ACCESS

vACCine covid-19 monitoring readinESS

Protocol

Background rates of Adverse Events of Special Interest for monitoring COVID-19 vaccines

Version 1.1
September 21 2020

Title	Background rates of Adverse Events of Special Interest for monitoring of COVID-19 vaccines
Protocol version identifier	1.1
Date of last version of protocol	September 21, 2020
EU PAS register number	*EUPAS37273*
Active substance	*NA*
Medicinal product	*NA*
Product reference	*NA*
Procedure number	*NA*
Marketing authorisation holder(s)	*NA*
Research question and objectives	This study will generate background incidence rates of adverse event of special interest over the period 2017 to 2020 from 10 healthcare databases in 7 European countries
Country(-ies) of study	Participating electronic health care databases among ACCESS partners.
Authors	Caitlin Dodd, University Medical Center Utrecht, The Netherlands Corinne Willame, University Medical Center Utrecht, The Netherlands Miriam Sturkenboom (University Medical Center Utrecht)
Contributors	Helga Gardarsdottir (University Utrecht) Hedvig Nordeng (University Oslo) Vera Ehrenstein (Aarhus University) Consuelo Huerta (AEMPS) Eugene van Puijenbroek (LAREB) Helle Wallach Kildemoes (University Oslo) Patrick Souverein (University Utrecht) Lynn Meurs (LAREB) Rosa Gini (Agenzia Regionale di Sanita) Andrea Margulis (RTI-HS) Carlo Giaquinto (SOSETE) Reimar Thomsen (University Aarhus) Ainara Mira (FISABIO) Satu Siiskonen (Utrecht University) Estel Plana (RTI-HS) Lia Gutierrez (RTI-HS) Joan Fortuny (RTI-HS) Patricia Garcia Poza Nicolas Thurin (University Bordeaux) Boni Bolibar (IDIAP-Jordi Gol) Josine Kuipers (PHARMO) Rachel Weinrib (RTI-HS)

Contents

List of 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
CEPI	Coalition for Epidemic Preparedness Innovations
CI	Confidence interval
DAP	Data Access Provider
DNA	Desoxyribonucleic acid
DRE	Digital Research Environment
ECDC	European Centre for Disease Prevention and Control
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
IMI	Innovative Medicines Initiative
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
TOPFA	Termination of Pregnancy for Fetal Anomaly
VAC4EU	Vaccine monitoring Collaboration for Europe

1.Title

Feasibility analysis of a European infrastructure for COVID-19 vaccine monitoring: Background rates of Adverse Events of Special Interest

2. Marketing authorisation holder

Not applicable

3. Responsible parties

University Medical Center Utrecht, The Netherlands
Caitlin Dodd, PhD, assistant professor
Corinne Willame, MSc, PhD student
Miriam Sturkenboom, PharmD, PhD, MSc, professor & ACCESS coordinator

Sponsor: European Medicines Agency Address	
name, degrees, job title	name, degrees, job title
name, degrees, job title	name, degrees, job title
name, degrees, job title	name, degrees, job title

Collaborating Institutions (by alphabetical order)	Study Sites	Key persons
Aarhus University	Denmark	Vera Ehrenstein, Reimar W. Thomsen
Spanish Agency of Medicines and Medical Devices (AEMPS)	Spain	Consuelo Huerta, Mar Martín-Pérez, Patricia García-Poza
Agenzia Regionale di Sanita Toscana (ARS)	Italy	Rosa Gini, Claudia Bartolini

Bordeaux PharmacoEpi (BPE), University of Bordeaux	France	Cecile Droz, Nicholas Moore
Liebniz Institute for Prevention Research and Epidemiology - BIPS	Germany	Ulrike Haug, Tania Schink
FISABIO	Spain	Javier Diez-Domingo
IDIAP-Jordi Gol	Spain	Bonaventura Bolibar, Ainhoa Gómez
PHARMO/STIZON	The Netherlands	Josine Kuipers & Michiel Meulendijk
RIVM	The Netherlands	Hester de Melker
RTI-HS	Spain & United States of America	Susana-Perez-Gutthann Alejandro Arana
SoSeTe-Pedianet	Italy	Carlo Giaquinto
University of Messina	Italy	Gianluca Trifiro
University of Oslo	Norway	Hedvig Nordeng
Utrecht University	The Netherlands	Olaf Klungel, Helga Gardarsdottir, Patrick Souverein, Satu Siiskonen
VAC4EU secretariat	Belgium	Patrick Mahy, Juul Klaassen, Nathalie Vigot

