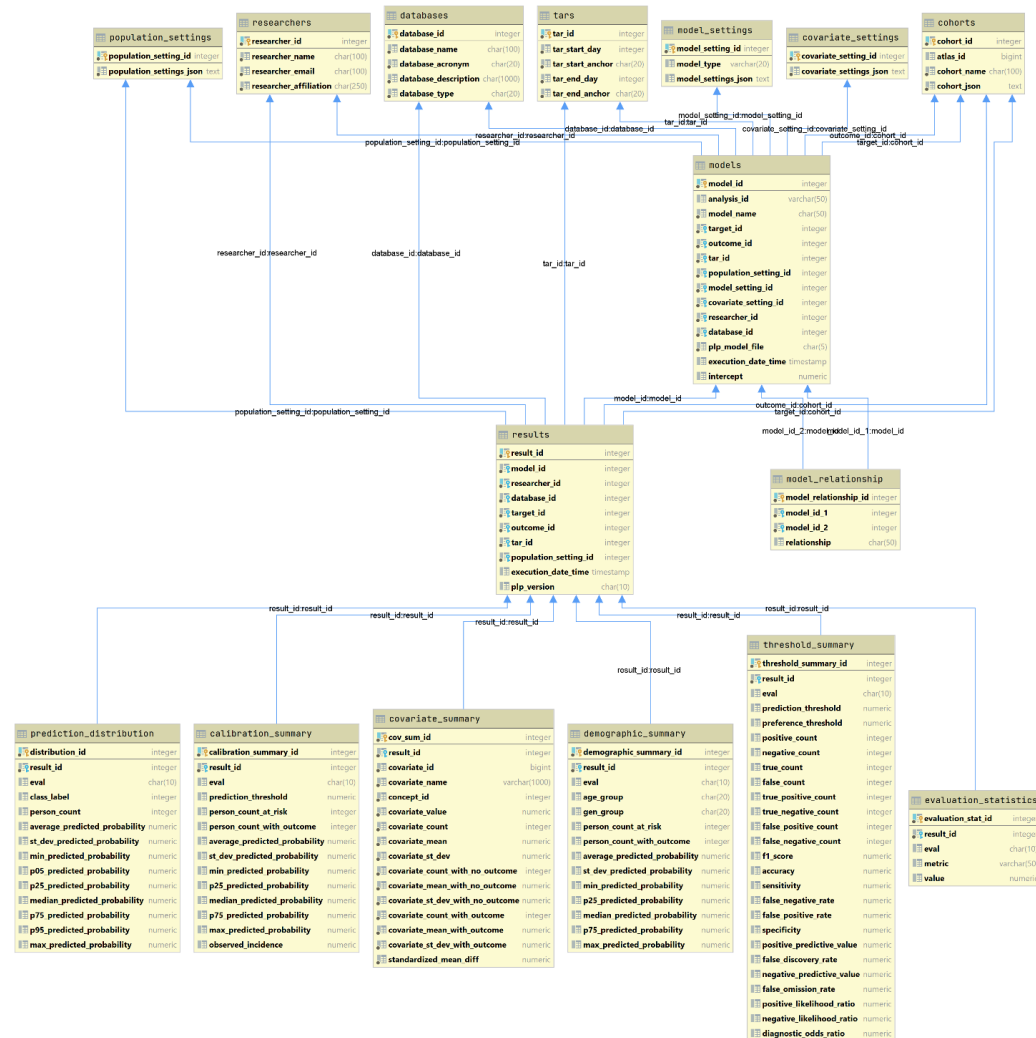


Figure 8. An entity relationship diagram for the database of the prediction model library.



The app we developed contains information on more than 600 models and 20,000 different validations. An example of the app can be seen in Figure 9. On the landing page, the development models are available with various important information displayed including the setting (target, outcome cohorts and time at risk), the model performance, and information about the developer and the database used for development. Once a model has been selected, more detailed information is then available on the discrimination performance. This is available for the model overall as well as at different thresholds. These thresholds can be explored interactively. For example a user can set a threshold and then see the sensitivity and specificity at this threshold. Further, calibration performance is available graphically as well as using various metrics (E50, E90 calibration-in-the-large). Plots to explore the calibration in different demographics are also available. Finally, a validation tab provides the ability to compare a development setting evaluation with an external validation of models. This comparison can be done both quantitatively using various performance metrics as well as qualitatively by comparing calibration and AUC plots.

