SAFETY-VAC

A framework for the post-authorization safety monitoring and evaluation of the vaccines in Europe.

Study protocol to provide and describe a network of real-world data sources for the evaluation of vaccine safety signals, and to assess its fitness-for-purpose in conducting vaccine safety studies.

Protocol V1.0

March 28, 2024

Title	SAFETY-VAC **A framework for the post-authorisation safety monitoring and evaluation of vaccines in the EU**
Protocol version identifier	1.0
Date of last version of protocol	March 28, 2024
EU PAS register number	
Active substance	*NA*
Medicinal product	*Vaccines J07*
Product reference	*NA*
Procedure number	*NA*
Marketing authorisation holder(s)	*NA*
Research question and objectives	The SAFETY-VAC project assesses the feasibility of participating data sources to participate in vaccine safety studies using electronic healthcare databases in European countries. The primary objective in this specific study is to provide and describe a network of real-world data sources for the evaluation of vaccine safety signals, and to assess its fitness-for-purpose through two specific objectives: **Objective a.** To assess data quality for conducting safety studies on the population, specific vaccines, and selected outcomes. **Objective b.** To assess whether data are fit-for-purpose for conducting safety studies on specific vaccines and outcomes in a near real-time monitoring manner in the future.
Country(-ies) of study	United Kingdom Spain Denmark Finland Norway Italy France
Author	Miriam Sturkenboom (University Medical Center Utrecht, the Netherlands) Carlos E. Durán (University Medical Center Utrecht, the Netherlands)

TABLE OF CONTENTS

1 TITLE

SAFETY-VAC: A framework for the post-authorization safety monitoring and evaluation of the vaccines in the EU.

Study protocol to provide and describe a network of real-world data sources for the evaluation of vaccine safety signals, and to assess its fitness-for-purpose in conducting vaccine safety studies.

Document version: V1.0

2 ABBREVIATIONS

ACCESS	vACCine covid-19 monitoring readinESS
ADVANCE	Accelerated Development of VAccine beNefit-risk Collaboration in Europe
AESI	Adverse Event of Special Interest
ARDS	Acute respiratory distress requiring ventilation
ATC	Anatomical Therapeutic Chemical
BMI	Body Mass Index
CDC	Centers for Disease Control and Prevention
CDM	Common Data Model
CI	Confidence interval
DAP	Data Access Provider
DRE	Digital Research Environment
EMA	European Medicines Agency
EMR	Electronic Medical Records
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance.
ETL	Extract, Transform, and Load
EU PAS	The European Union electronic Register of Post-Authorisation Studies
GDPR	General Data Protection Regulation
GP	General Practitioner
GPP	Good Participatory Practice
HIV	Human Immunodeficiency Virus
ICD	International Classification of Diseases
ICMJE	International Committee of Medical Journal Editors
ICU	Intensive Care Unit
MIS-C	Multisystem Inflammatory Syndrome in children
mRNA	messenger Ribonucleic acid
NHS	National Health Service
QC	Quality Control
RNA	Ribonucleic acid
SAP	Statistical Analysis Plan
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SPEAC	Safety Platform for Emergency vACcines
VAC4EU	Vaccine monitoring Collaboration for Europe

3 MARKETING AUTHORISATION HOLDER

Not applicable (N/A)

4 RESPONSIBLE PARTIES

University Medical Center Utrecht (UMCU), Utrecht, The Netherlands	Dr. Carlos E. Durán, Prof. Dr. Miriam Sturkenboom, Dr. Judit Riera, Nicoletta Luxi Vjola Hoxhaj,
Universiteit Utrecht (UU), Utrecht, The Netherlands CPRD data	Prof. Dr. Olaf Klungel, Dr. Patrick Souverein, Dr. Satu Johanna Siiskonen
VAC4EU	Dr. Sima Mohammadi
Teamit Institute	Dr. Fabio Riefolo, Dr. Irene Pazos
Agenzia Regionale di Sanitá Toscana (ARS) Data Tuscany region	Dr. Rosa Gini, Davide Messina, Dr. Giuseppe Roberto
Società Servizi Telematici -Pedianet	Prof. Dr. Carlo Giaquinto, Dr. Elisa Barbieri, Luca Stona
Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAP JGol)	Dr. Felipe Villalobos, Dr. Martín Solorzano, Carlo Alberto Bissacco
Instituto Aragonés de Ciencias de la Salud (IACS)	Dr. Antonio Gimeno, Dr. Beatriz Poblador, Dr. Mercedes Aza, Dr. Aida Moreno, Alejandro Santos
Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital	Prof. Vera Ehrenstein, Lise Skovgaard Svingel, Benjamin Randeris Johannesen
Bordeaux PharmacoEpi platform (BPE) & ADERA	Dr. Cécile Droz-Perroteau, Laure Carcaillon-Bentata,
University of Eastern Finland	Prof. Anna-Mija Tolppanen, Prof. Sirpa Hartikainen, Dr Thuan Vo, Dr Anne Paakinaho, Blair Rajamaki
University of Oslo (UiO), Norway Norwegian linked registry data	Prof. Dr. Hedvig Nordeng, Saeed Hayati, Mahmoud Zidan
Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO) – Valencia health system Integrated Database (VID)	Dr. Juan José Carreras Martínez, Dr. Arantxa Urchueguía Fornes, Elisa Correcher Martínez, Dr. Javier Díez-Domingo
Spanish Agency on Medicines and Medical Devices (AEMPS) -BIFAP database	Dr. Mar Martin, Dr. Patricia Garcia-Poza, Dr. Airam de Burgos, Belén Castillo-Cano, Dr. Elisa Martín-Merino

5 ABSTRACT

5.1 SAFETY-VAC Study: A framework for the post-authorisation SAFETY monitoring and evaluation of VACcines in the EU.

Study protocol to provide and describe a network of real-world data sources for the evaluation of vaccine safety signals, and to assess its fitness-for-purpose in conducting vaccine safety studies.

Document version: V1.0

Main authors:

- Prof. dr. M.C.J.M. Sturkenboom, University Medical Center Utrecht, The Netherlands
- Dr. CE Durán, University Medical Center Utrecht, The Netherlands

5.2 Rationale and background:

Numerous vaccines based on novel technologies targeting different diseases are continuously under development and obtaining marketing authorization. However, safety assessment is often limited to pre-authorisation clinical trials and new concerns are expected to arise during the post-authorisation phase. Thus, it is essential to create, assess, and describe a network of real-world data sources that are fit-for-purpose to address upcoming safety related questions, and to do so in a timely manner.

In May 2022, The European Medicines Agency (EMA) together with the European Centre for Disease Prevention and Control (ECDC) established the Vaccine Monitoring Platform (VMP) and the objective of generating a real-world evidence (RWE) framework for post-authorisation safety evaluation that can be leveraged in case of a new public health emergency or a safety concern occurring with a novel, or a more characterised, vaccine authorized in the European Union (EU) and the European Economic Area (EEA).

5.3 Research question and objectives:

To provide and describe a network of real-world data sources for the evaluation of vaccine safety signals, and to assess its fitness-for-purpose in conducting vaccine safety studies.

1. To assess the data quality for the purpose of conducting safety studies in the network and to describe data source population, capture of routine immunizations, and selected outcomes.
2. To assess whether data are fit for purpose for conducting vaccine safety studies.

5.4 Study design:

We will conduct a retrospective, multi-database, population-based cohort design during the study period from January 1st, 2017, till the last data availability. The work will start with accessing data from 10 different electronic health record data sources that have proven to be able to convert data (n=9) into the ConcePTION common data model (CDM) or are willing to do this (n=1).

5.5 Population and study size:

The source population of the 10 data sources comprises 68,300,000 persons. The source population comprises all persons in the data sources that participate in each objective.

5.5.1 Inclusion, exclusion criteria and follow-up

Persons of any age will be included in the study population when they comply with the following criteria:

- They have information on age and gender available.
- They have at least one day of follow in the study period (1/1/2017- latest availability).

Follow up will start on the latest of the following dates: day that one year of lookback is available during the study period, or at birth for those born during the study period.

Follow-up will finish at the earliest of the following dates: death, disenrollment, recommended end date. The recommended end date is the latest date that the Data Access Provider (DAP) recommends having information from data banks complete.

5.6 Variables:

All variables will be identified in the data sources by using diagnostic or procedure codes, medicines, and vaccines. Variables of interest will be:

Outcomes:

Thirty-nine events have been selected together with EMA to assess data sources preparedness for a wide range of outcomes. Clinical definition forms and codes' (including ICD-9, ICD-10, SNOMED, and ICPC codes) will be generated using the standardized VAC4EU process for identifying the events.

The selected thirty-nine outcomes are presented in table 1.

Table 1. List of selected events to be studied.

N	Name of the event
1.	Microangiopathy (MA)
2.	Acute coronary artery disease (CAD)
3.	Arrhythmia
4.	Myocarditis
5.	Pericarditis
6.	Venous thromboembolism (VTE)
7.	Arterial thrombosis (AMI /Ischemic stroke)
8.	TTS (VTE, arterial thrombosis, or CVST with thrombocytopenia in 10 days)
9.	Pulmonary embolism
10.	Haemorrhagic stroke (PE)
11.	Disseminated intravascular coagulation (DIC)
12.	Cerebral venous sinus thrombosis (CVST)
13.	Generalised convulsion
14.	Guillain Barré Syndrome (GBS)
15.	Diabetes (type 1)
16.	Single organ cutaneous vasculitis (SOCV)

N	Name of the event
17.	Erythema multiforme (EM)
18.	Meningoencephalitis
19.	Acute disseminated encephalomyelitis (ADEM)
20.	Narcolepsy
21.	Thrombocytopenia (TP)
22.	Transverse myelitis
23.	Bells’ palsy
24.	Kawasaki's disease (KD)
25.	Pancreatitis
26.	Rhabdomyolysis (RML)
27.	Severe cutaneous adverse reactions to drugs (SCARs)
28.	Sensorineural hearing loss (SNHL)
29.	Graves' disease (GD)
30.	Hashimoto's thyroiditis (HT)
31.	Auto-immune hepatitis (AIH)
32.	Polyarteritis nodosa (PAN)
33.	Rheumatoid arthritis (RA)
34.	Psoriatic arthropathies (PsA)
35.	Systemic lupus erythematosus (SLE)
36.	Idiopathic thrombocytopenic purpura (ITP)
37.	Erythema nodosum (EN)
38.	Multiple sclerosis
39	Ulcerative colitis (UC)

Exposure:

The following vaccines that will be included in the fit for purpose assessment and assessed in specific cohorts, nested in the study cohort:

- Measles-containing vaccines (doses 1, 2)
- Diphtheria, tetanus toxiod, and pertussis (dose 1, 2, 3)
- Haemophilus influenzae type B (doses 1, 2, 3)
- Hepatitis B (doses 1, 2, 3)
- Polio (doses 1, 2, 3)
- Pneumococcal conjugate vaccines (doses 1, 2)
- Varicella (dose 1)
- Bacille Calmette-Guérin vaccine (dose 1)
- Human papillomavirus vaccine (doses 1, 2)
- Rotavirus (doses 1, 2)
- Meningoccocal vaccine (doses 1, 2)
- Influenza vaccine (dose 1)
- COVID-19 vaccines (doses 1 to 6)

Covariates

The following covariates will be assessed:

- Age
- Gender
- Transplantation
- Immunocompromised status
- Pregnancy
- Hypertension

- Lipid abnormalities
- Malignancies
- HIV
- Cardiocerebrovascular disease
- Heart failure
- Diabetes
- Valvular heart disease
- Inflammatory bowel disease
- Coronary artery disease
- Myocardial infarction
- Arrhythmia
- VTE
- Infection
- Liver disease
- Alcohol abuse
- Sepsis
- Chronic renal disease
- Dementia
- Respiratory infections
- Herpes simplex
- Influenza
- Sleep disorders
- Mental health diseases
- Preeclampsia
- Hepatitis C
- Rheumatoid arthritis
- SLE
- Dermatomyositis
- Sjogren’s syndrome
- Gallstones
- Sickle cell disease
- Myasthenia gravis
- Pernicious anemia
- Autoimmune hepatitis
- Celiac disease
- Hepatitis B
- Psoriasis
- Gout
- Crohn's disease
- Ulcerative colitis
- Atopic dermatitis
- Immune thrombocytopenia
- Nonalcoholic fatty liver
- Obesity
- Dermatomyositis

5.7 Data sources:

The study will include data from 10 electronic health care databases that are population-based in 7 countries in Europe (UK, Spain, Denmark, Finland, Norway, Italy and France). The characteristics of each of the participating DAPs are summarized in the following table:

Table 2. Data sources to be included in this study.

