EUROPEAN HEALTH DATA & EVIDENCE NETWORK

806968 – EHDEN

European Health Data & Evidence Network

WP3 – Personalized Medicine

D3.4 Second 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 (1 – EMC) Daniel Prieto Alhambra (3 – UOXF) Sulev Reisberg (UTARTU)
Due date	31/10/2020
Delivery date	29/11/2020
Deliverable type	R
Dissemination level	PU
DoA - Version	V1
Date	12/11/2018

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

TABLE OF CONTENTS

Document History

Version	Date	Description
V1.0	15 November 2020	Final Draft for internal review
V1.1	27 November 2020	Final Version

DEFINITIONS

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

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

Grant agreement The agreement signed between the beneficiaries and the IMI JU for the undertaking of the EHDEN project (806968).

Project The sum of all activities carried out in the framework of the Grant Agreement.

Consortium The EHDEN Consortium, comprising the above-mentioned legal entities.

Consortium agreement 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.

Publishable Summary

There are large opportunities to use the massive amount of observational data in Europe for personalized decision-making. However, the lack of inter-operability of the data sources makes this a challenging task. The differences in structure (syntactic inter-operability) and terminology systems (semantic inter-operability) make the development of standardized analytical pipelines cumbersome. The European Health Data and Evidence Network (EHDEN) project is addressing this by standardizing a large amount of European data sources to the OMOP Common Data Model (CDM). 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 on WP3 activities, 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.

The current second report provides an update on the work done in the second year. This includes an overview of use cases in which the analytical pipelines have been applied and describes the advances made in methodological research, the start of a natural language processing pipeline, and work done to develop a pipeline for disease trajectories.

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).

1. Introduction

As discussed in depth in D3.1 “First Report on the implementation of the analytical pipeline for personalized medicine”, 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 D3.1 we focused on PLP and PLE, and presented some preliminary work on heterogeneity of treatment effect. In this deliverable we start by describing some use cases in which the PLP and PLE methods have been applied successfully. In addition, we describe the advances made with HTE and disease trajectory analysis, and present methodological work done by the WP3 team.

2. Use Cases

We start by presenting multiple use cases in which prediction models have been developed and then give an example of population-level effect estimation. These use cases have been developed and executed in close collaboration with WP1 “Evidence Workflow Development”, and with the Observational Health Data Sciences and Informatics (OHDSI) global data network.

2.1 Patient-Level Prediction

We first give a quick refresher on the prediction problem as described in more depth in D3.2 and https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html.

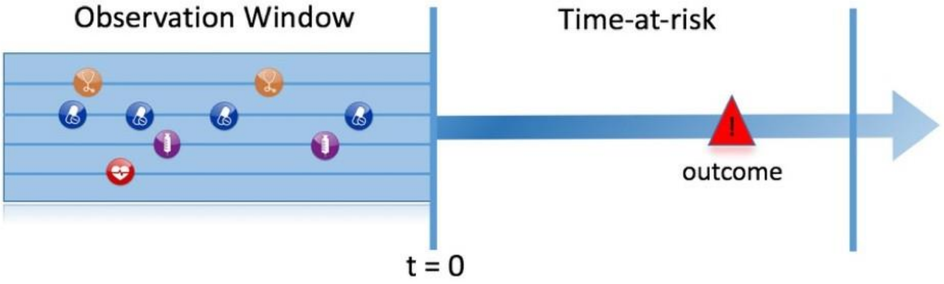


Figure 1: The prediction problem.

Figure 1 illustrates the prediction problem. 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.

As shown in Table 1, to define a prediction problem we have to define t=0 by a target cohort, the outcome we like to predict by an outcome cohort, and the time-at-risk. We define the standard prediction question as:

> Among *[target cohort definition]*, who will go on to have *[outcome cohort definition]* within *[time-at-risk period]*?

Furthermore, we have to make design choices for the model we like to develop and determine the observational datasets to perform internal and external validation.

Table 1: Main design choices in a prediction design

Choice	Description
Target cohort	How do we define the cohort of persons for whom we wish to predict?
Outcome cohort	How do we define the outcome we want to predict?
Time-at-risk	In which time window relative to t=0 do we want to make the prediction?
Model	What algorithms do we want to use, and which potential predictor variables do we include?

For more information see the R Package at https://github.com/PatientLevelPrediction, and the PLP framework paper [1].

2.1.1 COVID-19 Prediction Model (COVER)

Table 2. Problem definition COVER study

Choice	Definition
Target cohort	Patients with initial COVID-19 infection
Outcome cohort	Hospitalisation, Hospitalisation with intensive services, fatality
Time-at-risk	30 days
Model	LASSO logistic regression with custom covariates

Background

The majority of pandemic response planning has focused on population-level effects of likely disease spread and contain no information on how an individual's risk impacts their likely morbidity and mortality if they were to contract the virus. The WHO Risk Communication Guidance [2] distinguishes two categories of patients at high risk of severe disease: those older than 60 years and those with "underlying medical conditions" which is non-specific. Using general criteria to assess the risk of poor outcomes is a crude risk discrimination mechanism as entire patient groupings are treated homogeneously ignoring individual differences. Research has shown that COVID-19 does not impact all ages and sexes equally and as such a more personalised risk assessment can aid in improving outcomes. In a recent BMJ editorial [3], the authors conclude that the COVID-19 response "is about protecting lives and communities most obviously at risk in our unequal society". Quantifying a patient's risk of developing severe or critical illness when infected with COVID-19, could be used to help countries plan strategies to shield the most vulnerable patient populations. This is essential during the planning of de-confinement strategies. If a more personalised risk assessment were possible, then governments and healthcare authorities could advise based upon this and create a subtler lockdown procedure that reduces the personal and economic impact whilst continuing to keep healthcare utilisation at manageable levels.

In this research we aimed to develop COVID-19 Estimated Risk (COVER) scores to quantify a patient's risk of hospital admission (COVER-H), requiring intensive services (COVER-I), or fatality (COVER-F) due to COVID-19 using the OHDSI PLP framework. In order to allow for rapid development and to overcome the shortcoming of using small data for development, we made use of the abundant data from patients with influenza or flu-like symptoms to develop the models and then we tested whether the models transport to COVID-19 patients. Given the symptomatic similarities between the two diseases we hypothesized that the developed models will be able to transport between the two problem settings.

Methods

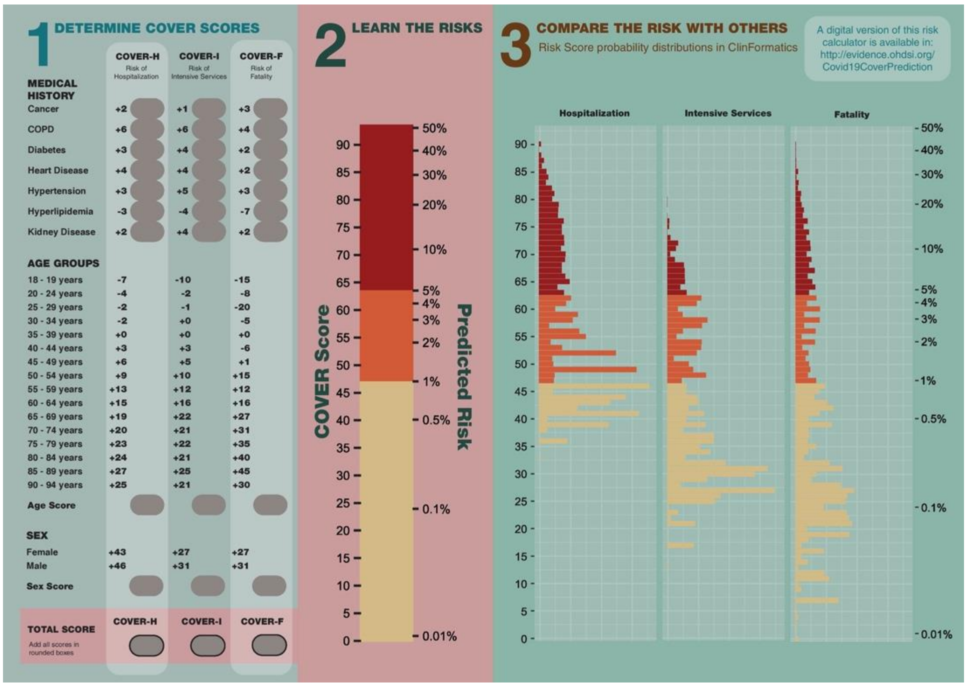
In this study, we analyzed a federated network of electronic medical records and administrative claims data from 14 data sources and 6 countries. We developed and validated 3 scores using 6,869,127 patients with a general practice, emergency room, or outpatient visit with diagnosed influenza or flu-like symptoms any time prior to 2020. The scores were validated on patients with confirmed or suspected COVID-19 diagnosis across five databases from South Korea, Spain and the United State. Outcomes included i) hospitalization with pneumonia, ii) hospitalization with pneumonia requiring intensive services or death, and iii) death in the 30 days after index date.

Results

Overall, 44,507 COVID-19 patients were included for model validation. We identified 7 predictors (history of cancer, chronic obstructive pulmonary disease, diabetes, heart disease, hypertension, hyperlipidemia, kidney disease) which combined with age and sex discriminated which patients would experience any of our three outcomes. The models achieved high performance in influenza. When transported to COVID-19 cohorts, the AUC ranges were: COVER-H: 0.69-0.81, COVER-I: 0.73-0.91, and COVER-F: 0.72-0.90. Calibration was overall acceptable.

Conclusions

A 9-predictor model performs well for COVID-19 patients for predicting hospitalization, intensive services and fatality. The models could aid in providing reassurance for low-risk patients and shield high risk patients from COVID-19 during de-confinement to reduce the virus' impact on morbidity and mortality (Figure 2).