4. Abstract

Title: Feasibility analysis of a European infrastructure for COVID-19 vaccine monitoring: Background rates of Adverse Events of Special Interest

Main authors:
Dr. C.N. Dodd, University Medical Center Utrecht, Utrecht, The Netherlands
Drs. C. Willame, University Medical Center Utrecht, Utrecht, The Netherlands
Prof. dr. M.C.J.M. Sturkenboom, University Medical Center Utrecht, The Netherlands.

Rationale and background: The global rapid spread of COVID-19 caused by the SARS-CoV2 triggered the need for developing vaccines to control for this pandemic. This study will generate background incidence rates of adverse events of special interest (AESI) that may be used to monitor benefit-risk profile of upcoming COVID-19 vaccines.

Research question and objectives:

Co-primary:

- To estimate the incidence rates of adverse events of special interest (AESI) in the general population by calendar year and data source over the period 2017 to 2020.
- To estimate the incidence of pregnancy outcomes among pregnant women aged between 12 to 55 years old by calendar year and data source over the period 2017 to 2020.
- To estimate the weekly and monthly incidence rates of COVID-19 (overall and by severity level) in 2020 by data source.
- To estimate the monthly incidence rates of multisystem inflammatory syndrome in children (MIS-C) aged between 0 to 19 years old in 2020 by data source.

Secondary:

- To estimate the incidence rates of AESI in the general population by calendar year, sex, age group, and data source over the period 2017 to 2020.
- To estimate the incidence rates of AESI in the general population by month, sex, age group, and data source over the period 2017 to 2020.
- To estimate the incidence rates of multisystem inflammatory syndrome (MIS-C) in children in 2020 by month, sex, age group, and data source.
- To estimate the prevalence of high-risk medical conditions for developing severe COVID-19 by year and data source over the period 2017 to 2020.
- To estimate the incidence rates of AESI in the at-risk population for developing severe COVID-19 by calendar year, sex, age group, and data source over the period 2017 to 2020.

Study design: A retrospective multi-database dynamic cohort study, conducted during the years 2017 to 2020, including the period of SARS-CoV-2 circulation in Europe until the date of last data availability for each data source.

Population: The study population will include all individuals observed in one of the participating data sources for at least one day during the study period (01 January 2017 - last data availability) and who has at least 1 year of data availability before cohort entry, except for individuals with data available since birth.

Variables:

Variables of interest will be

- Person-time: birth and death dates as well as periods of observation.
- Events: dates of medical and/or procedure and/or prescription/dispensing codes to identify AESI, pregnancy outcomes and at-risk medical conditions.

AESI:

Body system / Classification	**AESI (primarily based on May 27 list of SPEAC as endorsed by GACVS and the list discussed with EMA, updates may be needed if new AESI arrive)**
Auto-immune diseases	Guillain-Barré Syndrome (GBS)
	Acute disseminated encephalomyelitis (ADEM)
	Narcolepsy
	Acute aseptic arthritis

	Type I Diabetes
	Thrombocytopenia
Cardiovascular system	Acute cardiovascular injury including: Microangiopathy, Heart failure, Stress cardiomyopathy, Coronary artery disease, Arrhythmia, Myocarditis
Circulatory system	Coagulation disorders: Thromboembolism, Haemorrhage
	Single Organ Cutaneous Vasculitis
Hepato-gastrointestinal and renal system	Acute liver injury
	Acute kidney injury
Nerves and central nervous system	Generalized convulsion
	Meningoencephalitis
	Transverse myelitis
Respiratory system	Acute respiratory distress syndrome
Skin and mucous membrane, bone and joints system	Erythema multiforme
	Chilblain – like lesions
Other system	Anosmia, ageusia
	Anaphylaxis
	Multisystem inflammatory syndrome in children
	Death (any causes)
	COVID-19 (by level of severity)
	Sudden Death

Pregnancy outcomes:

Pregnancy outcome - Maternal	Gestational Diabetes
	Pre-eclampsia
	Maternal death
Pregnancy outcome - Neonates	Fetal growth restriction
	Spontaneous abortions
	Stillbirth