DAP	Data source	Country	Population size	Data banks available for this study	Vocabularies	Data update frequency
AEMPS	BIFAP	Spain	17.0 million	Primary care record, hospital discharge diagnosis, community pharmacy dispensing, date of death.	ICPC2, ICD9 and SNOMED for diagnosis. ATC for medicines.	2 months.
IDIAP J Gol	SIDIAP	Spain	5.8 million	Primary care record, outpatient specialist record, outpatient laboratory results, surveillance data, emergency room, hospital discharge diagnosis, long term facility diagnosis, date of death.	ICD10-CM for diagnosis. ATC for medicines. ATC and antigen for vaccines. ICD10-PCS for procedures.	6 months
FISABIO	VID	Spain	5.0 million	Primary care record, outpatient specialist record, outpatient laboratory results, surveillance data, emergency room visits, hospital discharge diagnosis, in-hospital prescribing, pharmacy dispensing outpatient, in-hospital prescription/dispensing, long term facility diagnosis, date and reasons of death.	ICD10-CM and ICD9-CM for diagnosis and procedures. ATC for medicines. Disease + text information for vaccines.	Instantaneous for outpatient data, every 6 months for inpatient data.
IACS	EPICHRON	Spain	1.3 million	Primary care record, outpatient laboratory results, emergency room visits, hospital discharge diagnosis, pharmacy dispensing outpatient, date of death.	ICPC, ICD9-CM and ICD10-CM for diagnosis. ATC for medicines. ICD10-CM for procedures.	3-6 months
SOSETE	PEDIANET	Italy	50.000	Primary care record, outpatient specialist diagnosis, surveillance data, emergency room visits, hospital discharge diagnosis, in-hospital prescribing (free text), outpatient prescription, date of death, reasons of death.	ICD9-CM and free text for diagnosis. ATC and free text for medicines. ATC and free text for vaccines. ICD9-CM and free text for procedures.	6 months
Utrecht University	CPRD-Aurum	UK	16 million	Primary care diagnoses, prescriptions, lab tests, hospital admissions and procedures CPRD death date	Read/Snomed for primary care diagnoses, BNF/product codes, but we have linked to ATC. ICD-10 for hospital diagnoses, OPCS for hospital procedures	New release scheme of primary care is quarterly. Lag time of hospital data (HES) difficult to say, currently available until 03/2021, used to annually updated.
Aarhus University	Danish registries	Denmark	5.9 million	Outpatient specialist diagnosis, laboratory results (hospital-based), emergency room visits, hospital discharge diagnosis, outpatient pharmacy dispensing, in-hospital prescription/dispensing, date of death, reasons of death (2 years lag time).	ICD-10 Danish modification for diagnosis. ATC and hospital internal codes for medicines. Internal code for vaccines. NOMESCO for procedures.	Depends on data source.

DAP	Data source	Country	Population size	Data banks available for this study	Vocabularies	Data update frequency
University of Eastern Finland	Finnish registries	Finland	2.9 million (50% random sample of total population)	Primary care record (with some restrictions), outpatient specialist diagnosis, outpatient laboratory results, surveillance data, emergency room visits, hospital discharge diagnosis, in-hospital laboratory results, outpatient pharmacy dispensing, long term facility diagnoses, date and reasons of death.	ICD-10 for diagnosis. ATC for medicines. ATC and free text for vaccines. NOMESCO for procedures.	Depending on data source, from 1 month to 1 year.
BPE & ADERA	SNDS	France	6.7 million (10% sample of the total population)	Outpatient healthcare (no results, no indication), pharmacy dispensing (quantity, dosage, name, no indication), public/private hospital stays with discharge diagnosis (no results), public hospital visits (no results, no indication), emergency room visits (with diagnosis if > 1 day, without if <=1 day), in-hospital dispensing/prescription (only for out-of-DRG drugs), date of death, reason of death.	ICD-10 for diagnosis. ATC for medicines and vaccines. CCAM for procedures, NABM for lab tests, LPP for (para)medical devices	1 year.
University of Oslo	Norwegian registries	Norway	5.3 million	Primary care record, outpatient specialist diagnosis, surveillance data (infectious diseases), emergency room visits, hospital discharge diagnosis, outpatient pharmacy dispensing, date and reasons of death.	ICD10, ICPC, ATC for medicines	No data update in this project. *The ETL'd data instance includes data on all residents in Norway between 1.1.2017- 31.12.2022.*

5.8 Study size

The study will include all eligible subjects in the data sources.

5.9 Data analysis:

For assessing the data quality, INSIGHT data quality level 1 and 2 checks are required as well as the running of tailored scripts on the incidence and prevalence of selected events, the prevalence of co-variates and the coverage of selected vaccines. Vaccine coverage estimations will be compared with available benchmarks if possible (e.g., WHO coverage estimates for routine vaccination, or ECDC COVID-19 tracker). The "The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework" (SPIFD) tool from Gatto et al. will be used to assess and summarize the results.

6 AMENDMENTS AND UPDATES

Number	Date	Section of study protocol	Amendment or update	Reason
N/A				

7 MILESTONES

Milestones and deliverables	Planned date
Contract signature	15 Feb 2024
Start of project	15 Feb 2024
D1 Study plan*	11 March 2024
D3 Study report (planned)	19 April 2024
D3 Study report (second realieased)	29 May 2024

8 RATIONALE AND BACKGROUND

The COVID-19 pandemic emphasised the public health need for comprehensive and rapid post-authorisation vaccine safety surveillance. An increasing number of vaccine products are based on novel technologies, for which safety experience is limited to pre-authorisation clinical trials until the recent COVID-19 pandemic. While new safety concerns are expected to arise with these novel vaccines, continuous monitoring and evaluation throughout the entire lifecycle remains necessary for authorized vaccines (1,2). To this aim, networks of real-world data sources that are fit-for-purpose and readily accessible are essential.

In May 2022, the European Medicines Agency (EMA) and the European Centre for Disease Prevention and Control (ECDC) established the Vaccine Monitoring Platform (VMP), in accordance with the regulation on EMA's reinforced role and ECDC's extended mandate. The VMP aims to generate real-world evidence (RWE) on the safety and effectiveness of vaccines in the European Union (EU) and the European Economic Area (EEA).

VAC4EU and the EU PE&PV research network (EU PE&PV) showed they can provide a framework for post-authorisation safety and effectiveness evaluation that can be leveraged in case of a new public health emergency or a safety concern occurring with a novel, or a more characterised vaccine.

9 RESEARCH QUESTION AND OBJECTIVES

9.1 Goal

The overarching goal of this project is to create a framework for the post-authorisation safety monitoring and evaluation of vaccines in Europe that can conduct near real-time studies on new or existing vaccines.

9.2 Objectives

The aim of this study is to provide and describe a network of real-world data sources for the evaluation of vaccine safety signals in Europe, and to assess its fitness-for-purpose through the following specific objectives.

Objective a. To assess the data quality for the purpose of conducting safety studies in the created network, including description of data source population, capture of routine vaccinations, and selected outcomes.

Objective b. To assess whether data are fit-for-purpose for conducting future safety studies on specific vaccines and selected outcomes in a near real-time monitoring manner.

10 RESEARCH METHODS

10.1 Study Design

A multi-database population-based cohort study design will be conducted from January 1st, 2017, till the last data availability.

To provide and describe a network of real-world data sources for the evaluation of vaccine safety signals, and to assess their fitness-for-purpose, we propose a list of 10 data sources (table 3) in seven countries based on available data banks, budget, and resources. The work will start with accessing data from 10 different electronic health record data sources that have proven to be able to convert data (n=9) into the ConcePTION common data model (CDM) or are willing to do this (n=1).

Table 3. Data access providers in the EU PE&PV and VAC4EU networks

Data access provider	Data source	CDM	Country
UU	CPRD	ConcePTION	UK
AEMPS	BIFAP	ConcePTION	Spain
IDIAPJGol	SIDIAP	ConcePTION	Spain
FISABIO	VID	ConcePTION	Spain

Data access provider	Data source	CDM	Country
IACS	EPICHRON	ConcePTION	Spain
Aarhus University	Danish registries	ConcePTION	Denmark
University of Eastern Finland	Finnish registers	ConcePTION ongoing	Finland
University of Oslo	Norwegian registers	ConcePTION	Norway
SOSETE	PEDIANET	ConcePTION	Italy
BPE	SNDS	ConcePTION	France

Participating data access providers (DAPs) will go through the following sequence of actions to create a quality checked data instance:

1. Review Protocol.
2. Update or create the Extraction Transformation Load (ETL) design document of local data to ConcePTION CDM for the data instance.
3. ETL their local data extraction into the ConcePTION Common Data Model (CDM).
4. Run off the shelf INSIGHT Level 1-3 level checks (see https://github.com/UMC-Utrecht-RWE) on the data instances of the ConcePTION CDM. Quality control of ConcePTION CDM (INSIGHT) allows to check for structural coherence and data completeness (level 1), semantic coherence and data characterization at value level (level 1b), relational data characterization (level 2), and data characterization at content level (level 3). When the specific data instance has already undergone quality checks, this step is not needed.
5. Review the INSIGHT level check 1, 2, 3 outputs with the quality assessor.
6. Run tailored study R-scripts to estimate disease incidence/prevalence rates, and coverage of vaccines.
7. Review and interpret the results.

10.2 Study Setting

This study will be conducted using electronic health record data from 10 data sources in 7 countries in Europe. The source population comprises 68,300,000 persons (table 4).

Table 4. Data access providers (DAP) and data characteristics

DAP	Data source	Country	Population size	Data banks available for this study	Vocabularies
AEMPS	BIFAP	Spain	17.0 million	Primary care record, hospital discharge diagnosis, community pharmacy dispensing, date of death.	ICPC2, ICD9 and SNOMED for diagnosis. ATC for medicines.
IDIAP J Gol	SIDIAP	Spain	5.8 million	Primary care record, outpatient specialist record, outpatient laboratory results, surveillance data, emergency room, hospital discharge diagnosis, long term facility diagnosis, date of death.	ICD10-CM for diagnosis. ATC for medicines. ATC and antigen for vaccines. ICD10-PCS for procedures.
FISABIO	VID	Spain	5.0 million	Primary care record, outpatient specialist record, outpatient laboratory results, surveillance data, emergency room visits, hospital discharge diagnosis, in-hospital prescribing, pharmacy dispensing outpatient, in-hospital prescription/dispensing, long term facility diagnosis, date and reasons of death.	ICD10-CM and ICD9-CM for diagnosis and procedures. ATC for medicines. Disease + text information for vaccines.
IACS	EPICHRON	Spain	1.3 million	Primary care record, outpatient laboratory results, emergency room visits, hospital discharge diagnosis, pharmacy dispensing outpatient, date of death.	ICPC, ICD9-CM and ICD10-CM for diagnosis. ATC for medicines. ICD10-CM for procedures.
SOSETE	PEDIANET	Italy	50.000	Primary care record, outpatient specialist diagnosis, surveillance data, emergency room visits, hospital discharge diagnosis, in-hospital prescribing (free text), outpatient prescription, date of death, reasons of death.	ICD9-CM and free text for diagnosis. ATC and free text for medicines. ATC and free text for vaccines. ICD9-CM and free text for procedures.
Utrecht University	CPRD-Aurum	UK	16 million	Primary care diagnoses, prescriptions, lab tests, hospital admissions and procedures CPRD death date	Read/Snomed for primary care diagnoses, BNF/product codes, but we have linked to ATC. ICD-10 for hospital diagnoses, OPCS for hospital procedures
Aarhus University	Danish registries	Denmark	5.9 million	Outpatient specialist diagnosis, laboratory results (hospital-based), emergency room visits, hospital discharge diagnosis, outpatient pharmacy dispensing, in-hospital prescription/dispensing, date of death, reasons of death (2 years lag time).	ICD-10 Danish modification for diagnosis. ATC and hospital internal codes for medicines. Internal code for vaccines. NOMESCO for procedures.
University of Eastern Finland	Finnish registries	Finland	2.9 million (50% random sample of total population)	Primary care record (with some restrictions), outpatient specialist diagnosis, outpatient laboratory results, surveillance data, emergency room visits, hospital discharge diagnosis, in-hospital laboratory results, outpatient pharmacy dispensing, long term facility diagnoses, date and reasons of death.	ICD-10 for diagnosis. ATC for medicines. ATC and free text for vaccines. NOMESCO for procedures.
BPE & ADERA	SNDS	France	6.7 million (10% sample of the total population)	Outpatient healthcare (no results, no indication), pharmacy dispensing (quantity, dosage, name, no indication), public/private hospital stays with discharge diagnosis (no	ICD-10 for diagnosis. ATC for medicines and vaccines. CCAM for procedures,

DAP	Data source	Country	Population size	Data banks available for this study	Vocabularies
				results), public hospital visits (no results, no indication), emergency room visits (with diagnosis if > 1 day, without if <=1 day), in-hospital dispensing/prescription (only for out-of-DRG drugs), date of death, reason of death.	NABM for lab tests, LPP for (para)medical devices
University of Oslo	Norwegian registries	Norway	5.3 million	Primary care record, outpatient specialist diagnosis, surveillance data (infectious diseases), emergency room visits, hospital discharge diagnosis, outpatient pharmacy dispensing, date and reasons of death.	ICD10, ICPC, ATC for medicines

10.3 Source and study population

The source population comprises all persons in the data sources. Table 4 gives an overview of the participating data sources and their characteristics. From the source population, we will select the study cohort.