A digital version of this risk calculator is available in: http://evidence.ohdsi.org/Covid19CoverPrediction

	COVER-H Risk of Hospitalization	COVER-I Risk of Intensive Services	COVER-F Risk of Fatality
MEDICAL HISTORY			
Cancer	+2	+1	+3
COPD	+6	+6	+4
Diabetes	+3	+4	+2
Heart Disease	+4	+4	+2
Hypertension	+3	+5	+3
Hyperlipidemia	-3	-4	-7
Kidney Disease	+2	+4	+2
AGE GROUPS			
18 - 19 years	-7	-10	-15
20 - 24 years	-4	-2	-8
25 - 29 years	-2	-1	-20
30 - 34 years	-2	+0	-5
35 - 39 years	+0	+0	+0
40 - 44 years	+3	+3	-6
45 - 49 years	+6	+5	+1
50 - 54 years	+9	+10	+15
55 - 59 years	+13	+12	+12
60 - 64 years	+15	+16	+16
65 - 69 years	+19	+22	+27
70 - 74 years	+20	+21	+31
75 - 79 years	+23	+22	+35
80 - 84 years	+24	+21	+40
85 - 89 years	+27	+25	+45
90 - 94 years	+25	+21	+30
Age Score			
SEX			
Female	+43	+27	+27
Male	+46	+31	+31
Sex Score			
TOTAL SCORE Add all scores in rounded boxes	COVER-H	COVER-I	COVER-F

Figure 2: COVER scoring system

Published Outputs:

Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network [4].

Results explorer (shiny app): http://evidence.ohdsi.org:3838/Covid19CoverPrediction/

2.1.2 Prediction Modelling in Rheumatoid Arthritis Patients

Table 3. Problem definition rheumatoid arthritis prediction study

Choice	Definition
Target cohort	Rheumatoid arthritis patients initiating first-line treatment of methotrexate monotherapy
Outcome cohort	Serious infections, myocardial infarction (MI) and stroke
Time-at-risk	90 days (serious infections), 2 years (MI and stroke), 5 years (cancer)
Model	L1 regularized logistic regression

Background

Rheumatoid arthritis (RA) is a common musculoskeletal disease, affecting approximately 0.5-1.0% of the adult population worldwide [5, 6]. While the management of RA has improved in recent decades, the risk of adverse health outcomes in addition to the traditional clinical manifestations in RA patients remains a major issue [7, 8]. Adverse health outcomes in RA include known complications such as cytopenia and comorbidities such as cardiovascular disease (CVD), infection, and cancer, which have a higher prevalence in RA patients compared to the general population [9-11]. Generally, periodic screening and monitoring for these adverse health outcomes throughout the course of therapy would allow for early interventions. Unfortunately, with a wide range of possible adverse health outcomes, this is a challenging task. Managing treatment of RA is complex and there is limited time available for direct interaction with patients. Evaluating patient-level risks for adverse health outcomes upon initiation of treatment would therefore allow clinicians to provide more personalized care. With methotrexate (MTX) adopted as the “anchor drug” since the 1990s [12], we aimed to develop and validate patient-level prediction models for risk of leukopenia, pancytopenia, infection (serious, opportunistic, all), cardiovascular disease (myocardial infarction (MI), stroke), and cancer (colorectal, breast, uterine) in RA patients initiating first-line treatment of MTX monotherapy.

Methods

Patient data were obtained from 15 claims and electronic health record (EHR) databases mapped to the OMOP CDM across 9 countries (Australia, Estonia, France, Germany, Japan, Netherlands, Spain, United Kingdom, and United States of America). All RA patients initiating first-line treatment of MTX monotherapy with at least one year of prior observation were included. Prediction models were developed on the Optum Clinformatics Data Mart Database using L1 regularized logistic regression to predict the risk of adverse health outcomes in 3 months (leukopenia, pancytopenia, infection), 2 years (MI and stroke), and 5 years (cancer) after initiating treatment. For each outcome, this allowed us to develop the models on approximately 20,000 RA patients, with 75% and 25% of the data used for training and testing the models, respectively. More than 143,000 RA patients from the other 14 databases were included for external validation. Performance was assessed using the area under the receiver operator characteristic curve (AUROC) and calibration plots.

Results

To the best of our knowledge, our study is the first to develop and externally validate patient-level prediction models for risk of a variety of adverse health outcomes in RA patients initiating first-line treatment of MTX monotherapy. For leukopenia, breast cancer, and uterine cancer, there were too few outcome events in our data to develop patient-level prediction models. For risk of pancytopenia, opportunistic infections, all infections, and colorectal cancer, we did not consider the internal validation AUROC < 0.70 high enough to warrant validating the model externally. However, for risk of serious infections, MI, and stroke, we were able to develop models that showed good performance on internal and external validation. Internal validation resulted in AUROCs of 0.75, 0.77, and 0.78, respectively, indicating good discrimination. The large sample size of the development dataset allowed for good calibration as expected. External validation results showed good discrimination and calibration across other databases, although for some databases the AUROC confidence intervals were wide because of few outcome events (see Table 4). In databases where the outcome incidence is substantially higher or lower than in the development database, the models may benefit from recalibration. Overall, the models for risk of serious infections, MI, and stroke demonstrated transportability to RA patients from 14 other databases, with particularly good performance across USA databases.

Conclusions

We developed and externally validated patient-level prediction models for risk of serious infections, MI, and stroke in RA patients initiating first-line MTX monotherapy. The models showed good performance and demonstrated transportability to RA patients from 14 other databases. The models could be used to identify high-risk patients and aid clinicians in providing better personalized care.

Published Outputs:

A full manuscript detailing the developed prediction models is currently in development for submission to a peer-reviewed journal.

Results Explorer (Shiny App): https://data.ohdsi.org/ehdenRaPrediction/

Table 4. Internal and external validation results

Outcome	Database	Target population	Outcome events	AUROC (95% confidence interval)
Serious infections	Optum DOD (internal validation)	5,251	79 (1,5%)	0.747 (0.686-0.807)
	MDCD	3,355	111 (3.3%)	0.632 (0.580-0.684)
	JMDC	3,278	11 (0.3%)	0.707 (0.546-0.868)
	MDCR	6,533	152 (2.3%)	0.675 (0.631-0.719)
	CCAE	27,877	216 (0.8%)	0.661 (0.622-0.700)
	Estonia	1,464	8 (0.5%)	0.816 (0.691-0.941)
	IQVIA US hospital	3,703	746 (20.2%)	0.607 (0.585-0.630)
	Optum EHR	41,072	397 (1.0%)	0.738 (0.712-0.764)
Myocardial infarction	Optum DOD (internal validation)	5,308	98 (1.8%)	0.775 (0.735-0.815)
	MDCD	3,427	78 (2.3%)	0.717 (0.665-0.770)
	JMDC	3,299	9 (0.3%)	0.487 (0.298-0.676)
	MDCR	6,613	210 (3.2%)	0.684 (0.649-0.718)
	CCAE	28,084	173 (0.6%)	0.730 (0.693-0.767)
	Estonia	1,465	18 (1.2%)	0.673 (0.547-0.799)
	IPCI	556	7 (1.3%)	0.683 (0.559-0.807)
	IQVIA Australia	560	14 (2.5%)	0.576 (0.438-0.713)
	IQVIA LPD France	3258	7 (0.2%)	0.685 (0.510-0.860)
	IQVIA Germany	7401	38 (0.5%)	0.644 (0.560-0.728)
	IQVIA THIN	6,935	44 (0.6%)	0.621 (0.552-0.690)
	IQVIA US ambulatory	32,524	115 (0.4%)	0.756 (0.717-0.795)
	IQVIA US hospital	4,140	191 (4.6%)	0.667 (0.632-0.703)
	Optum EHR	41,496	716 (1.7%)	0.764 (0.745-0.781)
	SIDIAP	3,614	15 (0.4%)	0.648 (0.501-0.796)
Stroke	Optum DOD (internal validation)	5,301	127 (2.4%)	0.783 (0.745-0.821)
	MDCD	3,415	108 (3.2%)	0.785 (0.743-0.828)
	JMDC	3,299	21 (0.6%)	0.753 (0.640-0.866)
	MDCR	6,609	297 (4.5%)	0.685 (0.653-0.716)
	CCAE	28,082	243 (0.9%)	0.731 (0.698-0.764)
	Estonia	1,464	24 (1.6%)	0.774 (0.704-0.845)
	IQVIA Germany	7,416	37 (0.5%)	0.703 (0.603-0.802)
	IQVIA THIN	6,937	21 (0.3%)	0.648 (0.551-0.745)
	IQVIA US ambulatory	32,561	131 (0.4%)	0.722 (0.667-0.751)
	IQVIA US hospital	4,127	199 (4.8%)	0.632 (0.595-0.669)
	Optum EHR	41,404	868 (2.1%)	0.779 (0.765-0.794)
	SIDIAP	3,615	7 (0.2%)	0.749 (0.579-0.919)

2.2 Population-Level Effect Estimation

We start with a quick refresher of population-level effect estimation (see D3.2 for more details).

With population-level effect estimation, we refer to the estimation of average causal effects of exposures (e.g. medical interventions such as drug exposures or procedures) on specific health outcomes of interest. In what follows, we consider two different estimation tasks:

- Direct effect estimation: estimating the effect of an exposure on the risk of an outcome, as compared to no exposure.
- Comparative effect estimation: estimating the effect of an exposure (the target exposure) on the risk of an outcome, as compared to another exposure (the comparator exposure).