	Preterm birth
	Major congenital anomalies
	Microcephaly
	Neonatal death
	Termination Of Pregnancy for Fetal Anomaly (TOPFA)
	Induced abortions

Control events: colonic diverticulitis, hypertension

Data sources: The study will include data from 10 data sources in 7 European countries (Denmark, Germany, France, Italy, Netherlands, Spain, United Kingdom). Data sources contain health insurance data (BIPS, SNDS), hospitalisation record linkage data (PHARMO, Danish registries, FISABIO, SIDIAP, ARS) or data from general practitioners (CPRD, PEDIANET, BIFAP).

Study size: The study population will comprise approximately 130.6 million individuals.

Data analysis: Incidence rates (and 95%CI) of AESI and pregnancy outcomes by calendar year will be calculated by dividing the number of incident (new) cases by the total person-time (for AESIs) or pregnancies (for pregnancy outcomes) at risk.

Prevalence rates (and 95%CI) of at-risk medical conditions for developing severe COVID-19 by calendar year will be calculated by dividing the number of existing cases in a year by the average of the total number of persons recorded monthly. Incidence rates (and 95%CI) of AESI among at-risk populations will also be computed.

Sensitivity analyses will be conducted according to the time prior to SARS-CoV2 circulation and during SARS-CoV2 circulation period to investigate the impact of circulating virus on incidence rates. Additionally, a sensitivity analysis will be conducting according to the time prior to, during, and after (dependent upon data availability) lock down measures limiting face-to-face healthcare encounters to assess the impact of changes in health care behaviours on the incidence rates. In addition, incidence rates of colonic diverticulitis and hypertension, a serious and non-serious control event, respectively, will be computed.

Milestones:

Milestone	**Planned date**
Protocol submitted to the EMA	14 August 2020
Data collection	October 2020
Planned analyses completed	15 November 2020

Final report of study results	15 December 2020

5. Amendments and updates

Date	Amendment	Justification	Protocol Section
21 September	Adding transverse myelitis	Request EMA	Events of special interest

6. Deliverables and Milestones

Deliverable	**Date**
D2. Protocol	14 August 2020
D4. Results report	December 15, 2020

7. Rationale and Background

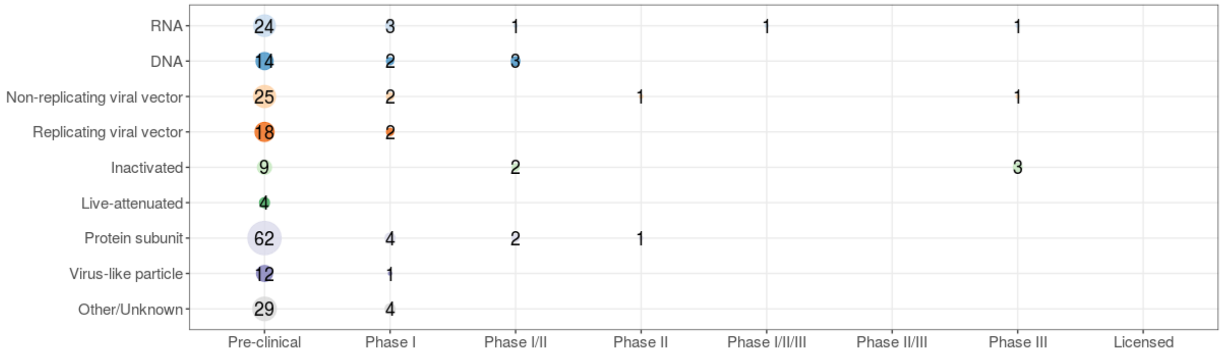
7.1 Background

COVID-19 vaccine development has been triggered on a global level following the release of the genetic sequence of SARS-CoV2 on 11 January 2020 (Le et al., 2020).

As of August 2020, the global COVID-19 vaccine research and development landscape includes 231 vaccine candidates (**Figure 1**). The most advanced vaccine candidates (28) have recently moved into clinical development, and 5 are already in clinical stage Phase 3 (i.e. encapsulated mRNA from Moderna, ChAdOx1-S from University of Oxford/AstraZeneca, and inactivated vaccines from Sinovac, Wuhan and Beijing Institute Biological Products/Sinopharm). Numerous other vaccine developers have indicated plans to initiate human testing in 2020.