10.3.1 Inclusion criteria, exclusion criteria and follow-up

Persons will be included in the dynamic study population when they have:

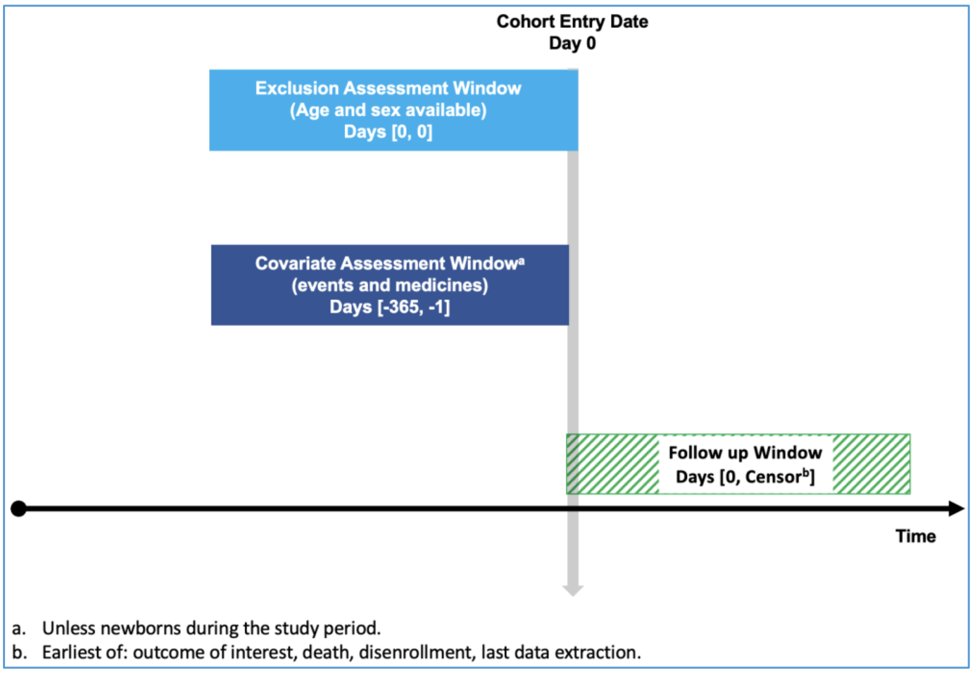
- Information on age and sex available.
- At least one day of follow in the study period (1/1/2017- latest availability).
- At least one year lookback history, with the exception of newborns.

Follow-up will start on the latest of the following dates:

- Study start date.
- Date at which they have time, except for newborns during the study period who will be included upon birth or when born in the year before the study start.

Follow-up finished at the earliest of the following dates: death, disenrollment, recommended end date by the DAP. The recommended end date is the date provided by the DAP until which they think the data of different databanks is complete (see figure 1 for study diagram).

Figure 1. Study diagram



a. Unless newborns during the study period.
b. Earliest of: outcome of interest, death, disenrollment, last data extraction.

10.4 Variables

10.4.1 Exposure Assessment

The Data Access Partners (DAPs) will convert their local data on vaccinations into the VACCINES table of the CDM. The exposure of interest in this project are all types of vaccines available during the study period and used in the participating study population in the participating countries, as part of the routine childhood, adulthood, or elderly vaccination program (see table 5).

Table 5. Vaccines of interest to assess feasibility.

Indicator	Cohorts
Measles-contaning vaccine, dose 1	Childhood from birth to 24 months
Measles-contaning vaccine, dose 2	Childhood from birth to 24 months
Diphteria tetanus toxiod and pertussis, dose 1	Childhood from birth to 12 months
Diphteria tetanus toxiod and pertussis, dose 2	Childhood from birth to 12 months
Diphteria tetanus toxiod and pertussis, dose 3	Childhood from birth to 24 months
Haemophilus influenzae type B, dose 1	Childhood from birth to 12 months
Haemophilus influenzae type B, dose 2	Childhood from birth to 12 months
Haemophilus influenzae type B, dose 3	Childhood from birth to 24 months
Hepatatis B, dose 1	Childhood from birth to 12 months
Hepatatis B, dose 2	Childhood from birth to 12 months
Hepatitis B, dose 3	Childhood from birth to 24 months
Polio, dose 1	Childhood from birth to 12 months
Polio, dose 2	Childhood from birth to 12 months
Polio, dose 3	Childhood from birth to 24 months
Pneumococcal conjugate vaccines, dose 1	Childhood from birth to 12 months
Pneumococcal conjugate vaccines, dose 2	Childhood from birth to 24 months
Varicella	Childhood from birth to 24 months
Bacille Calmette-Guérin (BCG) vaccine, dose 1	Childhood from birth to 24 months
Human papillomavirus vaccine, dose 1	Adolescents, 9 to 15-year-old, stratified per gender
Human papillomavirus vaccine, dose 2	Adolescents, 9 to 15-year-old, stratified per gender
Rotavirus, dose 1	Childhood from birth to 12 months
Rotavirus, dose 2	Childhood from birth to 12 months
Meningococcal vaccine, dose 1	Childhood from birth to 12 months
Meningococcal vaccine, dose 2	Childhood from birth to 15 months
Influenza vaccine	Seasonal cohorts, all ages, from 1 September to 30 April
COVID Vaccines, dose 1	Cohort entering 1 December 2020, stratified per age band at end of follow-up
COVID Vaccines, dose 2	Cohort entering 1 December 2020, stratified per age band at end of follow-up
COVID Vaccines, dose 3	Cohort entering 1 December 2020, stratified per age band at end of follow-up
COVID Vaccines, dose 4	Cohort entering 1 December 2020, stratified per age band at end of follow-up
COVID Vaccines, dose 5	Cohort entering 1 December 2020, stratified per age band at end of follow-up
COVID Vaccines, dose 6	Cohort entering 1 December 2020, stratified per age band at end of follow-up

Operationalization

ATC codes or the vaccine type (vxtype) will be utilized by DAPs to store records of vaccination in their CDM instances during the ETL phase. The vaccine type is the only case when a semantic mapping is adopted during conversion to the ConcePTION CDM (3). The investigators will retrieve from the level 1B checks the values utilised by the DAPs and map them to a three-digit code developed in the ADVANCE project (4). It maps all the antigens included in a vaccine to a code of three-digits (letters), and the vaccine is represented by the hyphen-separated alphabetic sequence of such three-letters codes. For example, all ATC codes of vaccine types associated to a trivalent diphtheria-pertussis-tetanus vaccine will be mapped to DIP-PER-TET.

10.4.2 Study Outcomes

Table 6 shows the outcomes that were selected in collaboration with EMA during the planning meeting and comprised: 1) AESI list for COVID-19 vaccines (ACCESS, SPEAC[1,2]), 2) AESI that might occur with vaccines in general, and 3) chronic immune-mediated events.

Table 6. List of selected events.

N	Name of the event
1.	Microangiopathy (MA)
2.	Acute coronary artery disease (CAD)
3.	Arrhythmia
4.	Myocarditis
5.	Pericarditis
6.	Venous thromboembolism (VTE)
7.	Arterial thrombosis (AMI /Ischemic stroke)
8.	TTS (VTE, arterial thrombosis, or CVST with thrombocytopenia in 10 days)
9.	Pulmonary embolism
10.	Haemorrhagic stroke (PE)
11.	Disseminated intravascular coagulation (DIC)
12.	Cerebral venous sinus thrombosis (CVST)
13.	Generalised convulsion
14.	Guillain Barré Syndrome (GBS)
15.	Diabetes (type 1)
16.	Single organ cutaneous vasculitis (SOCV)
17.	Erythema multiforme (EM)
18.	Meningoencephalitis
19.	Acute disseminated encephalomyelitis (ADEM)
20.	Narcolepsy
21.	Thrombocytopenia (TP)
22.	Transverse myelitis
23.	Bells’ palsy
24.	Kawasaki's disease (KD)
25.	Pancreatitis
26.	Rhabdomyolysis (RML)
27.	Severe cutaneous adverse reactions to drugs (SCARs)
28.	Sensorineural hearing loss (SNHL)
29.	Graves' disease (GD)
30.	Hashimoto's thyroiditis (HT)
31.	Auto-immune hepatitis (AIH)
32.	Polyarteritis nodosa (PAN)

[1] https://zenodo.org/communities/vac4eu/records?q=ACCESS&l=list&p=1&s=10&sort=bestmatch

[2] https://zenodo.org/communities/speac_project/records?q=&l=list&p=1&s=10&sort=newest

N	Name of the event
33.	Rheumatoid arthritis (RA)
34.	Psoriatic arthropathies (PsA)
35.	Systemic lupus erythematosus (SLE)
36.	Idiopathic thrombocytopenic purpura (ITP)
37.	Erythema nodosum (EN)
38.	Multiple sclerosis
39	Ulcerative colitis (UC)

Event definition forms and code lists including ICD-9, ICD-10, SNOMED, and ICPC codes will be used if available or generated using the following standard VAC4EU process. Event definition forms systematically capture the following items as a living document that will be closed and published upon study's end.

- Purpose of the event: covariate or outcome
- Version
- Document history
- Objective
- Clinical definition
- Synonyms/lay terms (for text mining purposes)
- Laboratory tests specific for diagnosing events
- Diagnostic tests specific for diagnosing events
- Drugs that are used to treat events
- Procedures used to treat events
- Setting where a condition is diagnosed (hospital, outpatient, GP)
- Literature review of diagnosis codes or algorithms used in other papers (health outcomes of interest)
- Code list used for study
- Algorithm proposal
- References

Code lists

Code lists to identify outcomes will be created using the VAC4EU Code Mapper tool (5), which maps concepts across vocabularies based on the Unified Medical Language System. Study variables are named in a standard VAC4EU hierarchical fashion based on the body system. The output of the Code Mapper is an Excel or CSV list. Each code is subsequently tagged as narrow or possible by two medical reviewers from the VAC4EU code list taskforce based on standard VAC4EU work instructions. Comments are consolidated in the VAC4EU code list task force.

The code lists are subsequently compiled in a CSV file through a standard R code which:

- Checks for ranges of codes in the Code Mapper outputs, and replacement with unique parent codes.
- Checks for odd characters in codes.
- Rounding of SNOMED codes.

10.4.3 Covariate Definition

Table 7 describes the list of selected covariates. Covariates were selected based on risk factors of the events of interest, based on a systematic literature review for the events, if they existed Code lists were created and tagged based on the same VAC4EU process as described under outcomes but now specific for covariates. Medication use may also be used as a proxy for comorbidities. Covariates will be assessed in the cohort within a lookback of 365 days for diagnoses codes and for medicines.