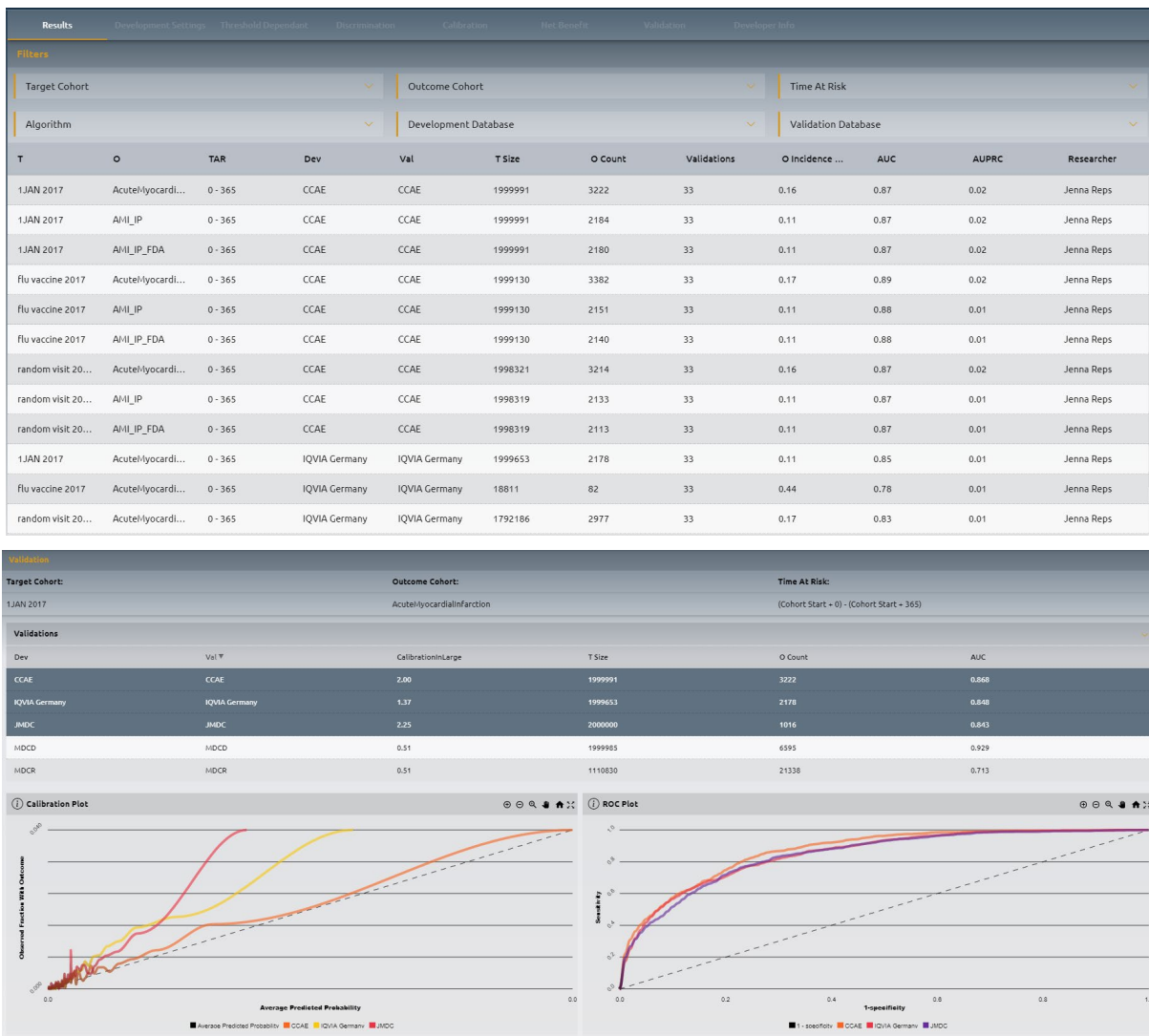



Figure 9. Screenshot of the library page and model validation risk exploration.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>		<b>Version:</b> v1.1 – Final
	<b>Author(s):</b> Peter Rijnbeek et al.		<b>Security:</b> PU




The library provides a unique environment to interactively explore the results and evidence for prediction models developed within the OHDSI PLP framework. This provides an improved user experience and allows for greater exploration of results by multiple stakeholder groups.

For a demo of a previous version of the tool see:

[https://www.youtube.com/watch?v=hi2Zs1Bfj54&ab\\_channel=OHDSI](https://www.youtube.com/watch?v=hi2Zs1Bfj54&ab_channel=OHDSI)

A publication is currently being written.



	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s):</b> Peter Rijnbeek et al.	<b>Security:</b> PU	25/27




### 3. NEXT STEPS

In the third year, the methods research and tool developed has progressed considerably and multiple manuscripts have been created. The analytical pipelines have also been used in use cases (see D3.7).

The next steps for WP3 are:

1. A standing item is to look into federated learning. Initial explorations and interactions with others in the OHDSI community have taken place but an implementation plan still needs to be made.
2. The research on the use of multilingual unstructured text in the context of prediction will continue with use cases.
3. Research on frequent-pattern mining can now be performed using the developed R package.
4. Further use cases for the disease trajectory pipeline will be explored.
5. We will look into the problem of class imbalance when building predictive models.
6. Development of educational material for the use of the developed pipelines will get a priority.