The landscape for COVID-19 vaccines is characterized by a wide range of technology platforms including nucleic acid (DNA and RNA), virus-like particle, peptide, viral vector (replicating and non-replicating), recombinant protein, live attenuated virus and inactivated virus approaches. Many of these platforms are not currently the basis for licensed vaccines. It is conceivable that some vaccine platforms may be better suited to specific population subtypes.

Figure 1 Pipeline of COVID-19 vaccine candidates by technology platform



(from https://vac-lshtm.shinyapps.io/ncov_vaccine_landscape/ August 10, 2020)

The scale of the impact of the COVID-19 pandemic has accelerated the development of vaccines and the first COVID-19 vaccine candidate entered human clinical testing with unprecedented rapidity on 16 March 2020. Moderna skipped animal model tests as reported by StatNews and went straight into human (StatNews website, March 2020). The vaccine developed by CanSino Biologics has, as of July 2020, been licensed for use in the Chinese military (Reuters website, July 2020). Instead of the regular 10-15 years to market a new vaccine, COVID-19 vaccines may now be available as early as in 12-18 months, with many remaining questions about the benefits and risks and an absolute need to have a system in place that monitors carefully potential adverse effects that may happen after regulatory approval and the effectiveness of vaccines.

As of August 11, 2020, twenty-eight candidate vaccines in clinical evaluation (**Table 1**). The majority utilize new technological approaches such as messenger RNA (mRNA) or viral vector, which have not been used in vaccines previously licensed in Europe. Six candidate vaccines make use of mRNA aiming to direct the synthesis of SARS-CoV2 antigens in the cells of the vaccine recipient, while few other candidate vaccines make use of an adenovirus vector, which may stimulate a heightened immune response. Each of these vaccine technologies, along with platforms and adjuvants, will inevitably lead to different safety profiles.

Table 1 COVID-19 candidate vaccines in clinical evaluation*

COVID-19 Vaccine developer/manufacturer	Vaccine platform	Type of candidate vaccine	Number of doses	Current stage of clinical evaluation/regulatory status- Coronavirus vaccine candidate
University of Oxford/AstraZeneca	Non-Replicating Viral Vector	ChAdOx1-S	1	Phase 1/2/3
Sinovac	Inactivated	Inactivated	2	Phase 1/2/3
Wuhan Institute of Biological Products/Sinopharm	Inactivated	Inactivated	2	Phase 1/2/3
Beijing Institute of Biological Products/Sinopharm	Inactivated	Inactivated	2	Phase 1/2/3
Moderna/NIAID	RNA	LNP-encapsulated mRNA	2	Phase 1/2/3
BioNTech/Fosun Pharma/Pfizer	RNA	3 LNP-mRNAs	2	Phase 1/2/3
CanSino Biological Inc./Beijing Institute of Biotechnology	Non-Replicating Viral Vector	Adenovirus Type 5 Vector	1	Phase 1/2
Anhui Zhifei Longcom Biopharmaceutical/Institute of Microbiology, Chinese Academy of Sciences	Protein Subunit	Adjuvanted recombinant protein (RBD-Dimer)	2 or 3	Phase 1/2
Institute of Medical Biology, Chinese Academy of Medical Sciences	Inactivated	Inactivated	2	Phase 1/2
Inovio Pharmaceuticals/ International Vaccine Institute	DNA	DNA plasmid vaccine with electroporation	2	Phase 1/2
Osaka University/ AnGes/ Takara Bio	DNA	DNA plasmid vaccine + Adjuvant	2	Phase 1/2
Cadila Healthcare Limited	DNA	DNA plasmid vaccine	3	Phase 1/2
Genexine Consortium	DNA	DNA Vaccine (GX-19)	2	Phase 1/2
Bharat Biotech	Inactivated	Whole-Virion Inactivated	2	Phase 1/2
Janssen Pharmaceutical Companies	Non-Replicating Viral Vector	Ad26COVS1	2	Phase 1/2
Novavax	Protein Subunit	Full length recombinant SARS CoV-2 glycoprotein nanoparticle vaccine adjuvanted with Matrix M	2	Phase 1/2