Table 7. List of study covariates and the CDM data that each covariate is based on

Covariate	Source ConcePTION CDM tables
Age	From PERSONS table
Gender	From PERSONS table
Race/ethnicity	From PERSONS table (if available)
Number of GP visits	From VISIT_OCCURRENCE table
Number of hospitalizations	From VISIT_OCCURRENCE table
Transplantation	From multiple tables: EVENTS, MEDICINES, SURVEY_OBSERVATIONS
Immunocompromised status	Algorithm from multiple tables: EVENTS - Inflammatory bowel disease - Diabetes type 1 - Gout - AIDS - Sjogren syndrome - Systemic Lupus Erythematosus - Transplant recipient - Psoriasis - Psoriatic arthropathy - Rheumatoid arthritis - Spondylarthritis - Multiple sclerosis - Hematological cancer - Multiple immunodeficiencies MEDICINES - Immunosuppressants
Pregnancy	From multiple tables (data source specific) using ConcePTION pregnancy algorithm (6)
Hypertension	EVENTS
Lipid abnormalities	EVENTS and MEDICINES
Malignancies	EVENTS and MEDICINES
HIV	EVENTS and MEDICINES
Decreased renal function	EVENTS
Cardiocerebrovascular disease	EVENTS and MEDICINES
Heart failure	EVENTS
Diabetes type II	EVENTS and MEDICINES
Valvular heart disease	EVENTS
Inflammatory bowel disease	EVENTS
Coronary artery disease	EVENTS
Myocardial infarction	EVENTS
Arrhythmia	EVENTS
VTE	EVENTS
Infection	EVENTS and MEDICINES
Liver disease	EVENTS
Alcohol abuse	EVENTS
Sepsis	EVENTS
Chronic renal disease	EVENTS
Dementia	EVENTS
Respiratory infections	EVENTS
Herpes simplex	EVENTS
Influenza	EVENTS
Sleep disorders	EVENTS
Mental health diseases	EVENTS and MEDICINES

Covariate	Source ConcePTION CDM tables
Preeclampsia	EVENTS
Hepatitis C	EVENTS
Rheumatoid arthritis	EVENTS
SLE	EVENTS
Dermatomyositis	EVENTS
Sjogren's syndrome EVENTS Gallstones	EVENTS
Sickle cell disease	EVENTS and MEDICINES
Myasthenia gravis	EVENTS
Pernicious anemia	EVENTS
Autoimmune hepatitis	EVENTS
Celiac disease	EVENTS
Hepatitis B	EVENTS
Psoriasis	EVENTS
Gout	EVENTS
Crohn's disease	EVENTS
Ulcerative colitis	EVENTS
Atopic dermatitis	EVENTS
Immune thrombocytopenia	EVENTS
Nonalcoholic fatty liver	EVENTS
Obesity	EVENTS and MEDICINES
Dermatomyositis	EVENTS

10.5 Data Sources

This study will use data from secondary electronic health record databases that are population-based. The characteristics of each of the participating DAPs are summarized below and in table 8.

Table 2. Data provider and data sources:

Country	Data Source	Data Access Provider	Estimated source population size	Start and end date of data instance*
Spain (ES)	BIFAP	BIFAP	17 million	1.1.2018-30.4.2022
Spain (ES)	SIDIAP	IDIAP JGol	5.8 million	1.1.2017-30.06.2023
Spain (ES)	VID	FISABIO	5.0 million	1.1.2018- 31.12.2022
Spain (ES)	EPICHRON	IACS	1.3 million	
Italy (IT)	PEDIANET	So.Se.Te	50.000	1.1.2011 (except hospitalizations, 1.1.2017) - 31.12.2022
Denmark (DK)	Danish national registries (DNR)	Aarhus University	5.9 million	1.1.2015- 31.12.2022
Norway (NO)	Norwegian national registers	University of Oslo	5.3 million	1.1.2017- 31.12.2022
United Kingdom (UK)	CPRD	Utrecht University	16 million	01.01.2017 - 31.12.2022
France (FR)	SNDS	BPE & ADERA	6.7 million (10% sample of the total population)	01.01.2017 – 31.12.2020
Finland (FI)	Finnish national registers	University of Eastern Finland	2.9 million (50% random sample of total population)	Not Available (n.a.)

*data instance is the subset of the data source that has been ETL'ed into the CDM at a certain point in time, this instance does not necessarily contain data from all databanks in the data source, but data required for one or more studies. Data quality checks are done for each data instance.

ES: BIFAP (SEVERAL REGIONS)

BIFAP (Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito público), a computerized database of medical records of primary care is a non-profit research project funded by the Spanish Agency for Medicines and Medical Devices (AEMPS). Information collected by PCPs includes administrative, socio-demographic, lifestyle, and other general data, clinical diagnosis and health problems, results of diagnostic procedures, interventions, and prescriptions/dispensations. Diagnoses are classified according to the International Classification of Primary Care (ICPC)-2, ICD-9 and SNOMEDCT system, and a variable proportion of clinical information is registered in "medical notes" in free text fields in the EMR. Additionally, information on hospital discharge diagnoses coded in ICD-10 terminology is linked to patients included in BIFAP for a subset of periods and regions participating in the database. All information on prescriptions of medicines by the PCP is incorporated and linked by the PCP to a health problem (episode of care), and information on the dispensation of medicines at pharmacies is extracted from the e-prescription system that is widely implemented in Spain.

The project started in 2001 and the current complete version of the database with information until December 2020 includes clinical information of 14,810 primary care practices (PCPs) and paediatricians. Nine participant autonomous regions send their data to BIFAP every year. BIFAP database currently includes anonymized clinical and prescription/dispensing data from around 20 million (17 active population) patients representing 92% of all patients of those regions participating in the database, and 32% of the Spanish population. Mean duration of follow-up in the database is 9 years. From several regions, hospitalization data can be linked, this subpopulation is called BIFAP-HOSP-PC.

ES: SIDIAP (CATALUNYA)

The Information System for Research in Primary Care (Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària (SIDIAP) in Catalonia, Spain, is a primary care database set up by the Institute of Research in Primary Care (Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina [IDIAP JGol]) and Catalan Institute of Health (Institut Català de la Salut). [ICS]). The database collects information from 278 primary health care centres and includes more than 5.8 million patients covered by the Catalan Institute of Health (approximately 78% of the Catalan population) and is highly representative of the Catalan population.

SIDIAP data comprise the clinical and referral events registered by primary care health professionals (i.e., GPs, paediatricians, and nurses) and administrative staff in electronic medical records, comprehensive demographic information, community pharmacy invoicing data, specialist referrals, and primary care laboratory test results. SIDIAP can also be linked to other data sources, such as the hospital discharge database, on a project-by-project basis. Health professionals gather this information using International Classification of Diseases, 10th Revision (ICD-10) codes, ATC codes, and structured forms designed for the collection of variables relevant to primary care clinical management, such as country of origin, sex, age, height, weight, body mass index, tobacco and alcohol use, blood pressure measurements, and blood/urine test results. In relation to vaccines, information on all routine childhood and adult immunisations is included in addition to the antigen and the number of administered doses.

SIDIAP is listed in the Catalogue of RWD sources and studies by EMA . SIDIAP was characterised in the IMI-ADVANCE project and considered fit for purpose for vaccine coverage, benefits, and risk assessment. An algorithm to identify pregnancies has been previously used within SIDIAP. The algorithm uses diagnosis codes recorded in primary healthcare records during pregnancy and information recorded in the sexual and reproductive healthcare registries, including LMP, gestational week, expected date of delivery, actual date of delivery or termination, and pregnancy outcomes. Approximately 50% to 60% of pregnant women in Catalonia are attended in the sexual and reproductive healthcare centres that contribute data to SIDIAP. Approximately 70% of infant records can be linked to maternal records and used for research. The protocol will be evaluated by the SIDIAP Scientific Committe and by the IDIAPJGol Ethics Committee, the approval can take up to 4 weeks. The timeframe for data availability after the approval by the two local Committees is one month.

ES: VID (VALENCIA)

The Valencia health system integrated database (VID) is a set of multiple, public, population-wide electronic databases for the Valencia Region, the fourth most populated Spanish region, with ≈5 million inhabitants and an annual birth cohort of 48000 new-borns, representing 10.7% of the Spanish population and around 1% of the European population. The VID provides exhaustive longitudinal information including sociodemographic and administrative data (sex, age, nationality, etc.), clinical (diagnoses, procedures, diagnostic tests, imaging, etc.), pharmaceutical (prescription, dispensation) and healthcare utilization data from hospital care, emergency departments, specialized care (including mental and obstetrics care), primary care and other public health services. It also includes a set of associated population databases and registries of significant care areas such as cancer, rare diseases, vaccines, congenital anomalies, microbiology and others, and public health databases from the population screening programmers. All electronic health systems in the VID use the ICD-9-CM and the ICD-10-CM. All the information in the VID databases can be linked at the individual level through a single personal identification code. The databases were initiated at different moments in time, but all in all the VID provides comprehensive individual-level data fed by all the databases from 2008 to date. Information on PCR test results as well as serological/antibody tests results for the whole population of the Valencia region is available and linkable from the Microbiological Surveillance Network (RedMIVA). The Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO) is Data Access Provider for Valencia Integrated Databases (VID).

ES: EPICHRON (ARAGON)

The EPICHRON database links sociodemographic and clinical anonymised information from 2010 to present for all the users of the public health system in Aragón (approximately 98% of the reference population). This database is built from the BIGAN platform, which integrates a technical infrastructure and a data lake gathering individual patient data from the regional health service information systems, including primary care, specialised care, hospitalisations, emergency department visits, drug prescriptions, image diagnosis, laboratory tests, diagnostics, vaccination, medical history, and demographics from the users of the public health system of Aragon, which comprises about 2 million individuals historic data and an active population of 1.3 million individuals.

IT: PEDIANET

Pedianet is a national population database that contains anonymous patient-level data of more than 500,000 children since 2004, corresponding to around 4% of the annual paediatric population who received healthcare from family paediatricians (FPs) in Italy who were part of the PEDIANET network. The network links FPs distributed throughout several Italian regions designated by the Italian NHS, including Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Veneto, Lazio, Marche, Toscana, Abruzzo, Campania, Sardegna, and Sicilia, and who use the same software (Junior Bit®) (Padova, Italy) in their professional practice. Only childredn in Friuli Venezia Giulia can be linked to the immunization registry.

According to the Italian NHS, each child is assigned to a FP, who is the primary referral for health-related matters. In Italy, there is a tax-funded public healthcare system with universal access, and patients do not incur direct costs related to primary care visits. The Pedianet database captures several types of patient-level information, including the reason for accessing healthcare, health status, demographic data, diagnosis and clinical symptoms (free text or ICD-9-CM codes), drugs (ATC codes), specialist appointments, diagnostic procedures, hospital or emergency room (ER) admissions, growth parameters, and clinical outcome data. Informed consent is required from children's parents to enter the data in the database. The data collected from the child's parents/tutors by paediatricians enters the dedicated cloud already encrypted and anonymised. Pedianet researchers do not know the process to anonymise the data and cannot know the owner of the data in any way.

DK: DANISH NATIONAL REGISTRIES (DNR OR DHR)

All Danish registries used in this study have a nationwide coverage and an almost 100% capture of contacts covering information on currently 5.9 million inhabitants plus historical information. Unambiguous person-level linkage across all data sources is possible via a unique identifier used in all Danish public records. Linked data from the following registries are available for the current project: the Danish Civil Registration System (identifier for linkage, age, sex, births, deaths, migrations); the Danish National Prescription Registry (outpatient dispensing in community pharmacies, no data on drugs administered in hospitals); the Danish National Health Service Register (GP contacts including vaccinations other than COVID-19); the Danish National Patient Registry (diagnoses and procedures from all hospital encounters); the Danish Vaccination Register (COVID-19 vaccinations only). Data are linked using a unique pseudonymized identifier on the servers of the Danish Health Data Authority (SDS). Individual-level data will be analysed by uploading and running of analytic scripts on the SDS servers and aggregate data that does not allow backtracking to individuals in accordance with the data regulation will be used for reporting. The Danish national registries are listed as a resource in the Catalogue of RWD sources and studies by EMA.