In both cases, the patient-level causal effect contrasts a factual outcome, i.e., what happened to the exposed patient, with a counterfactual outcome, i.e., what would have happened had the exposure not occurred (direct) or had a different exposure occurred (comparative). Since any one patient reveals only the factual outcome (the fundamental problem of causal inference), the various effect estimation designs employ different analytic devices to shed light on the counterfactual outcomes.

Use cases for population-level effect estimation include treatment selection, safety surveillance, and comparative effectiveness. Methods can test specific hypotheses one at a time (e.g. ‘signal evaluation’) or explore multiple-hypotheses-at-once (e.g. ‘signal detection’). In all cases, the objective remains the same: to produce a high-quality estimate of the causal effect.

We can specify the questions we wish to answer in a cohort study by making the five choices highlighted in Table 5.

Table 5. Main design choices in a comparative cohort design.

Choice	Description
Target cohort	A cohort representing the target treatment
Comparator cohort	A cohort representing the comparator treatment
Outcome cohort	A cohort representing the outcome of interest
Time-at-risk	At what time (often relative to the target and comparator cohort start and end dates) do we consider the risk of the outcome?
Model	The model used to estimate the effect while adjusting for differences between the target and comparator

For more information see the vignettes of the R Package at https://github.com/CohortMethod.

2.2.1 COVID-19 PLE studies

Back in March 2020 we noticed a striking uptake in the use of hydroxychloroquine, alone as well as in combination with azithromycin for the treatment of COVID-19 in Europe. Data from Spain suggested that >70% of patients hospitalized with COVID-19 were initiated on hydroxychloroquine, and about half of them were also given concomitant azithromycin. Both these drugs have known effects on cardiac repolarization, and drug-drug interactions had been reported previously. We therefore set out to investigate the cardiovascular safety of hydroxychloroquine in combination with azithromycin.

The resulting analyses were completed in a record time, with findings uploaded to MedRXiv on 10 April 2020 [13] and submitted to the European Medicines Agency (EMA) and other international regulators only 2 weeks after we initiated the study design. We demonstrated a doubled risk of short-term cardiovascular mortality when azithromycin is prescribed to patients previously taking hydroxychloroquine. This news was covered by Forbes, Science, and in regulatory warnings by the EMA that cited our preprint manuscript on 23 April

[14], therefore contributing to public health protection internationally. The details of this study are further described below, and in full in a recently published manuscript in Lancet Rheumatology [15].

At about the same time (22 April), the Spanish medicines regulator (AEMPS) published an additional warning suggesting a potential increase in the risk of neuropsychiatric symptoms amongst patients treated with hydroxychloroquine. This safety signal was further mentioned in a new warning by the EMA on 29 May 2020 [16]. In an unprecedented exercise, we leveraged the EHDEN and OHDSI COVID-19 data network and PLE analytical pipelines to analyse this signal. Our analyses were completed again within weeks, and uploaded to MedRXiv on 21 July [17]. More details on these analyses are reported below, but in brief, we did not find an increased risk of psychosis, depression or suicide/suicidal ideation amongst users of hydroxychloroquine.

Our novel methods have been praised in recent methodological guidelines published by the the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) [18].

Below two examples of studies are presented:

1. “Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study”. J E Lane et al. Lancet Rheumatol 2020. [15]

Shiny Application: https://data.ohdsi.org/Covid19EstimationHydroxychloroquine/

Background

Hydroxychloroquine, a drug commonly used in the treatment of rheumatoid arthritis, has received much negative publicity for adverse events associated with its authorisation for emergency use to treat patients with COVID-19 pneumonia. We studied the safety of hydroxychloroquine, alone and in combination with azithromycin, to determine the risk associated with its use in routine care in patients with rheumatoid arthritis.

Methods

In this multinational, retrospective study, new user cohort studies in patients with rheumatoid arthritis aged 18 years or older and initiating hydroxychloroquine were compared with those initiating sulfasalazine and followed up over 30 days, with 16 severe adverse events studied. Self-controlled case series were done to further establish safety in wider populations, and included all users of hydroxychloroquine regardless of rheumatoid arthritis status or indication. Separately, severe adverse events associated with hydroxychloroquine plus azithromycin (compared with hydroxychloroquine plus amoxicillin) were studied. Data comprised 14 sources of claims data or electronic medical records from Germany, Japan, the Netherlands, Spain, the UK, and the USA. Propensity score stratification and calibration using negative control outcomes were used to address confounding. Cox models were fitted to estimate calibrated hazard ratios (HRs) according to drug use. Estimates were pooled where the I2 value was less than 0·4.

Findings

The study included 956 374 users of hydroxychloroquine, 310 350 users of sulfasalazine, 323 122 users of hydroxychloroquine plus azithromycin, and 351 956 users of hydroxychloroquine plus amoxicillin. No excess risk of severe adverse events was identified when 30-day hydroxychloroquine and sulfasalazine use were compared. Self-controlled case series confirmed these findings. However, long-term use of hydroxychloroquine appeared to be associated with increased cardiovascular mortality (calibrated HR 1·65 [95% CI 1·12–2·44]). Addition of azithromycin appeared to be associated with an increased risk of 30-day cardiovascular mortality (calibrated HR 2·19 [95% CI 1·22–3·95]), chest pain or angina (1·15 [1·05–1·26]), and heart failure (1·22 [1·02–1·45]).

Interpretation

Hydroxychloroquine treatment appears to have no increased risk in the short term among patients with rheumatoid arthritis, but in the long term it appears to be associated with excess cardiovascular mortality. The addition of azithromycin increases the risk of heart failure and cardiovascular mortality even in the short term. We call for careful consideration of the benefit–risk trade-off when counselling those on hydroxychloroquine treatment.

Choice	Description
Target cohort	1.Hydroxychloroquine; 2.Hydroxychloroquine + Azythromycin
Comparator cohort	1.Sulfasalazine; 2.Hydroxychloroquine + Amoxicillin
Outcome cohort	Long list of serious adverse events, including cardiovascular mortality.
Time-at-risk	A) 30-day post therapy initiation B) As long as continuously on the same index treatment
Model	Propensity scorematching + Cox regression

2. “Risk of depression, suicidal ideation, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multi-national network cohort study”. J.C.E. Lane et al. MedRXiv. [17]

Shiny Application: https://data.ohdsi.org/Covid19EstimationHydroxychloroquine2/

Objectives

Concern has been raised in the rheumatological community regarding recent regulatory warnings that hydroxychloroquine used in the COVID-19 pandemic could cause acute psychiatric events. We aimed to study whether there is risk of incident depression, suicidal ideation, or psychosis associated with hydroxychloroquine as used for rheumatoid arthritis (RA).

Methods

This cohort study used claims and electronic medical records from 10 sources and 3 countries (Germany, UK and US). RA patients aged 18+ and initiating hydroxychloroquine were compared to those initiating sulfasalazine (active comparator) and followed up in the short (30-day) and long term (on treatment). Study outcomes included depression, suicide/suicidal ideation, and hospitalization for psychosis. Propensity score stratification and calibration using negative control outcomes were used to address confounding. Cox models were fitted to estimate database-specific calibrated hazard ratios (HRs), with estimates pooled where I2<40%.

Results

918,144 and 290,383 users of hydroxychloroquine and sulfasalazine, respectively, were included. No consistent risk of psychiatric events was observed with short-term hydroxychloroquine (compared to sulfasalazine) use, with meta-analytic HRs of 0.96 [0.79-1.16] for depression, 0.94 [0.49-1.77] for suicide/suicidal ideation, and 1.03 [0.66-1.60] for psychosis. No consistent long-term risk was seen, with meta-analytic HRs 0.94 [0.71-1.26] for depression, 0.77 [0.56-1.07] for suicide/suicidal ideation, and 0.99 [0.72-1.35] for psychosis.

Conclusions

Hydroxychloroquine as used to treat RA does not appear to increase the risk of depression, suicide/suicidal ideation, or psychosis compared to sulfasalazine. No effects were seen in the short or long term. Use at higher dose or for different indications needs further investigation.

Choice	Description
Target cohort	1.Hydroxychloroquine; 2.Hydroxychloroquine + Azythromycin
Comparator cohort	1.Sulfasalazine; 2.Hydroxychloroquine + Amoxicillin
Outcome cohort	Depression, suicide/suicidal ideation, hospitalization for psychosis.
Time-at-risk	A) 30-day post therapy initiation B) As long as continuously on the same index treatment
Model	Propensity score matching + Cox regression

3. Analytical Pipeline Development

In the next sections an update is provided for the analytical pipeline of heterogeneity of treatment effects, disease trajectories, and the work performed in natural language processing to enable the use of unstructured text.

3.1 Risk Stratified Effect Estimation

Analysis of heterogeneity of treatment effect (HTE), i.e. non-random variation in the direction or magnitude of a treatment effect for individuals within a population, is the cornerstone of precision medicine; its goal is to predict the optimal treatments at the individual level, accounting for an individual's risk for harm and benefit outcomes [19]. In D3.2, we described the work performed on the development of the HTE analytical pipeline. This work has progressed considerably and here we provide a brief update and present a pre-print publication that is currently under review at a high-impact journal.

The proposed framework defines five distinct steps that enable a standardized approach for risk-based assessment of treatment effect heterogeneity for databases mapped to the OMOP-CDM. These are: 1) general definition of the research aim; 2) identification of the database within which the analyses will be performed; 3) a prediction step where internal or external prediction models are used to assign patient-level risk predictions; 4) an estimation step where absolute and relative treatment effects are estimated within risk strata; 5) presentation and evaluation of the results. We created an R-package that can easily perform this kind of analyses, and made it publicly available (https://github.com/OHDSI/RiskStratifiedEstimation).