Kentucky Bioprocessing, Inc	Protein Subunit	RBD-based	2	Phase 1/2
Arcturus/Duke-NUS	RNA	mRNA		Phase 1/2
Gamaleya Research Institute	Non-Replicating Viral Vector	Adeno-based	1	Phase 1
Clover Biopharmaceuticals Inc./GSK/Dynavax	Protein Subunit	Native like Trimeric subunit Spike Protein vaccine	2	Phase 1
Vaxine Pty Ltd/Medytox	Protein Subunit	Recombinant spike protein with Advax™ adjuvant	1	Phase 1
University of Queensland/CSL/Seqirus	Protein Subunit	Molecular clamp stabilized Spike protein with MF59 adjuvant	2	Phase 1
Institute Pasteur/Themis/Univ. of Pittsburg CVR/Merck Sharp & Dohme	Replicating Viral Vector	Measles-vector based	1 or 2	Phase 1
Imperial College London	RNA	LNP-nCoVsaRNA	2	Phase 1
Curevac	RNA	mRNA	2	Phase 1
People's Liberation Army (PLA) Academy of Military Sciences/Walvax Biotech.	RNA	mRNA	2	Phase 1
Medicago Inc.	VLP	Plant-derived VLP adjuvanted with GSK or Dynavax adjs.	2	Phase 1
Medigen Vaccine Biologics Corporation/NIAID/Dynavax	Protein Subunit	S-2P protein + CpG 1018	2	Phase 1

*Table reproduced from World Health Organization (WHO), 12 August 2020 (https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines)

7.2 Rationale for the study

When new vaccines are launched on a market and used at a large scale, monitoring of adverse events post-immunisation are necessary to ensure a proper evaluation of the benefit-risk profile of vaccines. Different methods for signal evaluation such as observed versus expected analysis and signal detection exist to identify safety signal and to assess the relationship between vaccine exposure and the occurrence of an event. These methods rely on accurate background rates of the event under evaluation. In the absence of these background rates, occurrence of rare events or an apparent increase in more common events can be interpreted as a signal of an unsafe vaccine. This stresses the importance of generating background rates of potential adverse events of special interest (AESI) in regions or countries where upcoming COVID-19 vaccines may be used (Black, et al., 2009).

This study will generate background rates of AESI that may be used to contextualize data from prospective monitoring studies and spontaneous reporting databases, and thereby, to help identify potential safety signals.

8. Research question and objectives

8.1. Co-Primary objectives

- To estimate the incidence rates of adverse events of special interest (AESI) in the general population by calendar year and data source over the period 2017 to 2020.
- To estimate the incidence of pregnancy outcomes among pregnant women aged between 12 to 55 years old by calendar year and data source over the period 2017 to 2020.
- To estimate the weekly and monthly incidence rates of COVID-19 (overall and by severity level) in 2020 by data source.
- To estimate the monthly incidence rates of multisystem inflammatory syndrome in children (MIS-C) aged between 0 to 19 years old in 2020 by data source.

8.2. Secondary objectives

- To estimate the incidence rates of AESI in the general population by calendar year, sex, age group, and data source over the period 2017 to 2020.
- To estimate the incidence rates of AESI in the general population by month, sex, age group, and data source over the period 2017 to 2020.
- To estimate the incidence rates of multisystem inflammatory syndrome (MIS-C) in children in 2020 by month, sex, age group, and data source.
- To estimate the prevalence of high-risk medical conditions for developing severe COVID-19 by year and data source over the period 2017 to 2020.
- To estimate the incidence rates of AESI in the at-risk population for developing severe COVID-19 by calendar year, sex, age group, and data source over the period 2017 to 2020.

9. Research methods

9.1 Study design

The study will be a retrospective multi-database dynamic cohort study. The study will be conducted during the years 2017 to 2020, including the period of SARS-CoV2 circulation in Europe until the date of last data availability for each data source.