NO: NORWEGIAN NATIONAL LINKED REGISTERS AT UIO (NHR)

The core data that the University of Oslo (UiO) has access to are the health care administrative data banks of the entire Norwegian population, which amounts to approximately 5.3 million inhabitants. Norway has a universal public health care system, consisting of primary health care services and specialist healthcare services. Many population-based health registries were established in the 1960s, with use of unique personal identifiers facilitating linkage between registries. The mandatory national health registries were established to maintain national functions. They are used for health analysis, health statistics, improving the quality of healthcare, research, administration and emergency preparedness. The Norwegian data sources, in this project are the national, mandatory Norwegian Surveillance System for Communicable Diseases (MSIS), which will be linked to five national health registries, i.e. the Medical Birth Registry, the National Patient Register, Norway Control and Payment of Health Reimbursement, the Norwegian Immunisation Registry, and the National Prescription Registry. Information about all Norwegian National Registries can be found here: www.fhi.no/en/more/access-to-data/about-the-national-health-registries2/.
In this project, University of Oslo is Data Access Provider for Norwegian national registry data. Their current Norwegian health registry data will be used, capitalizing on the existing ETL's and quality checked data instance. In specific, UiO will contribute with ETL'd data on all residents in Norway between 1.1.2017- 31.12.2022, with historical data on these individuals back to 2010. Consequently, we will not be able to provide analysis as a near real-time analysis. Some ICD-10 codes are not at the 4-digit level.

UK: CPRD

The Clinical Practice Research Database (CPRD) from the UK collates the computerised medical records of GPs in the UK who act as the gatekeepers of health care and maintain patients' life-long electronic health records. Accordingly, GPs are responsible for primary health care and specialist referrals, and they also store information about specialist referrals and hospitalisations. General practitioners act as the first point of contact for any non-emergency health-related issues, which may then be managed within primary care and/or referred to secondary care, as necessary. Secondary care teams also provide information to GPs about their patients, including key diagnoses. The data recorded in the CPRD include demographic information, prescription details, clinical events, preventive care, specialist referrals, hospital admissions, and major outcomes, including death. Most of the data is coded using Read or SNOMED codes. Data validation with original records (specialist letters) is also available. The population in the data bank is generalisable to the UK population based on age, sex, socioeconomic class, and national geographic coverage CPRD Aurum versions is used. There are currently approximately 59 million individuals (acceptable for research purposes) -17 million of whom are active (ie, still alive and registered with the GP practice)- in over 2,000 primary care practices (https://cprd.com/Data). Data include demographics, all GP/health care professional consultations (e.g., phone calls, letters, e- mails, in surgery, at home), diagnoses and symptoms, laboratory test results, treatments (including all prescriptions), all data referrals to other care providers, hospital discharge summary (date and Read/SNOMED codes), hospital clinic summary, preventive treatment and immunisations, and death (date and cause). For a proportion of the CPRD panel practices (> 80%), the GPs have agreed to permit the CPRD to link at the patient level to HES data. Access to CPRD data will be provided by University Utrecht).

FRANCE: SNDS

The *Système National des Données de Santé* (SNDS) is the French nationwide healthcare database. It currently covers the overall French population (about 67 million persons) from birth (or immigration) to death (or emigration), even if a subject changes occupation or retires. Using a unique pseudonymized identifier, the SNDS merges all reimbursed outpatient claims from all French health care insurance schemes (SNIIRAM database), hospital-discharge summaries from French public and private hospitals (PMSI database), and the national death registry. SNDS data are available since 2006 and contains information on:

- General characteristics: gender, year of birth, area of residence, deprivation index, etc;
- Death: month, year and cause;
- Long-term disease registration associated with an ICD-10 diagnostic code;
- Outpatient reimbursed healthcare expenditures with dates and codes (but not the medical indication nor result): visits, medical procedures, nursing acts, physiotherapy, lab tests, dispensed drugs and medical devices, etc. For each expenditure, associated costs, prescriber and caregiver information (specialty, private/public practice) and the corresponding dates are provided;
- Inpatient details: primary, related and associated ICD-10 diagnostic codes resulting from hospital discharge summaries with the date and duration of the hospital stay, the performed medical procedures (but no results), lab tests (but no results) and the related costs. Drugs included in the diagnosis related group cost are not captured. However, expensive drugs (i.e., the one charged in addition to the group cost) are.

Outpatient data (SNIIRAM) are uploaded to the SNDS throughout the year. It is admitted that a lag of around 6 months is required to catch 90% of the dispensing. Inpatient data (PMSI) are uploaded in one

time, at the end of the following year. Hence, we consider that complete SNDS data of year Y are available in January of the year Y+2.
SNDS access is regulated: each study involving the human person with or without data extraction from the SNDS needs approval from the *Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé* (CESREES) in charge of assessing scientific quality of the project, and authorization from the *Commission Nationale de l'Informatique et des Libertés* (CNIL) which is the French data protection authority, and then an agreement with the SNDS data holder (CNAM) for data extraction.

FI: FINNISH NATIONAL REGISTERS

Finnish national data registers account for a total population of 5.4 million inhabitants. Main linkable data banks are: 1. *Hospital discharge register*: use of in- and outpatient services. Diagnoses for each admission are made by the attending physician. The register contains the following information on each hospital visit: dates, reason for hospital stay, specialty of the caring unit, date of operation, up to five operational codes (NOMESCO classification), where the patient was discharged to and assessment of need for assistance in activities of daily life. Since 2009, the data bank contains outpatient visits to specialised healthcare and since 2011 to primary healthcare. Laboratory and physiological measurements are available since 2015. 2. *Kanta electronic prescriptions:* all prescribed medicines purchased by an individual. Medicines used in hospitals are not included, but the register covers prescriptions written by hospital physicians and dispensed in community settings. Data on dispensing date, number of packages, tablets and defined daily dose (DDD) are available. Medicines are classified according to Anatomical Therapeutic Chemical (ATC)–classification system. 3. *Special reimbursement register:* entitlement to special reimbursement due to severe chronic diseases such as Alzheimer’s disease, diabetes, psychosis, epilepsy, asthma, c*hronic obstructive pulmonary disease and* several cardiovascular diseases. The diagnoses are based on explicit predefined criteria. 4. *Statistics Finland* is the statistical authority of Finland, producing the majority of official statistics and conducting the population census, which has solely been based on the register data since 1990. These censuses include indicators of socioeconomic position (e.g. education, occupational status and taxable income). The causes of death register are compiled from death certificate data containing underlying, direct, intervening, and contributing causes. Death certificates are issued by physicians and if an autopsy is required, by a medicolegal officer.

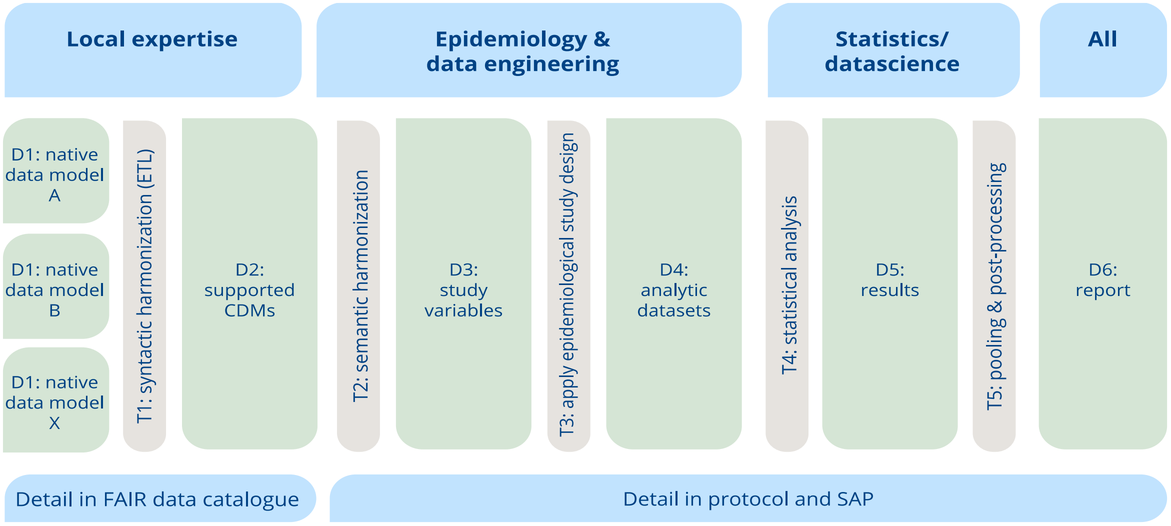
10.6 Study size

The study will include all subjects eligible in the data sources that could be utilized for the report.

10.7 Data Management

The study will be conducted in a distributed manner using the UMCU and VAC4EU tools, procedures, and pipeline. This pipeline can be viewed from a programming perspective (see figure 2) or tool perspective (figure 3). Figure 2 specifies the data sets (D) and transformation processes (T), programming follows this pipeline, with involvement of different types of experts.

Figure 2. Data Management from the data transformation perspective



D1: Original data can be in any native format

The RWD-RWE pipeline used by VAC4EU and EU PE&PV starts with data banks that are controlled by the DAP, these can be in any format. This stays local. The ETL design is shared in a searchable FAIR VAC4EU catalogue. The VAC4EU FAIR Molgenis data catalogue is a meta-data management tool designed to contain searchable meta-data describing organisations that can provide access to specific data sources.

T1: Syntactic harmonisation (ETL)

T1: Syntactic harmonisation is conducted through an extraction, transformation, and loading (ETL) process of native data into the requested CDM. To harmonise the structure of the data sets stored and maintained by each data partner, a shared syntactic foundation is used. The ETL process has various structured steps as described by Thurin et al. (3):

- DAPs are asked to share the data dictionaries of their data banks (selected tables and variable names/structure)
- Metadata (descriptive data about the data sources and data banks) & data dictionaries, are uploaded in the VAC4EU metadata catalogue
- DAPs make an ETL design
- The design is reviewed
- ETL is deployed

D2: Common data model

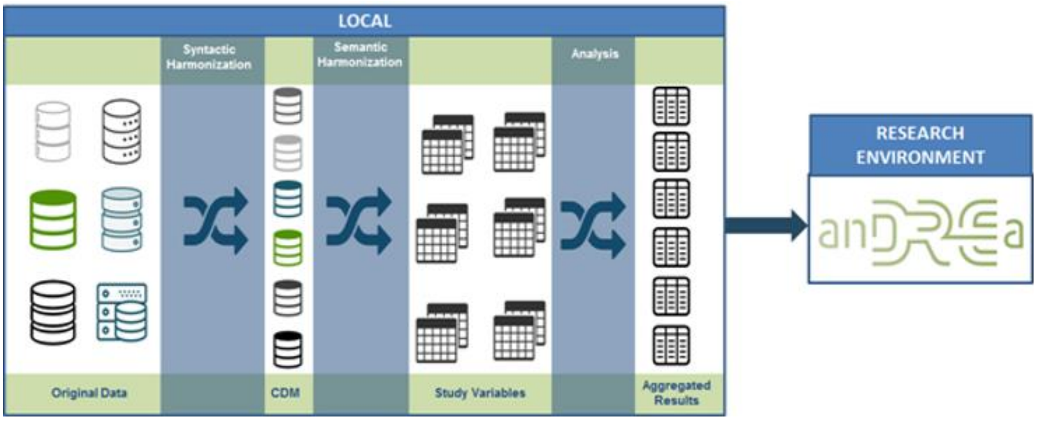
For this project we will use the ConcePTION CDM version v2.2. In the ConcePTION CDM the data is only syntactically harmonised, allowing for the data to remain in its original language (e.g., presence of different medical diagnostic systems such as ICD-9, ICD-10, SNOMED etc.).

T2: Semantic harmonisation

In this step we conduct time anchoring (observation periods, look back periods), clean the data such as the dose of vaccines, sort on record level, aggregate across multiple records, and combine concepts for implantation of algorithms, and rule-based creation of study variables. Based on the relevant diagnostic medical codes and keywords, as well as other relevant concepts (e.g., medications), one or more phenotype algorithms are constructed to operationalise the identification and measurement of each event or covariate.

In this phase of creation of study variables, the semantic mapping is conducted. This semantic mapping across different vocabularies is conducted as part of the R-study script using different functionalities. To reconcile differences between different terminologies and native data availability, machine readable code lists are used that comprise the terminologies that are used in the network (e.g. ICD-9, ICD10, SNOMED, ICPC and DAP specific adaptations). This is combined with the BRIDGE metadata file that defines risk windows, look back periods, and algorithms for each study variable (7).