Step 1: Problem definition. The typical research aim is: "to compare the effect of treatment T to a comparator treatment C in patients with disease D with respect to outcomes $O_1, \dots, O_n$". At least three cohorts are defined: a single **treatment cohort** (T) which includes patients with disease D receiving the target treatment of interest; a single **comparator cohort** (C) which includes patients with disease D receiving the comparator (control) treatment; one or more **outcome cohorts** ($O_1, \dots, O_n$) that contain patients developing the outcomes of interest.

Step 2: Identification of the database. The aim of this step is the inclusion of databases that represent the patient population of interest. The inclusion of multiple databases potentially increases the generalizability of results. Furthermore, the cohorts should preferably have adequate sample size to ensure precise effect estimation, even within smaller risk strata (typically 4 risk quarters).

Step 3: Prediction. The prediction framework requires the definition of two essential cohorts: a target cohort and an outcome cohort. To generate the target cohort we pool the already defined treatment cohort T and comparator cohort C. However, for risk-based analysis of treatment effects it is necessary to develop the patient-level prediction model in a patient sample where treatment assignment is well balanced. Hereto, we use a propensity score matched patient subset on which we develop the prediction model. The propensity scores are based on LASSO logistic regression for modeling the association between treatment assignment and all available demographics, drug exposures, diagnoses, measurements and medical procedures. Finally, we need to define the time horizon for which we aim to make predictions and we need to select the machine-learning algorithm we want to use to generate patient-level predictions.

Step 4: Estimation. We use the patient-level prediction model to divide the target population into a set of equally-sized risk strata, typically 4 risk quarters. Then, we estimate propensity scores within risk strata. These propensity scores are used when estimating treatment effects, either by matching of patients from different treatment cohorts, by stratification of patients into groups with similar propensity scores, or by weighing patients' contribution to the estimation process. Within risk strata we estimate treatment effect both on the relative and the absolute scale. We use Cox proportional hazards regression to estimate relative treatment effects. We estimate absolute treatment effects by the difference between Kaplan-Meier estimates at the end of the time at risk.

Step 5: Result presentation and evaluation. Our framework provides standardized output for each step of the analysis: A) The number of patients and person years by treatment arm along with the number of outcomes; B) A performance overview of the derived prediction models, including discrimination and calibration both in the propensity score matched subset, the entire population and separately for treated and comparator patients; C) Propensity score distributions by treatment group and covariate balance plots for each risk stratum; D) Event rates, hazard ratios and absolute risk differences in risk strata for a selected outcome, both in tables and in graphs; E) Hazard ratios and absolute risk differences for all analyzed outcomes by risk stratum. A shiny application can be generated to enable easy sharing of the results.

As a proof of concept we evaluated treatment effect heterogeneity of ACE inhibitors compared to beta blockers in patients with hypertension. We considered 3 main outcomes (hospitalization with heart failure, acute myocardial infarction, stroke) and 6 safety outcomes (hypokalemia, hyperkalemia, hypotension, angioedema, cough, abnormal weight gain). Here we only present the results from CCAE, a US claims database (Table 6).

Table 6: Number of patients, person years and events within quarters of predicted risk for hospitalization with heart failure for the 3 main outcomes of the study (acute myocardial infarction, hospitalization with heart failure, and ischemic or hemorrhagic stroke).

		ACE inhibitors			Beta blockers		
Outcome	Risk quarter	Patients	Person years	Events	Patients	Person years	Events
Acute myocardial infarction	1	161,099	276,171	203	133,977	220,633	135
	2	204,882	372,197	534	90,193	169,231	321
	3	214,413	393,583	1,117	80,662	150,035	535
	4	204,167	351,727	2,095	90,908	154,419	1,520
Heart failure (hosp)	1	146,259	249,809	228	126,387	206,706	378
	2	188,006	341,014	457	84,280	158,425	340
	3	218,052	399,394	826	83,421	155,222	570
	4	230,226	400,330	2,012	98,380	169,139	1,773
Stroke (ischemic or hemorrhagic)	1	146,069	294,484	299	126,264	206,453	320
	2	187,524	340,234	554	84,000	157,913	351
	3	217,070	397,830	947	83,038	154,587	521
	4	226,128	393,861	1,718	97,628	167,810	1,077

Relative treatment effects of ACE-inhibitors vs beta blockers increased (hazard ratios decreased) with increasing acute MI risk, resulting in more pronounced increases of absolute risk difference (ARD) with increasing acute MI risk. Patients in the low-risk quarter did not receive absolute treatment benefit (ARD -0.03%) while absolute risk was 0.54% lower (95% confidence interval 0.36%-0.71%) for patients in the high-risk quarter. In contrast, the absolute and relative effects of ACE-inhibitors on safety outcomes (e.g., cough and angioedema) are approximately constant or even slightly decreasing with increasing acute MI risk (Figure 3).

This example nicely illustrates heterogeneity of absolute treatment effects, i.e., differences in absolute benefits and harms of ACE-inhibitors vs beta blockers for patients with different baseline risk. The results suggest that treatment with ACE-inhibitors, compared to treatment with beta blockers, may be focused on the higher risk patients, in whom the benefits outweigh the harms. However, treatment with beta blockers may be a viable option in lower risk patients, in whom the benefit-harm tradeoff is in favor of beta blockers.

This is in accordance with earlier findings that beta blockers should be considered as first-line treatment for younger hypertensive patients. More thorough evaluation of these results is required in future research.

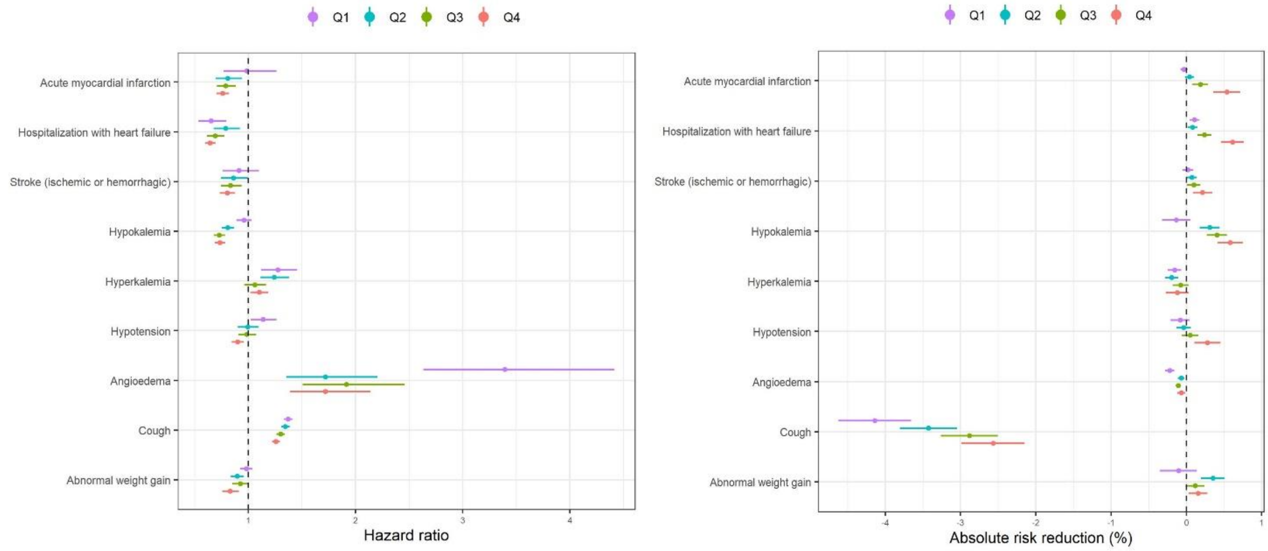


Figure 3: (Left) Hazard ratios (relative treatment effects) for the main and safety outcomes, estimated by fitting stratified Cox regression models within quarters of predicted risk of acute myocardial infarction. (Right) Absolute risk reduction for the main and safety outcomes, estimated as the difference in Kaplan-Meier estimates within quarters of predicted risk for acute MI. The four risk quarters (Q1-Q4) are defined using the internally developed model for acute MI.

The proof-of-concept study demonstrated the feasibility and power of the approach. The next step is to scale up to more drugs and outcomes as evaluated in OHDSI's LEGEND study in which a systematic, multinational, large-scale analysis was performed on the comparative effectiveness and safety of all first-line antihypertensive drug classes [20].

Published Outputs:

The developed framework has been published as pre-print and is submitted to a journal [21].
Results Explorer (Shiny App): https://data.ohdsi.org/AceBeta9Outcomes/

3.2 Disease trajectories

Patient journeys/trajectories – sequences of health events, including diseases, procedures, and visits, which patients follow – have gained more and more attention during recent years. It is now more important than ever to analyse and continuously improve such journeys to keep patients engaged with the medical care and have the best output of the treatment. Detecting the sequences of health events may also help us to better understand disease aetiology (what happened before the disease) and predict events for the future (what happened after). It can also reveal and enable to analyse different treatment options that are used in practice for the same diseases.

Observational health data is a great source for analysing such trajectories. Visits, diagnoses, lab analyses, drugs/prescriptions, etc. are all common elements in most of these datasets. These are also the key elements in OMOP CDM. This makes it extremely useful to run health event trajectory analyses also on OMOP CDM data, as the same analysis could be easily run on various databases.

In the EHDEN project, we have started developing an R-package "Trajectories" to investigate disease trajectories in OMOP CDM. It follows the principles previously published by Søren Brunak's group [22] by detecting pair-wise temporal associations between diseases. However, we have added lots of useful features to our package. It is not limited to diseases only but can also use any other health event in OMOP CDM such as observations, procedures, or drug exposures. Also, one could run the package on a specific cohort instead of the whole database. For instance, if one is interested in hypertension patients, it allows to generate trajectories for these types of patients. Perhaps the greatest value of the package is the fact that it will be completely open source as soon as it reaches a generally stable state so that anyone can run the package on top of their own CDM, evaluate its validity, and contribute to make further improvements to the tool.