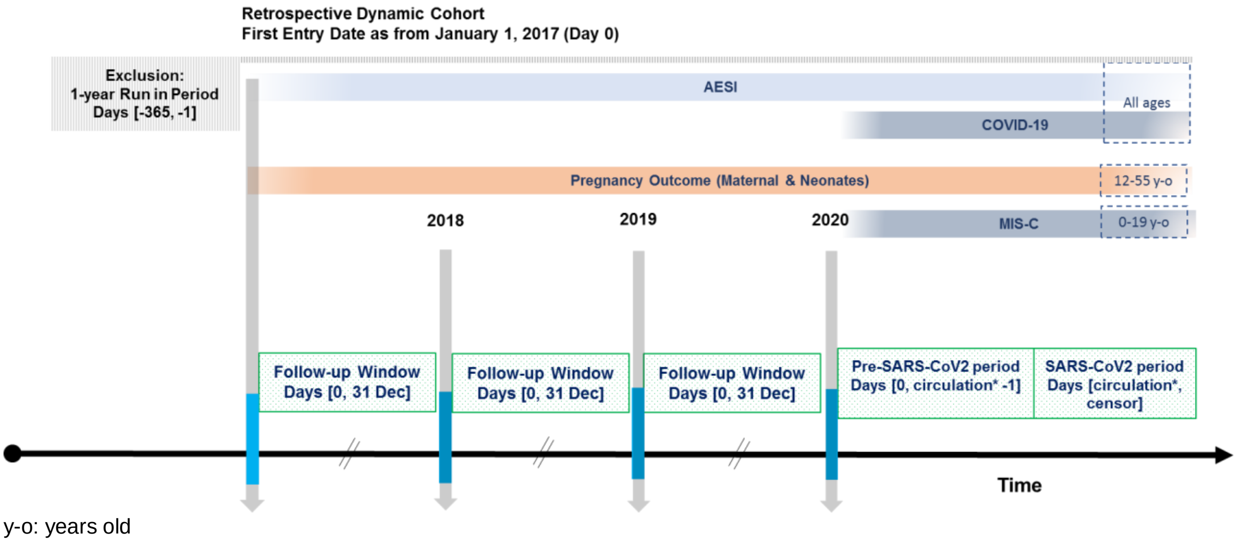
The base population will include all individuals observed in one of the participating data sources for at least one day during the study period (01 January 2017 - last data availability) and who have at least 1 year of data availability before cohort entry, except for individuals with data available since birth.

Per event, for calculation of incidence, individuals will be followed until the earliest of date of the event, death, exiting the data source, or last data draw-down. Because person-time will be censored at the occurrence of the event, person-time may vary between events.

For calculation of prevalence, individuals will be followed until death, exiting the data source, or last data draw-down.

Sub-populations such as pregnant women or children will be created according to the outcome under assessment (**Figure 2**).

Figure 2 Study design



y-o: years old
Study period from January 2017 until last data collected (eg. October 2020)
*Start of SARS-CoV2 circulation is country-specific and will be based on ECDC data
Censoring will occur at event date, last data collected, last data draw-down, or death, whichever occurs first

9.2 Setting

The study results will include data from 10 data sources in 7 European countries, comprising 130.6 million individuals (**Table 2**). Data sources are described in section 9.4.

Table 2. Overview of data sources to be used for the study

Country	Data Access Provider	Name Data source	Active population	Type of data source	Types of encounters for diagnoses	Availability of medical birth registry/data	Availability of mortality registry
Germany	BIPS	GePaRD	16 million	Health insuranc	GP, Hospital	Yes	No
Netherlands	PHARMO	PHARMO	6 million	Record linkage	GP, Hospital	Yes	Yes
Denmark	Aarhus University	Danish Registries	5.8 million	Record linkage	Hospital	Yes	Yes
Spain	AEMPS	BIFAP	8 million	GP medical	GP	No	No

Spain-Valencia	FISABIO	FISABIO	5 million	Record linkage	GP, Hospital	Yes	Yes
Spain-Catalunya	IDIAPJGol	SIDIAP	5.7 million	Record linkage	GP	No	No
Italy	SoSeTe	PEDIANET	0.5 million	Pediatric medical record	Family Pediatricians , Hospital	No	No
Italy	ARS	ARS data	3.6 million	Record linkage	Hospital	Yes	Yes
United Kingdom	Utrecht University	CPRD/HES	13 million	GP & Hospital medical	GP, Hospital	No	No
France	BPE*	SNDS	67 million	Health insuranc	Hospital	No	Yes
Total			**130.6 million**				

GP: General practitioner
*BPE, University of Bordeaux is a subcontractor of VAC4EU. BPE, University of Bordeaux may have delay in data delivery due to approval processes.