Figure 3. Data management from a systems' and location's perspective



D3: Study variables

D3 datasets are interim data sets with information on study variables for each study participant. The unit may be a person, a medicine or vaccine record, or episode of time. The design of these datasets is described in codebooks.

T3: application of epidemiological design

In the T3 step, epidemiological designs are applied such as sampling, matching (on specific variables and/or propensity scores), and selection based on inclusion and exclusion criteria using the study variables in the D3 datasets. The designs will be implemented for the various study objectives using R-scripts, and these may use the existing functions (R-cran) or functions that have been developed in the VAC4EU community.

D4: Analytical data set

D4 is an analytical dataset, and multiple D4 data sets may be produced based on the objectives of the study. The format is described initially in a code book for communication between programmers and statisticians.

T4: Statistical analysis

This step in the data transformation pipeline will produce statistical estimates such as descriptives (counts, percentages), distributions (mean, percentiles), rates (prevalence, incidence), regression coefficients, or other relevant estimates.

As per VAC4EU policy, the analytical code will publicly release with an open-source licence once the final report is made publicly available. The code will be published in this GitHub repository https://github.com/VAC4EU/ROC18.

D5: Results

D5 is the set of estimands, tables or aggregate data that is transferred from the DAPs to the Digital Research Environment (DRE) (see figure 3). The DRE is made available through UMCU. The DRE is a cloud-based, globally available research environment where data are stored and organised securely and where researchers can collaborate. All researchers who need access to the DRE will be granted access to study-specific secure workspaces by UMCU. Access to the workspaces will be possible only after double authentication using an identification code and password together with the user's mobile phone for authentication. Downloading of files will be possible only after requesting and receiving permission from a workspace member with an "owner" role, who will be a UMCU team member.

10.8 Data analysis

The statistical analyses and methods for this study are descriptive and comprise the INSIGHT data quality checks and indicators, plus the additional scripts to describe vaccine coverage, prevalence of covariates and incidence of outcomes. All analyses will be conducted using R version R-4.03 or higher (Foundation for Statistical Computing, Vienna Austria) or STATA.

Descriptive analysis
The attrition table will provide an overview of the reasons for exclusions and the final study population for this study. Demographic characteristics of the study population as well as the prevalence of covariates will be provided as part of the descriptive analysis.

Vaccine coverage estimations
Vaccination coverage is the cumulative risk of being vaccinated with a particular antigen at a certain age. Vaccination schedules differ per country, but the WHO has benchmark data, related to certain antigens at certain ages. These data will be used as an external benchmark and is provided below in table 9. To estimate vaccine coverage in a dynamic population is challenging because of left and right censoring. In the IMI-ADVANCE project, methodological work was conducted to explore the best methods to estimate coverage in a dynamic population. We refer to Braeye et al. (4) for a simulation study that compared different methods. The authors concluded that the IPW/Cumulative distribution function methods were generally the least biased. Preference for a specific method should be based on the type of censoring and type of dependence between completeness of follow-up and vaccination.
The following outcome parameters will be estimated:

- Number of doses administered per vaccine during the study period.
- Vaccine coverage curves (cumulative incidence) by birth year for childhood vaccines, for HPV from 9 years of age. Estimates will be provided at certain ages (see table 9).

For the coverage estimation, we will follow the methodological approaches developed in the ADVANCE project (8). The PP.fu-method relies on the assumption that the age-specific coverage estimated from the part of the population in follow-up at any age in weeks represent the age-specific coverage of the population. The IPW-method relies on the assumption that the proportion of persons in follow-up receiving a vaccine during a certain age equals the proportion of persons not in follow-up receiving a vaccine. Since this assumption is likely violated for the influenza-vaccine study population, as in older age groups death is a common cause of loss to follow-up, the IPW-method was not applied to influenza-vaccine. A general summary of these assumptions is that with the PP.fu-method we assume that the observed coverage equalled the study population coverage, while with the IPW-method we estimate the coverage. The IPW-method accounts for both left and right censoring of vaccinations but can produce unstable estimates when weights are very small or large and bias can accumulate as the method sums over the weekly estimated number of vaccinations (8). In this study we will apply the IPW

method for early childhood vaccinations, since we will condition on start of birth for left censoring and address right censoring with IPW. For HPV vaccination, COVID-19 and influenza vaccine, we applied the PP_{FU} method due the potentially high proportion of incomplete follow-ups over a longer period.
For childhood vaccines (see table 9) coverage will be estimated by birth year over age in months using only those persons that were born and in follow-up during the study period. The number of persons in follow-up for at least one day during an age in months will be counted. Then, the number of persons who received a vaccination during that month will be counted as well, and those who had a registered vaccination during that age-month. A letter (A, B, C, D, E) will be assigned to every age-month of every person.

$$A_i = in\ follow{-}up\ (FU)\ during\ age\ i,\ vaccinated\ during\ age\ i$$

$$B_i = in\ FU\ during\ age\ i,\ vaccination\ recorded\ before\ age\ i$$

$$C_i = in\ FU\ during\ age\ i,\ no\ recorded\ vaccination\ before\ age\ i$$

$$D_i = Not\ in\ FU\ during\ age\ i,\ vaccination\ recorded\ before\ age\ i$$

$$E_i = Not\ in\ FU\ during\ age\ i,\ no\ recorded\ vaccination\ before\ age\ i$$

From the data aggregated by birth year, we will calculate and produce coverage curves applying the following methods:

Period Prevalence: Follow-Up (PP_{FU})

The PP_{FU} estimate for month i is the number of vaccinated persons in follow-up divided by the number of persons in follow-up during month i.

$$PPFU,i=Ai + BiAi + Bi + CiPPFU,i=Ai + BiAi + Bi + Ci$$

$PP_{FU,i}$ calculates the vaccination coverage by dividing the number of vaccinated persons in follow-up prior and during month i over the total number of persons in follow-up during month i.

Inverse probability weighted (IPW)

When using the IPW method, to address right censoring inverse probability weighting is applied

$$IPWi=\sum 0\rightarrow iVaccIPW,iNinFUIPWi=\sum 0\rightarrow iVaccIPW,iNinFU$$

IPW_i is the coverage estimated by the IPW_i-method during month i.

$$VaccIPW,i=VACCobserved,iFUproportion,iVaccIPW,i=VACCobserved,iFUproportion,i$$

$VaccIPW_{,i}$ is the estimated number of vaccinations during month i; $VACC\ observed,\ i$ is the observed number of vaccinations during month i; and $FU_{proportion,i}$ is the proportion of persons in follow-up during month i. $NinFU$ is the total number of persons in the birth cohort.

The age at which the coverage will be estimated, the method to estimate coverage, and the benchmark indicator per vaccine are provided in table 9.

Table 3. Vaccine, age of dose assessment (months), method and main reference indicator for the selected vaccines

Vaccine	Dose 1 (months)	Dose 2 (months)	Dose 3 (months)	Method to estimate coverage	WHO/ECDC assessment indicator
Measles-Mumps-Rubella	12	23		IPW	MCV1[3] (Measles-containing-vaccine first dose)
Diphtheria-Pertussis-Tetanus	12	12	23	IPW	DTP3[4] (Diphteria tetanus toxiod and pertussis, dose 3)
Hib	12	12	23	IPW	Hib3[5] (Haemophilus influenzae type B, dose 3)
Hepatitis B	12	12	24	IPW	HepB3[6] (Hepatitis B, dose 3)
Polio	12	12	23	IPW	Pol 3[7] (Polio, dose 3)
Pneumococcal	12	12		IPW	PCV[8] (Pneumococcal conjugate vaccines, dose 2)
Influenza	Yearly			PP_{FU}	n.a.
Varicella	24			IPW	n.a.
HPV	14 years, women	14 years, women		PP_{FU}	HPV[9] (HPV, dose 2, in women)
Rotavirus	12	12		IPW	RotaC[10] (Rotavirus, dose 2)
Tuberculosis	24			IPW	BCG[11] (Bacille Calmette-Guérin (BCG) vaccine, dose 1)

[3] Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/measles-containing-vaccinefirst-dose-(mcv1)-immunization-coverage-among-1-year-olds-(-)

[4] Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/diphtheria-tetanus-toxoidand-pertussis-(dtp3)-immunization-coverage-among-1-year-olds-(-)

[5] Hib (Hib3) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hib-hib3-immunization-coverage-among-1-year-olds-(-)

[6] Hepatitis B (HepB3) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hepatitisb-(hepb3)-immunization-coverage-among-1-year-olds-(-)

[7] Polio (Pol3) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/polio-(pol3)-immunization-coverage-among-1-year-olds-(-)

[8] Pneumococcal conjugate vaccines (PCV3) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/pneumoccocal-conjugatevaccines-(pcv3)-immunization-coverage-among-1-year-olds-(-)

[9] HPV immunization coverage estimates among primary target cohort (9-14 years old girls) (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/girls-aged-15-years-old-thatreceived-the-recommended-doses-of-hpv-vaccine

[10] Rotavirus vaccines completed dose (RotaC) immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/rotavirus-vaccinescompleted-dose-(rotac)-immunization-coverage-among-1-year-olds-(-)

[11] BCG immunization coverage among 1-year-olds (%) [Internet]. [cited 2024 May 8]. Available from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/bcg-immunizationcoverage-among-1-year-olds-(-)

Vaccine	Dose 1 (months)	Dose 2 (months)	Dose 3 (months)	Method to estimate coverage	WHO/ECDC assessment indicator
Meningococcal	15	15		IPW	n.a.
Coronavirus				PP_{FU}	Covid-19[12] (we will report the coverage in age bands)

To assess completeness of vaccination data in the specific data sources, we will compare the coverage estimates with the most recently published estimates from WHO. For COVID-19 vaccines, we will use data from the ECDC. A priori, we decide that if coverage estimates in the databases deviate more than 10% from reference data, this is of concern. It is of note that coverage reported to WHO may have varying origin, birth cohorts and years of assessment, as described in the foot page links. We have used the most recent data available, but changes during lockdown were not considered.

Covariate prevalence

Prevalence of diagnosis and medicines defined as covariates will be measured at the start of follow-up and prior to start of COVID-19 vaccine roll-out (within a lookback of 365 days prior to 01-12-2020).

Incidence & prevalence rates of events

Incidence rates will be calculated based on the first occurrence of an event and requiring absence of that event in the year prior (population at risk). Upon occurrence of the event, follow-up time will be censored.

Point prevalence estimates will be calculated at the start of each year: the numerator is the persons with the disease in the year prior, denominator is all persons present at the start of each calendar year.

One-year period prevalence estimates will be calculated, the numerator comprises all persons who at the start of the year either had the disease in the year prior or developed the disease during the calendar year. The denominator comprises the person-years of follow-up in that year as an estimate of the average number of patients in that year (we do not condition on being fully available to avoid immortal time bias).

All estimates will be age-standardized to the Eurostat population. Age-specific estimates will be calculated in the following categories: 0-1, 2-4, 5-11, 12-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+ years.

10.9 Quality control

Rigorous quality-control (QC) procedures will be used. Data transformation into the ConcePTION CDM will be conducted by each subcontracted research partner in its associated database, using the processes described in the following sections (see below), each of these steps is fully transparent and will be signed of/reviewed by local and central teams. Standard operating procedures or internal process guidance at each research centre will be used to guide the conduct of the study. These procedures include rules for secure and confidential data storage, backup, and recovery; methods to maintain and archive project documents; QC procedures for programming; standards for writing analysis plans; and requirements for scientific review by senior staff.

[12] Durán CE, Messina D, Gini R, Riefolo F, Aragón M, Belitser S, et al. Rapid Safety Assessment of SARS-CoV-2 Vaccines in EU Member States using Electronic Health Care Data Sources (COVID Vaccine Monitor-CVM study): Final Study Report for WP3 (electronic health record data). 2023 Aug 24;

Quality of protocol, SAP, reports and manuscript

All these products are and will be reviewed by the entire consortium with many domain experts, using version control on SharePoint.