Finally, the EHDEN data network has grown considerably, and we can invite these European data partners to participate in the upcoming use cases to further improve the analytics.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>		<b>Version:</b> v1.1 – Final
	<b>Author(s):</b> Peter Rijnbeek et al.		<b>Security:</b> PU



## REFERENCES

1. He, J., et al., *The practical implementation of artificial intelligence technologies in medicine*. Nature medicine, 2019. **25**(1): p. 30-36.
2. Rijnbeek, P.R. and J.A. Kors, *Finding a short and accurate decision rule in disjunctive normal form by exhaustive search*. Machine learning, 2010. **80**(1): p. 33-62.
3. Markus, A.F., J.A. Kors, and P.R. Rijnbeek, *The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies*. Journal of Biomedical Informatics, 2021. **113**: p. 103655.
4. Fournier-Viger, P., et al., *A survey of itemset mining*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017. **7**(4): p. e1207.
5. Fournier-Viger, P., et al., *A survey of sequential pattern mining*. Data Science and Pattern Recognition, 2017. **1**(1): p. 54-77.
6. Agrawal, R. and R. Srikant. *Fast algorithms for mining association rules*. in *Proc. 20th int. conf. very large data bases, VLDB*. 1994. Citeseer.
7. Zaki, M.J., *Scalable algorithms for association mining*. IEEE transactions on knowledge and data engineering, 2000. **12**(3): p. 372-390.
8. Han, J., et al., *Mining frequent patterns without candidate generation: A frequent-pattern tree approach*. Data mining and knowledge discovery, 2004. **8**(1): p. 53-87.
9. Zaki, M.J., *SPADE: An efficient algorithm for mining frequent sequences*. Machine learning, 2001. **42**(1): p. 31-60.
10. Reps, J.M., et al., *Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data*. Journal of the American Medical Informatics Association, 2018. **25**(8): p. 969-975.
11. Timmons, A.C. and K.J. Preacher, *The importance of temporal design: How do measurement intervals affect the accuracy and efficiency of parameter estimates in longitudinal research?* Multivariate behavioral research, 2015. **50**(1): p. 41-55.
12. Campbell, E.A., E.J. Bass, and A.J. Masino, *Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma*. Journal of the American Medical Informatics Association, 2020. **27**(4): p. 558-566.
13. Zhao, J., et al., *Learning from heterogeneous temporal data in electronic health records*. Journal of biomedical informatics, 2017. **65**: p. 105-119.
14. Batal, I., et al., *An efficient pattern mining approach for event detection in multivariate temporal data*. Knowledge and information systems, 2016. **46**(1): p. 115-150.
15. Moskovitch, R., et al., *Procedure prediction from symbolic Electronic Health Records via time intervals analytics*. Journal of biomedical informatics, 2017. **75**: p. 70-82.
16. Moskovitch, R., et al., *Prognosis of clinical outcomes with temporal patterns and experiences with one class feature selection*. IEEE/ACM transactions on computational biology and bioinformatics, 2016. **14**(3): p. 555-563.
17. Li, Y., et al., *Graph Neural Network-Based Diagnosis Prediction*. Big Data, 2020. **8**(5): p. 379-390.
18. Maragatham, G. and S. Devi, *LSTM model for prediction of heart failure in big data*. Journal of medical systems, 2019. **43**(5): p. 1-13.

	<b>D3.6 – Third Report on the implementation of the analytical pipeline for personalized medicine</b>		
	<b>WP3 – Personalized Medicine</b>	<b>Version: v1.1 – Final</b>	
	<b>Author(s): Peter Rijnbeek et al.</b>	<b>Security: PU</b>	<b>27/27</b>



19. Eyre, H., et al., *Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python*. arXiv preprint arXiv:2106.07799, 2021.
20. Vaswani, A., et al. *Attention is all you need*. in *Advances in neural information processing systems*. 2017.
21. Choi, E., et al., *Retain: An interpretable predictive model for healthcare using reverse time attention mechanism*. arXiv preprint arXiv:1608.05745, 2016.
22. Kodialam, R.S., et al., *Deep Contextual Clinical Prediction with Reverse Distillation*. arXiv preprint arXiv:2007.05611, 2020.
23. Steyerberg, E.W. and F.E. Harrell Jr, *Prediction models need appropriate internal, internal-external, and external validation*. *Journal of clinical epidemiology*, 2016. **69**: p. 245.
24. Collins, G.S., et al., *External validation of multivariable prediction models: a systematic review of methodological conduct and reporting*. *BMC medical research methodology*, 2014. **14**(1): p. 1-11.
25. Reps, J.M., et al., *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*. *BMC medical research methodology*, 2020. **20**(1): p. 1-10.
26. Lehne, M., et al., *Why digital medicine depends on interoperability*. *NPJ digital medicine*, 2019. **2**(1): p. 1-5.
27. Kent, S., et al., *Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment*. *PharmacoEconomics*, 2021. **39**(3): p. 275-285.
28. Moons, K.G., et al., *Risk prediction models: II. External validation, model updating, and impact assessment*. *Heart*, 2012. **98**(9): p. 691-698.
29. Riley, R.D., et al., *External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges*. *bmj*, 2016. **353**.
30. Ramspek, C.L., et al., *External validation of prognostic models: what, why, how, when and where?* *Clinical Kidney Journal*, 2021. **14**(1): p. 49-58.