In the following figure, the general workflow inside the package is shown. It includes all the events in the cohort in OMOP CDM and analyses all event pairs to detect their temporal association. As there might be lots of significant pairs, it is vital to visualize the findings effectively. In order to do so, it constructs a graph from the significant event pairs and then aligns true patient journeys based on the actual data along that graph. As a result, it generates a graph of actual patient trajectories along significantly associated temporal event pairs.

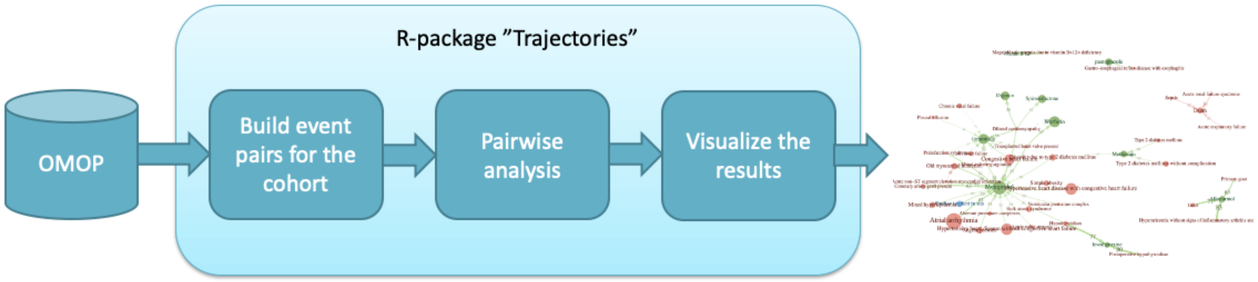


Figure 4: Workflow of Trajectories R-Package

We have tested the package currently on two health datasets available for the University of Tartu (1 million patients and 170K patients from Estonia). We are currently running it also on General Practitioner Data from Erasmus MC, results will be shared in the upcoming pipeline deliverable. In the following figure, an example output (all significant event pairs) of the run on Estonian data is shown.

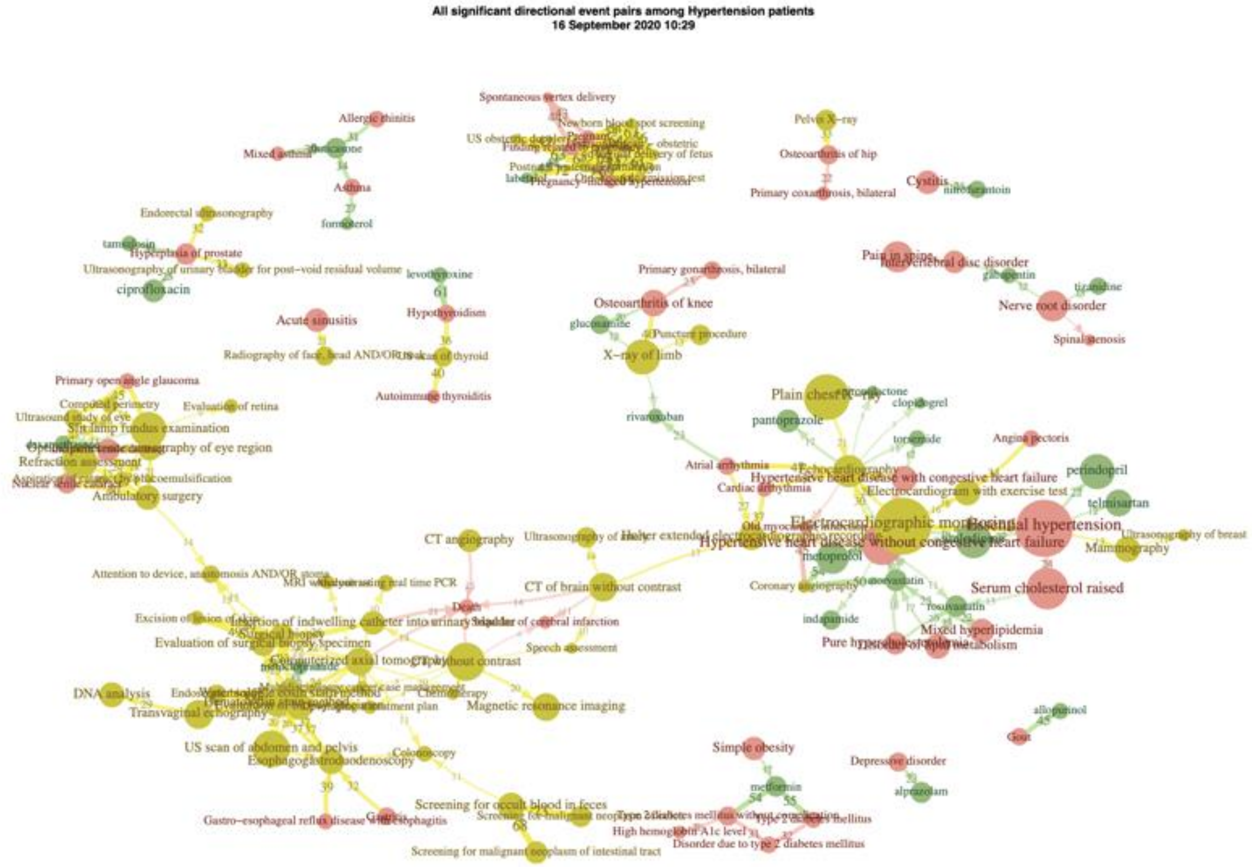

All significant directional event pairs among Hypertension patients
16 September 2020 10:29
Spontaneous vertex delivery
Newborn blood spot screening
US obstetric doppler
Pregnant
– obstetric
delivery of fetus
Pregnancy-induced hypertension
labetalol
Allergic rhinitis
Mixed asthma
fluticasone
Asthma
formoterol
Endorectal ultrasonography
tamsulosin
Hyperplasia of prostate
Ultrasonography of urinary bladder for post-void residual volume
ciprofloxacin
levothyroxine
Hypothyroidism
Acute sinusitis
Radiography of face, head AND/OR neck
US scan of thyroid
Autoimmune thyroiditis
Primary open angle glaucoma
Computed perimetry
Evaluation of retina
Ultrasound study of eye
Slit lamp fundus examination
Optical coherence tomography of eye region
Refraction assessment
phacoemulsification
Nuclear senile cataract
Ambulatory surgery
Pelvis X-ray
Osteoarthritis of hip
Primary coxarthrosis, bilateral
Cystitis
nitrofurantoin
Pain in spine
Intervertebral disc disorder
gabapentin
tizanidine
Nerve root disorder
Spinal stenosis
Primary gonarthrosis, bilateral
Osteoarthritis of knee
glucosamine
Puncture procedure
X-ray of limb
rivaroxaban
Plain chest X-ray
spironolactone
clopidogrel
pantoprazole
torsemide
Angina pectoris
Atrial arrhythmia
Echocardiography
Hypertensive heart disease with congestive heart failure
Cardiac arrhythmia
Electrocardiogram with exercise test
perindopril
telmisartan
Electrocardiographic monitoring
Essential hypertension
Old myocardial infarction
Holter extended electrocardiographic recording
Hypertensive heart disease without congestive heart failure
Ultrasonography of breast
Mammography
metoprolol
Coronary angiography
atorvastatin
Serum cholesterol raised
rosuvastatin
indapamide
Mixed hyperlipidemia
Pure hypercholesterolemia
Disorder of lipid metabolism
CT angiography
Ultrasonography of
CT of brain without contrast
Attention to device, anastomosis AND/OR stoma
MRI
real time PCR
Death
Excision of lesion of skin
Insertion of indwelling catheter into urinary bladder
of cerebral infarction
Surgical biopsy
Evaluation of surgical biopsy specimen
Speech assessment
Computerized axial tomography
CT without contrast
DNA analysis
Chemotherapy
treatment plan
Magnetic resonance imaging
Transvaginal echography
US scan of abdomen and pelvis
Esophagogastroduodenoscopy
Colonoscopy
Gastro-esophageal reflux disease with esophagitis
Gastritis
Screening for occult blood in feces
Screening for malignant neoplasm
Screening for malignant neoplasm of intestinal tract
Simple obesity
metformin
Type 2 diabetes mellitus without complication
Type 2 diabetes mellitus
High hemoglobin A1c level
Disorder due to type 2 diabetes mellitus
Depressive disorder
alprazolam
allopurinol
Gout

Figure 5: Example of disease trajectories showing in yellow procedures, in green drugs, and in red conditions

If one is interested in a particular event – for instance, essential hypertension – it can be fixed in the graph and actual patient journeys through this event are then aligned to the graph. The example result of this is shown in the next figure. We can see that 33% of essential hypertension patients are directly followed by electrocardiographic monitoring and 1% will get atorvastatin as a next step after monitoring.