Quality of data checks

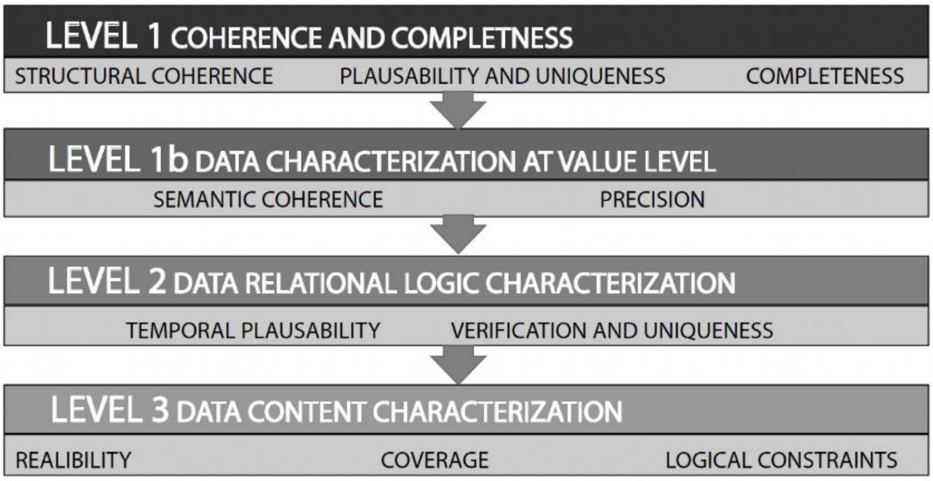
The INSIGHT data quality assessment R-tool for data converted in the ConcePTION CDM will allow for a detailed characterization of the data source instance that will be used for this study, including an overview of the anticipated availability and quality of exposure to selected vaccines of interest. INSIGHT is a public set of R tools that identifies potential data quality issues in ConcePTION CDM-standardized instances through the systematic execution and summary of over 588 configurable data quality assessments (9). All INSIGHT scripts are publicly available on https://github.com/UMC-Utrecht-RWE.

For the INSIGHT level 1-3 quality checks, detailed statistical analysis plans are available on public repositories:

- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1 (10)
- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level2 (11)
- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level3 (12)

Level 1 focuses on compliance with the ConcePTION CDM specifications and data completeness. Level 2 evaluates the temporal plausibility of events and the uniqueness of records. Level 3 provides an overview of distributions, outliers, and trends over time. The data quality assessments are run locally by the DAP and assessed centrally by a data quality revisor together with the DAP's representatives. INSIGHT is a tool that aligns with and operationalizes the five dimensions of the EMA data quality framework: reliability, extensiveness, coherence, timeliness, and relevance (figure 4). Data quality is the sum of several internal and external features of data. An important feature of the VAC4EU procedures is that each data instance (a version of the original data that is converted into CDM) will undergo through the quality assessment using the INSIGHT pipeline, prior to running an analysis study script, to ensure quality of data, since this may vary largely between different instances.

Figure 4. Hierarchy and dimensions of data quality assessment in the INSIGHT tools mapped to the EMA data quality framework

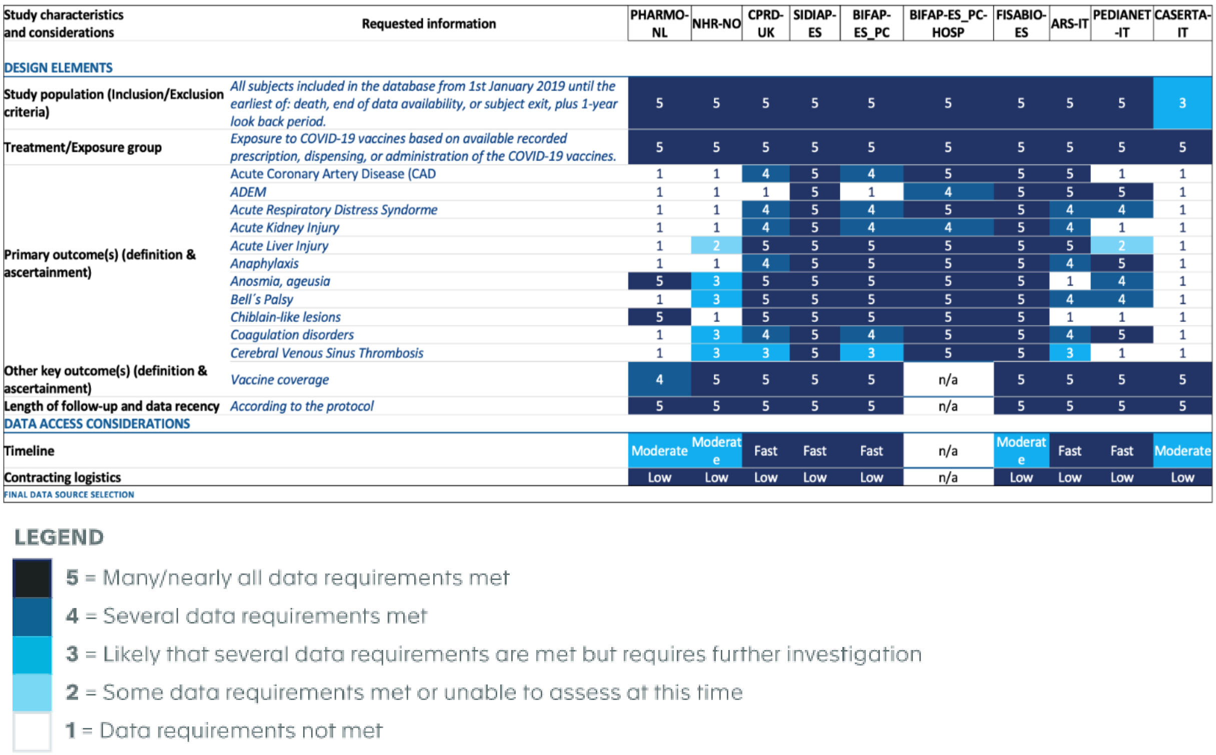

LEVEL 1 COHERENCE AND COMPLETNESS
STRUCTURAL COHERENCE
PLAUSABILITY AND UNIQUENESS
COMPLETENESS
LEVEL 1b DATA CHARACTERIZATION AT VALUE LEVEL
SEMANTIC COHERENCE
PRECISION
LEVEL 2 DATA RELATIONAL LOGIC CHARACTERIZATION
TEMPORAL PLAUSABILITY
VERIFICATION AND UNIQUENESS
LEVEL 3 DATA CONTENT CHARACTERIZATION
REALIBILITY
COVERAGE
LOGICAL CONSTRAINTS

To ensure that data quality indicators can be inspected, results will be presented in a HTML format for each level, facilitating their understanding and sharing. These reports will contain summary tables that allow for a concise representation of data quality indicators and graphs that provide a visual representation of trends and patterns. The INSIGHT data quality assessments are an iterative process for each data instance. Each level can be rerun until the required quality is attained or all constraints are noted.

After the quality indicators of the data instances and outputs of the analytical scripts have been generated, the fitness-for-purpose of the data instance will be performed by using, implementing, and adapting the feasibility framework *The Structured Process to Identify Fit-For-Purpose RWD sources* (SPIFD) proposed by Gatto et al.(13). The SPIFD assessment is a tool aimed at conducting feasibility assessment to determine whether a data source is fit-for-purpose for specific real-world effectiveness and safety study. The SPIFD will allow us to tabulate different evaluation items and therefore to score them (figure 5).

Eventually, the list of DAPs included in the network will contain an overview of data source characteristics, such as availability of information on vaccine exposure, rates of events of interest and relevant covariates, geographical coverage, healthcare settings covered, number of patients and median follow-up time, lag time, and other aspects informing their suitability for vaccine safety studies. Under this approach we will be able to produce a list of data sources that could be included in the network of real-world data sources for future vaccine safety studies.

Figure 5. Example of the application of the Structured Process to Identify Fit-For-Purpose RWD sources (SPIFD) to assess fitness for purpose of data instances for specific studies in the COVID-19 Vaccine Monitor study.



Study characteristics and considerations	Requested information	PHARMO-NL	NHR-NO	CPRD-UK	SIDIAP-ES	BIFAP-ES_PC	BIFAP-ES_PC-HOSP	FISABIO-ES	ARS-IT	PEDIANET-IT	CASERTA-IT
DESIGN ELEMENTS											
Study population (Inclusion/Exclusion criteria)	*All subjects included in the database from 1st January 2019 until the earliest of: death, end of data availability, or subject exit, plus 1-year look back period.*	5	5	5	5	5	5	5	5	5	3
Treatment/Exposure group	*Exposure to COVID-19 vaccines based on available recorded prescription, dispensing, or administration of the COVID-19 vaccines.*	5	5	5	5	5	5	5	5	5	5
Primary outcome(s) (definition & ascertainment)	Acute Coronary Artery Disease (CAD	1	1	4	5	4	5	5	5	1	1
	ADEM	1	1	1	5	1	4	5	5	5	1
	Acute Respiratory Distress Syndorme	1	1	4	5	4	5	5	4	4	1
	Acute Kidney Injury	1	1	4	5	4	4	5	4	1	1
	Acute Liver Injury	1	2	5	5	5	5	5	5	2	1
	Anaphylaxis	1	1	4	5	5	5	5	4	5	1
	Anosmia, ageusia	5	3	5	5	5	5	5	1	4	1
	Bell´s Palsy	1	3	5	5	5	5	5	4	4	1
	Chiblain-like lesions	5	1	5	5	5	5	5	1	1	1
	Coagulation disorders	1	3	4	5	4	5	5	4	5	1
	Cerebral Venous Sinus Thrombosis	1	3	3	5	3	5	5	3	1	1
Other key outcome(s) (definition & ascertainment)	*Vaccine coverage*	4	5	5	5	5	n/a	5	5	5	5
Length of follow-up and data recency	*According to the protocol*	5	5	5	5	5	n/a	5	5	5	5
DATA ACCESS CONSIDERATIONS											
Timeline		Moderate	Moderate	Fast	Fast	Fast	n/a	Moderate	Fast	Fast	Moderate
Contracting logistics		Low	Low	Low	Low	Low	n/a	Low	Low	Low	Low
FINAL DATA SOURCE SELECTION											

LEGEND

5 = Many/nearly all data requirements met

4 = Several data requirements met

3 = Likely that several data requirements are met but requires further investigation

2 = Some data requirements met or unable to assess at this time

1 = Data requirements not met

Figure 5 shows an example of criteria for population, exposure groups and outcomes, as well as data access considerations (lag time or contracting delays). We will produce a similar heat map for this project, with outcomes and vaccines specific for this study.

General approach to quality of R-coding

Data Management will follow standard operating procedures. R programs will be made available by UMCU (INSIGHT data quality checks) and ARS. ARS will create clear documentation (graphical and in Excel spreadsheet) of the data management steps, beginning with describing the required variables from the CDM and each of the subsequent transformation steps and intermittent data tables.

Coding conventions (process quality)

VAC4EU GitHub (and the underlying GitHub version control system) will be used to collaborate with multiple parties. At its core, GitHub tracks all changes and shows which, when, who and why changes were made. In the chain of events, any previous state can be recovered easily. Regarding proposed changes or potential bugs, GitHub provides a platform to discuss details. Using GitHub Actions, standard workflows will be defined and executed after a submitted change. An example is executing unit tests to ensure that scripts are correct.

The main coordinator of the GitHub is VAC4EU who creates a repository for the study and provides the *main* functions to be used in each study:

- A .readme file is initialized with relevant information about the scripts.
- For each study, a ‘branch’ is created in which scripts are tailored to the respective study.
- After each version update, the coordinator requests from all teams to incorporate the changes using ‘merge’. One responsible from each team is appointed and allowed access to the repository. In case the main scripts contain an error, the ‘issues’ functionality is used to report the bug. If possible, a bug fix can be proposed by creating a ‘pull request’. The ‘issues’ platform also provides a means to ask for further clarification regarding new versions.

We will use one set of standard conventions for all parties to facilitate collaboration and minimize bugs in scripts. Coding conventions are categorized into three parts:

- Notation (e.g., name scripts, functions, and objects).
- Syntax (e.g., spacing, braces, indentation).
- Documentation (e.g., writing comments, dividing code into sections).