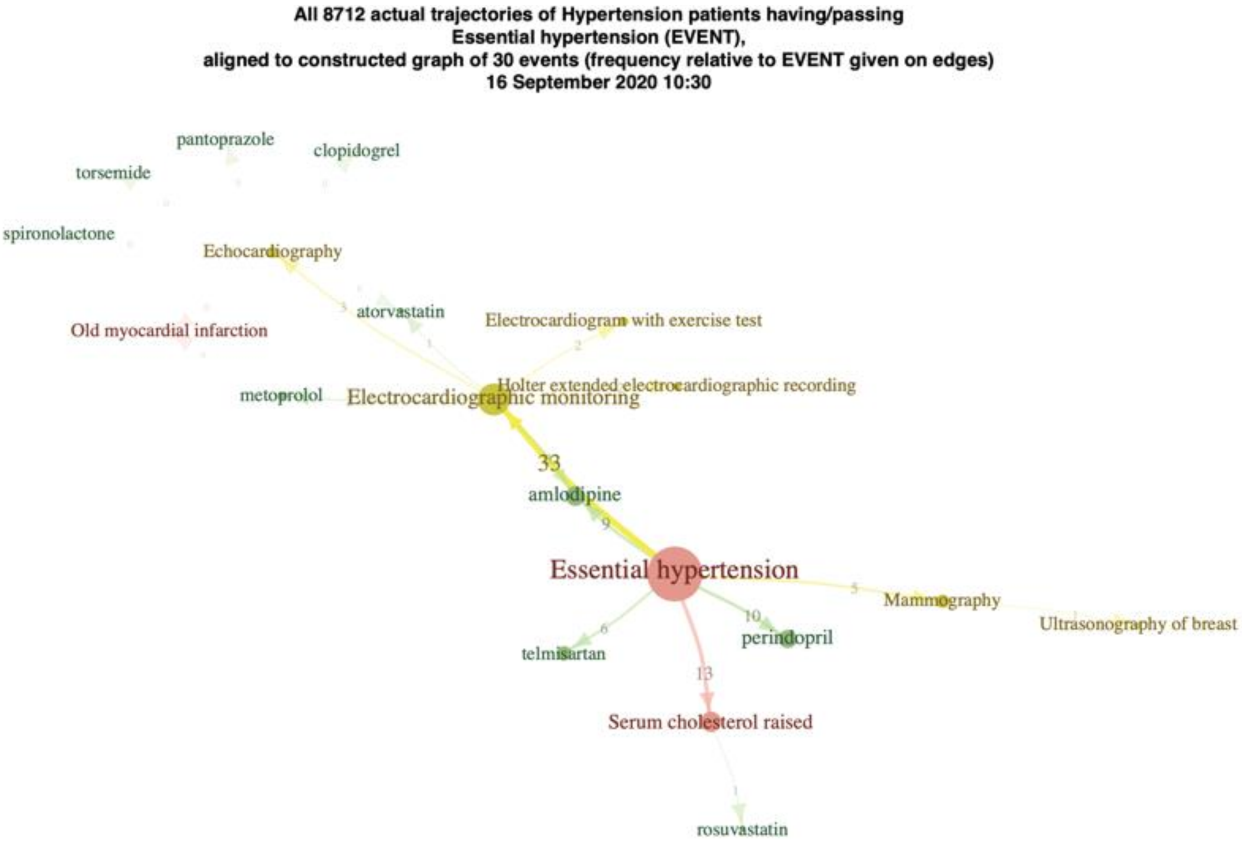

All 8712 actual trajectories of Hypertension patients having/passing
Essential hypertension (EVENT),
aligned to constructed graph of 30 events (frequency relative to EVENT given on edges)
16 September 2020 10:30
pantoprazole
clopidogrel
torsemide
spironolactone
Echocardiography
atorvastatin
3
Electrocardiogram with exercise test
1
2
Old myocardial infarction
metoprolol
Electrocardiographic monitoring
Holter extended electroeardiographic recording
33
amlodipine
9
Essential hypertension
5
Mammography
Ultrasonography of breast
10
perindopril
6
telmisartan
13
Serum cholesterol raised
1
rosuvastatin

Figure 6: Example trajectory of essential hypertension, showing in yellow procedures, in green drugs, and in red conditions

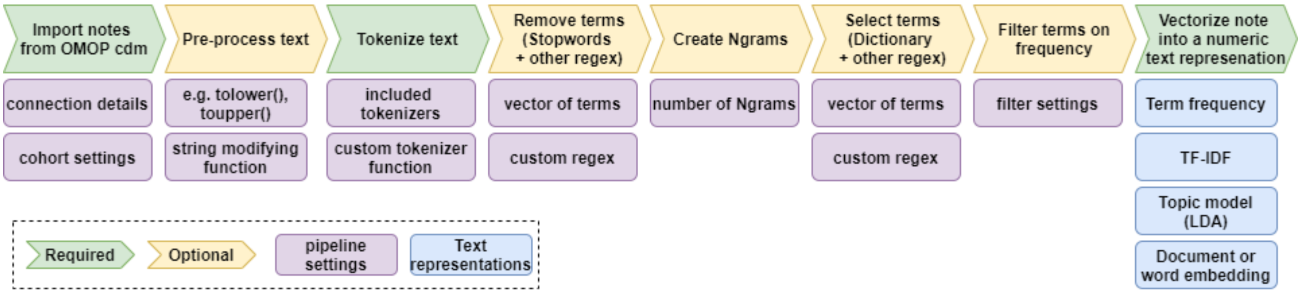
Before releasing the package to the public, we need to test it on various databases and database engines. We are currently working on this. Also, we are currently adding package documentation and unit tests to ensure the validity of its outputs as described in the Software Validity Chapter of the Book of OHDSI (https://ohdsi.github.io/TheBookOfOhdsi/SoftwareValidity.html)

3.3 Natural Language Processing

Electronic health record (EHR) databases are a rich source of data for building patient level prediction models. Currently, most prediction models use only the structured data in the EHR, such as coded conditions, measurements, vital signs, and drug prescriptions, as features [1]. However, EHRs commonly also store vast amounts of unstructured textual data (e.g., physician's and nurse's notes and discharge letters) [23]. Using natural language processing (NLP) methods the information hidden in the unstructured clinical text can be extracted and incorporated in PLP models.

We developed a standardized 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. The tool contains a modular NLP pipeline for the pre-processing, tokenization and vectorization of text documents, that can be fully customized to specific needs. The pipeline settings and customizations are saved with the result for sharing and reproducibility. The tool is called *Text Represented In Terms Of Numeric-features* (TRITON) and is now publicly available on GitHub at github.com/mi-erasmusmc/Triton.

TRITON can pre-process the text in any way (for example a conversion to lowercase), tokenize it using various (or custom) tokenization methods, remove language dependent stop words, create term ngrams, and filter terms based on their absolute and relative frequency, before creating a vectorized text representation (see figure 7 for an overview). Two bag-of-word text representations are currently supported, the term frequency (TF) and the term frequency-inverse document frequency (TFIDF [24]). More text representations such as topic models (latent Dirichlet allocation [25]) and word and document embeddings (GloVe [26], doc2vec [27]) are planned to be implemented. Furthermore, extra options for dictionary-based approaches will be added in the future, to find specific words, together with spelling correction options and the possibility of detecting contextual information (e.g., negation, experiencer, temporal aspects, severity).



Import notes from OMOP cdm
Pre-process text
Tokenize text
Remove terms (Stopwords + other regex)
Create Ngrams
Select terms (Dictionary + other regex)
Filter terms on frequency
Vectorize note into a numeric text represenation
connection details
cohort settings
e.g. tolower(), toupper()
string modifying function
included tokenizers
custom tokenizer function
vector of terms
custom regex
number of Ngrams
vector of terms
custom regex
filter settings
Term frequency
TF-IDF
Topic model (LDA)
Document or word embedding
Required
Optional
pipeline settings
Text representations

Figure 7. An overview of the TRITON natural language processing pipeline. Each step in the pipeline has specific settings.

We provide a language independent, scalable, and customizable tool for the generation of numeric text representation covariates, that can readily be generated from any OMOP CDM database that contains unstructured text data. Subsequently, the created covariates can be used by the analysis tools within the OHDSI framework, e.g., in population-level estimation and patient-level prediction.

In the upcoming period we will further develop the NLP pipeline and will apply it to multiple use cases.

4 Methods Research

In addition to analytical pipeline development as described above, WP3 is also performing methods research driven by EHDEN use cases. In the next sections the work conducted in the second year is described.

4.1 Predictive analytics using unstructured data

In contrast to coded data, text data lacks an organized structure and terminology. Moreover, text data can be very large and often contains patient-sensitive information. This makes it difficult to manage, analyse, and use the text data in the development of prediction models [28]. However, the information that is captured in the clinical text is generally more extensive and detailed than the coded data. Previous studies have described the value of text data in addition to structured data in tasks such as phenotyping or case identification [29]. However, the value of this information to improve patient-level prediction models has largely been unexplored.

The aim of this EHDEN use case is to determine the added value of textual data in EHRs for improving patient-level prediction models. 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 (see section 4.3) and tested for their predictive value. First, as a baseline, bag-of-word approaches will be considered, including raw term frequencies and TFIDF values. Second, dictionary-based approaches will be used to identify and extract relevant concepts (e.g., symptoms, conditions, procedures) in the texts. For concept recognition, deep-learning based language models can be explored (e.g., BERT [30], UMLFit [31]). Thirdly, topic modelling (latent

Dirichlet allocation) and word or document embedding methods (e.g. GloVe, doc2vec) will be used to generate a specific number of new features that capture semantic information.

Different text pre-processing steps can be performed. From simple processing steps, such as converting sentences to lower case, to the detection of contextual information (e.g., negation, and temporality). Furthermore, spelling correction algorithms can be applied [32]. The impact of these pre-processing steps on predictive performance will be assessed.

Two important aspects that will be addressed are multi-linguality and scalability. Since EHR data may stem from different (European) countries, the investigated methods need to be able to cope with several different European languages. Validating the prediction models between countries (with different languages) will require a standardized multilingual dictionary. SNOMED CT and the Unified Medical Language System (UMLS) will be used as the main resources for English and non-English dictionaries. If a dictionary is not or only partially available in a non-English language of interest, we will apply automatic machine translation.

Using the TRITON framework, it is be possible to apply the methods on a large scale, on every CDM database (that contains unstructured text data), and for many different outcomes and covariates. Because the methods will need to be scalable, they require the use of unsupervised and semi-supervised methods as much as possible. Most of the proposed methods fall in these categories. Furthermore, we will investigate the use of pre-trained models and unlabelled data sets (BERT, UMLFit).

The primary data source for method development and evaluation will be the Dutch Integrative Primary Care Information (IPCI) database, comprising general practitioner records and specialist letters of more than 2 million patients in the Netherlands. Most of the information in IPCI is in unstructured, textual format. Currently, only the coded information in IPCI has been mapped to the OMOP CDM. Additional data sources with textual data, in Dutch or other European languages, will be included when they become available in the project. Using the TRITON framework and NLP pipeline and other pre-trained models the potential value of textual information for patient-level prediction will be assessed for a wide range of outcomes and target cohorts.

4.2 Learning Curves

EHDEN's federated data network opens up possibilities to develop clinical prediction models on massive amounts of patient data which can serve large patient populations in a timely manner. In practice this could manifest in the development of several hundreds or even thousands of prediction models for the various target-outcome pairs and for the many different databases that are currently being mapped to the OMOP CDM.

However, models developed on these large amounts of observational health data run the risk of being more complex than needed. These models can include many more features without achieving substantially better discrimination than smaller models. As a result, these models may become harder to interpret, more difficult to implement in clinical practice, and more susceptible to overfitting. In addition, developing prediction models on such large data sources can put strong demands on computing resources and may require computation times that can become prohibitive. Reducing the sample size of a large and unwieldy dataset to an "adequate" sample size that is still sufficient to achieve nearly the same performance as the full dataset, may facilitate the development of less complex clinical prediction models with less computing resources.

The objective of this study is to provide guidance on sample size considerations for developing predictive models by empirically establishing the adequate sample size, which balances the competing objectives of improving model performance and reducing model complexity as well as computational requirements.

For this study we empirically assessed the effect of sample size on prediction performance and model complexity by generating learning curves for 81 prediction problems in three large observational health databases, requiring training of 17,248 prediction models. The adequate sample size was defined as the sample size for which the performance of a model was equal to the maximum model performance minus a small threshold value.

Figure 8 exemplifies the approach for a single learning curve showing the reduction in the number of events and the number of predictors of an adequate model as compared to a full model.

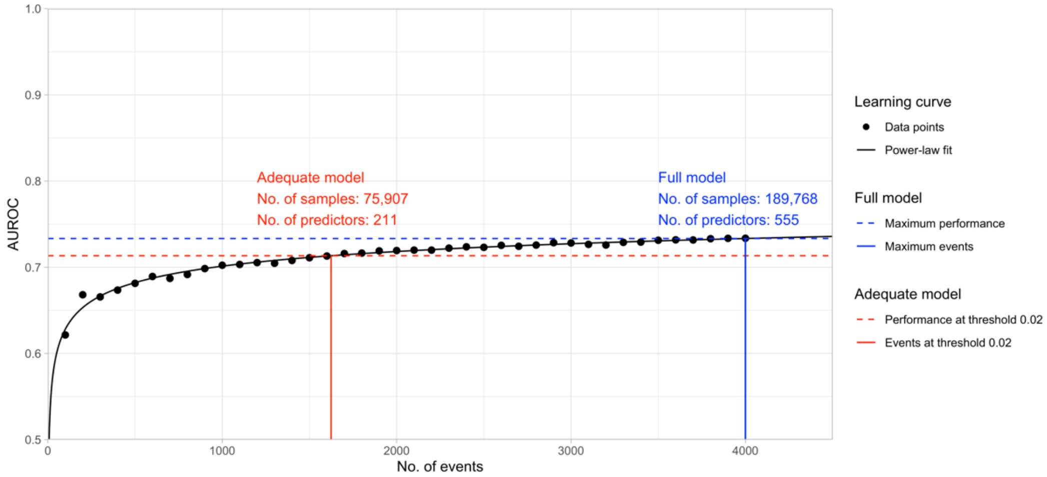


Figure 8. Exemplary learning curve where the horizontal lines indicate the maximum performance of full model (blue) and the performance of the adequate model at a threshold of 0.02 (red). The vertical lines denote the maximum number of events (blue) and the adequate number of events (red).

In Figure 9, we observe that the adequate sample size achieves a median reduction of the number of events between 9.5% and 78.5% for threshold values between 0.001 and 0.02. Moreover, the median reduction of the number of predictors in the models at the adequate sample size varies between 8.6% and 68.3%, respectively.

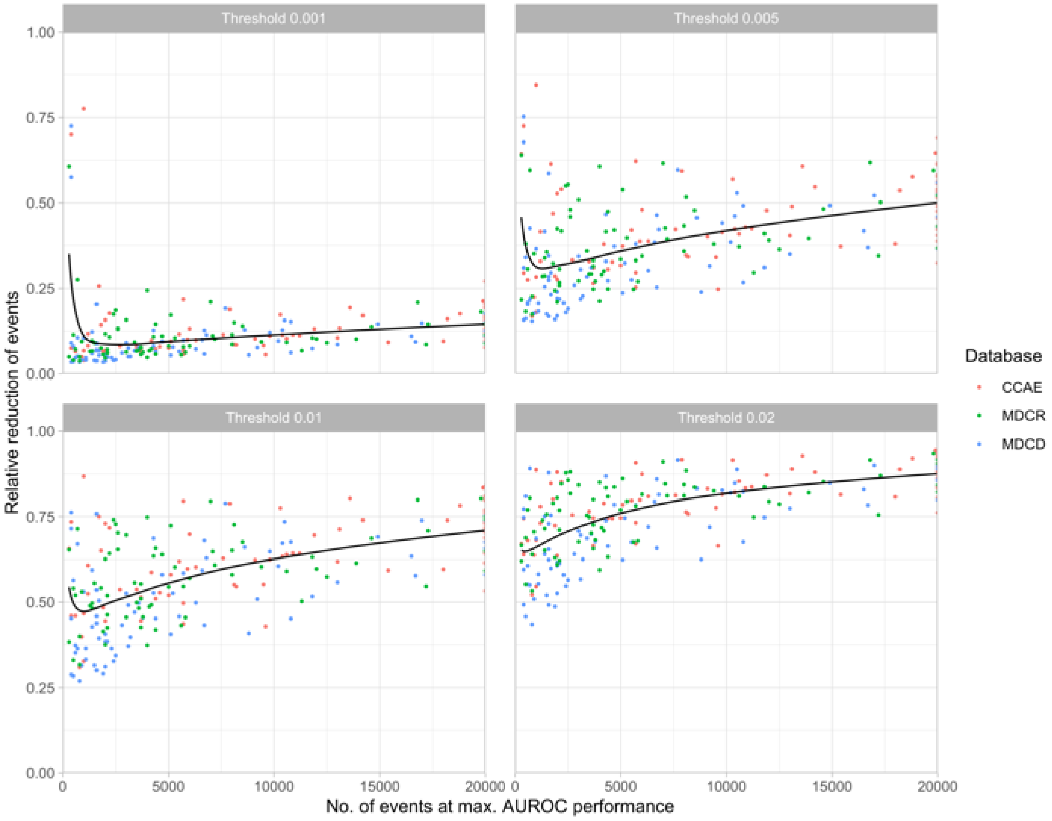


Figure 9. Relative reduction of the number of events at thresholds of 0.001, 0.005, 0.01, and 0.02 plotted against the number of events at the maximum AUROC performance for multiple prediction problems in three large databases.

Therefore, our results suggest that in most cases only a fraction of the available data was sufficient to produce a model close to the performance of one developed on the full data set, but with a substantially reduced model complexity.

Published Output

A preprint of this work is available in the arXiv repository [33].

4.3 Explainable AI

Lack of transparency is identified as one of the main barriers to implement patient-level prediction models in clinical practice [34, 35]. As it is the responsibility of clinicians to give the best care to each patient, they should be confident that AI systems (i.e., the prediction model and other parts of the implementation) can be trusted. A possible step towards trustworthy AI is to develop explainable AI. The field of explainable AI aims to create insight into how and why models produce predictions, while maintaining high predictive performance levels. Although the field of explainable AI has promising prospects for health care, it is not fully developed yet. We have reviewed the current literature to provide guidance on the design of explainable AI systems for the health-care domain.

For an AI system to be *explainable*, we argue we need both interpretability and fidelity. The *interpretability* of an explanation captures how understandable an explanation is for humans. The *fidelity* of an explanation expresses how accurately an explanation describes model behavior, i.e. how faithful an explanation is to the task model. The task model is the model generating predictions.

There are different methods to achieve explainability. One way to achieve explainable AI is by *explainable modelling*, i.e., by developing an AI model where the internal functioning is directly accessible to the user, so that the model is intrinsically interpretable. Alternatively, post-hoc explanations can accompany the AI model to make it explainable. *Post-hoc explanations* can be motivated by the trade-off between predictive performance and interpretability. Hence, instead of developing an intrinsically interpretable model with the risk of a lower predictive performance, post-hoc explanations accompany the AI model and provide insights without knowing how the AI model works. We further classify explainable AI techniques according to the type of explanation and the scope of explanation. We distinguish three types of explanations in the literature: model-based explanations (e.g., rule-based model), attribution-based explanations (e.g., feature importance), and example-based explanations (e.g., counterfactual explanation). Furthermore, the scope of explanation can be local (i.e. explaining an individual prediction) or global (i.e., explaining the entire model).