Script names will be informative, where words are separated with an underscore or a hyphen. For scripts that are executed sequentially, the names are prefixed with numbers that indicate the order. For naming functions and objects, we suggest adopting the “snake style”, where words are separated with an underscore. For syntax rules, we will implement the tidy verse style guide found at https://style.tidyverse.org/. To facilitate implementing these rules, we will use the ‘formatR’ R package. This package automatically restyles R code to adhere to these rules. For documentation, comments will be provided that explain each part of the code. Each script file will start with a title, author, date, and version number. Comments are placed to describe functions and objects.

Scripts will be private during development and will be made public though the VAC4EU Github and Zenodo upon acceptance of the final report.

Standard/bespoke analyses script creation, testing and release

Study scripts connect and package functions using a structured design and follow the statistical analysis plan. Study scripts will be created in 4 steps:

1. Defining a map of the script, which includes specification of the folder structure, data model, graphical representation of the steps, use of functions, allocation of responsibilities and timelines, plus review schedules.
2. Programming of the code by a programmer plus statistician. Test with code profiler to monitor bottlenecks in the code.
3. Test script on one real data partner before making it available for all the DAPS.
4. Take the script into deployment.

For each update, these steps are repeated.

Quality of study conduct

The work in this proposal will be conducted according to the guidelines for Good Pharmacoepidemiology Practice (GPP) (International Society for Pharmacoepidemiology 2008) and according to the ENCePP code of conduct (European Medicines Agency 2018). All partners and principal investigators have experience in conducting pharmacovigilance/pharmacoepidemiological research and research is done by researchers trained in pharmacoepidemiology or pharmacovigilance. Utrecht University and University Medical Center Utrecht (data management), and Teamit (project management) are working according to a quality management system based on ISO 9001 principles and are certified.

All partners are ENCePP centers and the majority are members of the VAC4EU network, an ENCePP-listed network. Collectively these partners have registered more than 100 PASS studies on the EU PAS register, several with ENCePP seal.

The quality management system is system and process-oriented and based on continuous improvement. The system is based upon standard operating procedures implemented throughout the divisions with regular internal audits as well as external audits that lead to certification. The quality management system is based on national and international external quality requirements where available and pertinent, including the guidelines for Good Pharmacoepidemiological Practices, RECORD-PE, ENCePP Guide on Methodological Standards in Pharmacoepidemiology, Good Clinical Practice, and Good Clinical Data management. Practice as well as national and international guidelines and legislation concerning data-handling and privacy issues.

All deliverables will be reviewed by project partners, we have multiple expert organizations in public health and epidemiology who will contribute to methods and review.

10.10 Limitations of the research methods

10.10.1 Limitations related to the data sources.

This proposal uses available data sources, which capture different databanks. Not all data sources capture the same type of information which may impact on the ability to identify chronic diseases and their flares. Therefore, we conducted a pre-feasibility assessment and included data sources which mostly cover both general practices as well as outpatient and inpatient diagnosis, medicines and vaccine data. This data is not collected for research purposes and their fitness for purpose will be assessed using a series of analysis: first the data quality levels 1-3 through the 588 INSIGHT indicators, then the dedicated incidence and vaccine coverage estimates, which will be followed by a formal fit for purpose assessment based on the SPIFD assessment is a tool.

Capture of over-the-counter medications, potentially indicative of short-term disease status (e.g., painkillers, cough medicines, and fever reducers) may not be captured reliably.

10.10.2 Limitations in the methodology

For the descriptive analyses we do not foresee limitations, differences between data sources will be described transparently to allow for the choice of fit for purpose data sources for future studies. All the proposed methods have already been applied in the COVID-19 Vaccine Monitor (14) and in the IMI-ADVANCE (4,5) studies, which will allow for rapid adaptation and deployment.

10.11 Protection of human participants

This is a non-interventional study using secondary data collection and does not pose any risks for individuals. Each data source research partner will apply for an independent ethics committee review according to local regulations. Data protection and privacy regulations will be observed in collecting, forwarding, processing, and storing data from study participants.

10.11.1 Patient information

This study involves data that exists in an anonymized structured format and contains no patient personal information.

All parties will comply with all applicable laws, including laws regarding the implementation of organisational and technical measures to ensure the protection of patient personal data. Such measures will include omitting patient names or other directly identifiable data in any reports, publications, or other disclosures, except where required by applicable laws.

Patient personal data will be stored at DAPs in encrypted electronic form and will be password protected to ensure that only authorised study staff have access.

DAPs will implement appropriate technical and organisational measures to ensure that personal data can be recovered in the event of a disaster. In the event of a potential personal data breach, DAPs shall be responsible for determining whether a personal data breach has in fact occurred and, if so, providing breach notifications as required by law.

10.11.2 Patient consent

As this study does not involve data subject to privacy laws according to applicable legal requirements, obtaining informed consent from individuals is not required.

10.11.3 Ethical aspects

This study will adhere to the Guidelines for Good Pharmacoepidemiology Practices (GPP) and has been designed in line with the ENCePP Guide on Methodological Standards in Pharmacoepidemiology. The ENCePP Checklist for Study Protocols will be completed.

The study is a post-authorisation study of vaccine effectiveness and will comply with the definition of the non-interventional (observational) study referred to in the International Conference on Harmonisation tripartite guideline Pharmacovigilance Planning E2E and provided in the

EMA Guideline on Good Pharmacovigilance Practices (GVP) Module VIII: Post Authorisation Safety Studies and with the 2012 EU pharmacovigilance legislation, adopted June 19, 2012.
The study will be registered in the EU PAS Register before data collection commences.
The research team and study sponsor should adhere to the general principles of transparency and independence in the ENCePP Code of Conduct and the ADVANCE Code of Conduct.
The study will be conducted in accordance with legal and regulatory requirements, as well as with scientific purpose, value, and rigour, and will follow accepted research practices described in the Guidelines for Good Pharmacoepidemiology Practices (GPP) issued by the International Society for Pharmacoepidemiology (ISPE), and Good Epidemiological Practice guidelines issued by the International Epidemiological Association.

10.11.4Institutional review board (IRB)/Independent ethics committee (IEC)

Each DAP will be following the local country and data custodian requirements to apply for access to the data. All correspondence with the institutional review board or independent ethics committee and applicable documentation will be retained as part of the study materials.

11 MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS

As per EMA GVP Module VIII, the study and its protocol will be registered in the EU PAS Register prior to the start of data collection.

Results of analyses and interpretation will be delivered in report form.

Study results will be published following guidelines, including those for authorship, established by the ICMJE. When reporting the results of this study, the appropriate Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist and the RECORD-PE extension will be followed. Independent publication rights will be granted to the research team in line with Section VIII.B.5., Publication of study results, of the EMA *Guideline on Good Pharmacovigilance Practices (GVP) Module VIII: Post-Authorisation Safety Studies*.

Upon study completion and finalisation of the study report, the results of this study will be submitted for publication, preferably in a relevant peer-reviewed journal and posted in the EMA-HMA catalog.

Analytical programs will be posted in a public Github & Zenodo repository after acceptance of the report, reports will also be publicly posted on Zenodo (VAC4EU, EU PE&PV) and cross linked to the EU PAS register.

12 PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

For studies in which the research team uses only data from automated healthcare databases, according to the International Society for Pharmacoepidemiology Guidelines for GPP.

"Aggregate analysis of database studies can identify an unexpected increase in risk associated with a particular exposure. Such studies may be reportable as study reports, but typically do not require reporting of individual cases. Moreover, access to automated databases does not confer a special obligation to assess and/or report any individual events contained in the databases. Formal studies conducted using these databases should adhere to these guidelines."

For non-interventional study designs that are based on secondary use of data, such as studies based on medical chart reviews or electronic health records, systematic reviews, or meta-analyses, reporting of adverse events/adverse drug reactions is not required. Reports of adverse events/adverse drug reactions should only be summarized in the study report, where applicable.

According to the EMA Guideline on GVP, Module VI – Management and Reporting of Adverse Reactions to Medicinal Products,

"All adverse events/reactions collected as part of [non-interventional post-authorization studies with a design based on secondary use of data], the submission of suspected adverse reactions in the form of [individual case safety reports] is not required. All adverse events/reactions collected for the study should be recorded and summarized in the interim safety analysis and in the final study report."

Module VIII – Post-Authorization Safety Studies echoes this approach. Legislation in the EU further states that for certain study designs such as retrospective cohort studies, particularly those involving electronic health records, it may not be feasible to make a causality assessment at the individual case level.

13 REFERENCES

1. Salmon DA, Lambert PH, Nohynek HM, Gee J, Parashar UD, Tate JE, et al. Novel vaccine safety issues and areas that would benefit from further research. BMJ Glob Health [Internet]. 2021 May 1 [cited 2024 May 8];6(Suppl 2):e003814. Available from: https://gh.bmj.com/content/6/Suppl_2/e003814

2. Ricotta EE, Rid A, Cohen IG, Evans NG. Observational studies must be reformed before the next pandemic. Nature Medicine 2023 29:8 [Internet]. 2023 Jun 7 [cited 2024 May 8];29(8):1903–5. Available from: https://www.nature.com/articles/s41591-023-02375-8

3. Thurin NH, Pajouheshnia R, Roberto G, Dodd C, Hyeraci G, Bartolini C, et al. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. Clin Pharmacol Ther [Internet]. 2022 Jan 1 [cited 2024 Jun 18];111(1):321–31. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/cpt.2476

4. Braeye T, Bauchau V, Sturkenboom M, Emborg HD, García AL, Huerta C, et al. Estimation of vaccination coverage from electronic healthcare records; methods performance evaluation – A contribution of the ADVANCE-project. PLoS One [Internet]. 2019 Sep 1 [cited 2024 May 10];14(9):e0222296. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0222296

5. Becker BFH, Avillach P, Romio S, van Mulligen EM, Weibel D, Sturkenboom MCJM, et al. CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE

project. Pharmacoepidemiol Drug Saf [Internet]. 2017 Aug 1 [cited 2024 May 23];26(8):998–1005. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/pds.4245

6. ARS-toscana/ConcePTIONAlgorithmPregnancies: Repository of the script of the ConcePTION Algorithm for Pregnancies [Internet]. [cited 2024 May 10]. Available from: https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies

7. Royo AC, Elbers R, Weibel D, Hoxhaj V, Kurkcuoglu Z, Sturkenboom MC, et al. Real-World Evidence BRIDGE: a tool to connect protocol with code programming. medRxiv [Internet]. 2024 May 8 [cited 2024 May 10];2024.05.08.24306833. Available from: https://www.medrxiv.org/content/10.1101/2024.05.08.24306833v1

8. Braeye T, Emborg HD, Llorente-García A, Huerta C, Martín-Merino E, Duarte-Salles T, et al. Age-specific vaccination coverage estimates for influenza, human papillomavirus and measles containing vaccines from seven population-based healthcare databases from four EU countries – The ADVANCE project. Vaccine. 2020 Apr 3;38(16):3243–54.

9. Hoxhaj V, Navarro CLA, Riera-Arnau J, Elbers RJ, Alsina E, Dodd C, et al. INSIGHT: A Tool for Fit-for-Purpose Evaluation and Quality Assessment of Observational Data Sources for Real World Evidence on Medicine and Vaccine Safety. medRxiv [Internet]. 2023 Oct 30 [cited 2024 May 8];2023.10.30.23297753. Available from: https://www.medrxiv.org/content/10.1101/2023.10.30.23297753v1

10. UMC-Utrecht-RWE/INSIGHT-Level1: V1. [cited 2024 May 8]; Available from: https://zenodo.org/records/10035167

11. UMC-Utrecht-RWE/INSIGHT-Level2: V1. [cited 2024 May 8]; Available from: https://zenodo.org/records/10035169

12. UMC-Utrecht-RWE/INSIGHT-Level3: V1. [cited 2024 May 8]; Available from: https://zenodo.org/records/10035171

13. Gatto NM, Campbell UB, Rubinstein E, Jaksa A, Mattox P, Mo J, et al. The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. Clin Pharmacol Ther [Internet]. 2022 Jan 1 [cited 2024 May 8];111(1):122–34. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/cpt.2466

14. Durán CE, Messina D, Gini R, Riefolo F, Aragón M, Belitser S, et al. Rapid Safety Assessment of SARS-CoV-2 Vaccines in EU Member States using Electronic Health Care Data Sources (COVID Vaccine Monitor-CVM study): Final Study Report for WP3 (electronic health record data). 2023 Aug 24;