We argue that the reason to demand explainability determines what should be explained as this determines the relative importance of the properties of explainability (i.e., interpretability and fidelity). We identify three reasons why explainability can be required: 1) to assist in verifying (or improving) other model desiderata, 2) to manage social interaction, or 3) to discover new insights. However, explanations can be costly and might only be needed when the cost of misclassification is high or the AI system has not yet proven to work well in practice. In the health-care domain both situations often apply.

We propose the following step-by-step guide to select the most appropriate class of explainable AI methods:

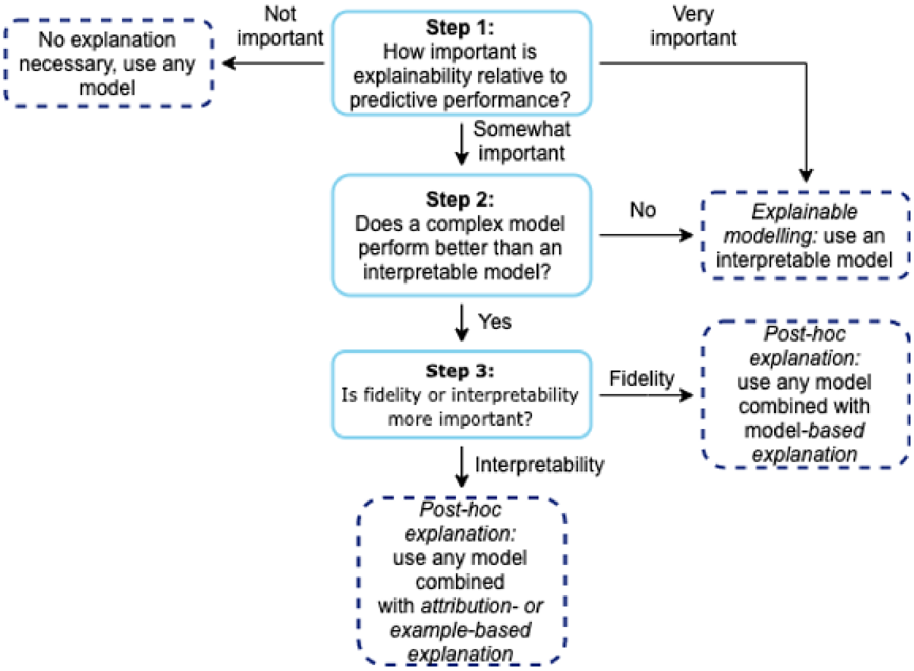

Step 1:
How important is explainability relative to predictive performance?
Not important
No explanation necessary, use any model
Very important
Somewhat important
Step 2:
Does a complex model perform better than an interpretable model?
No
Explainable modelling: use an interpretable model
Yes
Step 3:
Is fidelity or interpretability more important?
Fidelity
Post-hoc explanation: use any model combined with model-based explanation
Interpretability
Post-hoc explanation: use any model combined with attribution- or example-based explanation

Figure 10: Step-by-step guide with recommendations to choose between classes of explainable AI methods.

As we believe explainability is very important to create trustworthy AI for health care, we conclude that explainable modelling might be preferred over post-hoc explanations. If one wants to opt for a post-hoc explanation, model-based explanations are the preferred type of explanation. In the near future, we aim to extend the patient-level prediction pipeline with more explainable models (e.g., rule-based learners) and investigate whether a trade-off between predictive performance and interpretability occurs. Moreover, we will explore how we can enhance the interpretability of patient-level prediction models by incorporating expert knowledge in the model development process.

Published Output

A preprint of this work is available in the arXiv repository [36].

5. Next Steps

In the second year, again major steps have been taken in the development of analytical pipelines for personalized medicine and a significant amount of work has been done in methods research.

Most importantly we have applied the patient-level prediction and population-level effect estimation pipelines for several high-impact clinical problems, which resulted in multiple manuscripts. The studies for COVID-19 are especially good examples of the power of the standardization to the OMOP CDM and the strength of a strong collaborative community.

The next steps for WP3 are:

1. Extending the methods research to build predictive models that transport well to other settings. This includes training on subsets of variables that are frequent in all database, federated learning, and addition of functionality to perform re-calibration in the framework.
2. Further research on the use of multilingual unstructured text in the context of prediction.
3. Research on frequent pattern mining has recently started and will be implemented against the CDM and tested in use cases.
4. Further development of the disease trajectory pipeline and application across the data network.
5. Application of the expertise gained in Explainable AI in the field of patient-level prediction.

Furthermore, we will apply the analytical pipeline to more use cases in the upcoming year. The pipelines will be further updated in an agile manner driven by these use cases. EHDEN continues to collaborate with OHDSI in large-scale studies such as the recently started SARS-Cov-2 Large-scale Longitudinal Analyses (SCYLLA) project on the comparative safety and effectiveness of treatments under evaluation for COVID-19 across an international observational data network. The EHDEN data network will grow considerably in the coming period and we look forward to invite these European data partners to participate in the upcoming use cases.

REFERENCES

1. 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.
2. World Health Organization, *Coronavirus disease 2019 (COVID-19) Situation report*, in *World Health Organization*. 2020.
3. Scally, G., B. Jacobson, and K. Abbasi, *The UK's public health response to covid-19*. 2020, British Medical Journal Publishing Group.
4. Williams, R.D., et al., *Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network.* medRxiv, 2020.
5. Gabriel, S.E. and K. Michaud, *Epidemiological studies in incidence, prevalence, mortality, and comorbidity of the rheumatic diseases.* Arthritis research & therapy, 2009. **11**(3): p. 229.
6. Rudan, I., et al., *Prevalence of rheumatoid arthritis in low–and middle–income countries: A systematic review and analysis.* Journal of global health, 2015. **5**(1).
7. Dougados, M., et al., *Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA).* Annals of the rheumatic diseases, 2014. **73**(1): p. 62-68.
8. Turesson, C., *Comorbidity in rheumatoid arthritis.* Swiss medical weekly, 2016. **146**(1314).
9. Listing, J., K. Gerhold, and A. Zink, *The risk of infections associated with rheumatoid arthritis, with its comorbidity and treatment.* Rheumatology, 2013. **52**(1): p. 53-61.
10. Peters, M., et al., *EULAR evidence-based recommendations for cardiovascular risk management in patients with rheumatoid arthritis and other forms of inflammatory arthritis.* Annals of the rheumatic diseases, 2010. **69**(2): p. 325-331.
11. Turesson, C. and E.L. Matteson, *Malignancy as a comorbidity in rheumatic diseases.* Rheumatology, 2013. **52**(1): p. 5-14.
12. Visser, K. and D. van der Heijde, *Optimal dosage and route of administration of methotrexate in rheumatoid arthritis: a systematic review of the literature.* Annals of the rheumatic diseases, 2009. **68**(7): p. 1094-1099.
13. Lane, J.C., et al., *Safety of hydroxychloroquine, alone and in combination with azithromycin, in light of rapid wide-spread use for COVID-19: a multinational, network cohort and self-controlled case series study.* medRxiv, 2020.
14. Dimitrova, E., *COVID-19: reminder of risk serious side effects with chloroquine and hydroxychloroquine.* Eur. Med. Agency, 2020.
15. Lane, J.C., et al., *Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study.* The Lancet Rheumatology, 2020. **2**(11): p. e698-e711.
16. EMA, *COVID-19: reminder of the risks of chloroquine and hydroxychloroquine.* European Medicines Agency, 2020.
17. Lane, J.C., et al., *Risk of depression, suicidal ideation, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multi-national network cohort study.* medRxiv, 2020.
18. EMA, *The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology.* Eur. Med. Agency, 2020.
19. Rekkas, A., et al., *Predictive approaches to heterogeneous treatment effects: a scoping review.* BMC Medical Research Methodology, 2020. **20**(1): p. 1-12.
20. Suchard, M.A., et al., *Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis.* The Lancet, 2019. **394**(10211): p. 1816-1826.

21. Rekkas, A., et al., *A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases.* arXiv preprint arXiv:2010.06430, 2020.
22. Afshar, M., et al., *Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies.* Journal of the American Medical Informatics Association, 2019. **26**(11): p. 1364-1369.
23. Ramos, J. *Using tf-idf to determine word relevance in document queries*. in *Proceedings of the first instructional conference on machine learning*. 2003. New Jersey, USA.
24. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* Journal of machine Learning research, 2003. **3**(Jan): p. 993-1022.
25. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
26. Lau, J.H. and T. Baldwin, *An empirical evaluation of doc2vec with practical insights into document embedding generation.* arXiv preprint arXiv:1607.05368, 2016.
27. Kreimeyer, K., et al., *Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review.* Journal of biomedical informatics, 2017. **73**: p. 14-29.
28. Ford, E., et al., *Extracting information from the text of electronic medical records to improve case detection: a systematic review.* Journal of the American Medical Informatics Association, 2016. **23**(5): p. 1007-1015.
29. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
30. Howard, J. and S. Ruder, *Universal language model fine-tuning for text classification.* arXiv preprint arXiv:1801.06146, 2018.
31. Fivez, P., S. Suster, and W. Daelemans. *Unsupervised context-sensitive spelling correction of clinical free-text with word and character n-gram embeddings*. in *BioNLP 2017*. 2017.
32. John, L.H., et al., *How little data do we need for patient-level prediction?* arXiv preprint arXiv:2008.07361, 2020.
33. He, J., et al., *The practical implementation of artificial intelligence technologies in medicine.* Nature medicine, 2019. **25**(1): p. 30-36.
34. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nature medicine, 2019. **25**(1): p. 44-56.
35. 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.* arXiv preprint arXiv:2007.15911, 2020.