9.3 Variables

Variables of interest for the calculation of background incidence rates and prevalence rates will be those relevant for creation of:

- Person-time: birth and death dates as well as periods of observation.
- Events: dates of medical and/or procedure and/or prescription/dispensing codes to identify AESI and at-risk medical conditions.

9.3.1 Person-time & Follow-up

For incidence rates of non-pregnancy AESI, start of follow-up time will be defined as the latest of having one year of valid data in the data source, or 01 January 2017, for those who are not in the data source at birth; or as the latest between birth and 01 January 2017 otherwise. End of follow-up will be defined per event as the earliest of date of event, death, last data draw-down, or exiting the data source. Individual person-time will vary according to the event under evaluation.

For incidence rates of pregnancy outcomes, start of follow-up time will be defined at the start date of the pregnancy. For subjects pregnant on 01 January 2017 with one year of valid data prior to 01 January 2017, 01 January 2017 will be used as the start of follow-up. For subjects reaching one year of valid data in the data source during a pregnancy, the date of one year of valid data will be used as the start of follow-up. End of follow-up will be defined per pregnancy as the date of the event, end date of pregnancy (this may be equal to the date of the event), death, last data draw-down, or exiting the data source. Subjects may contribute more than one pregnancy during the study period.

For prevalence of at-risk medical conditions, start of follow-up time for identification of at-risk conditions will be defined as the latest of 01 January 2016, for those who are not in the data source at birth; or as the latest between birth and 01 January 2016 otherwise. Start of follow-up for inclusion in at risk group will be defined as the latest of having one year of valid data in the data source, or 01 January 2017, for those who are not in the data source at birth; or as the latest between birth and 01 January 2017 otherwise. End of follow-up will be defined as the earliest of date of death, last data draw-down, or exiting the data source.

9.3.2 AESI, At-risk medical conditions & Operationalization

9.3.2.1 AESI

The list of AESI has been defined based on events that are or are potentially related to marketed vaccines, events related to vaccine platforms or adjuvants, and events that may be associated with COVID-19. As part of the harmonization of COVID-19 vaccine safety monitoring during clinical development phase, the Coalition for Epidemic Preparedness Innovations (CEPI) has created a preliminary list of AESI for COVID-19 vaccine safety monitoring together with the Brighton Collaboration (SPEAC, 2020). This preliminary list did not yet include AESI related to adjuvants nor maternal/neonatal outcomes. In addition, since AS03 adjuvant will be made available for COVID-19 vaccine development and a potential association between vaccine containing AS03 (ie. Pandemrix, GSK Vaccines, Belgium) and narcolepsy has been identified during the H1N1 pandemic, the list of AESI will also include narcolepsy (Miller et al., 2013). The list of AESI has been discussed and agreed with the European Medicine Agency (EMA) advisory group monitoring committee on 9th July 2020.

Serious adverse events following immunization that will be observed during clinical development phase of COVID-19 vaccines, or AESI associated with disease may be considered for inclusion in the AESI list. Any new AESI and rational for inclusion will be added in a protocol amendment/update. Any new AESI will undergo an expedited operationalization process (see section 9.3.2.3) for availability of code lists prior to the scheduled extraction, transformation, and loading (ETL) in October 2020.

Table 3. List of AESI (based on list discussed with EMA and advisory forum, sudden death added)

Body system / Classification	AESI
Auto-immune diseases	Guillain-Barré Syndrome
	Acute disseminated encephalomyelitis
	Narcolepsy
	Acute aseptic arthritis
	Diabetes (type 1 and broader)
	(Idiopathic) Thrombocytopenia
Cardiovascular system	Acute cardiovascular injury including: Microangiopathy, Heart failure, Stress cardiomyopathy, Coronary artery disease, Arrhythmia, Myocarditis

Circulatory system	Coagulation disorders: Thromboembolism, Hemorrhage
	Single Organ Cutaneous Vasculitis
Hepato-gastrointestinal and renal system	Acute liver injury
	Acute kidney injury
Nerves and central nervous system	Generalized convulsion
	Meningoencephalitis
	Transverse myelitis
Respiratory system	Acute respiratory distress syndrome
Skin and mucous membrane, bone and joints system	Erythema multiforme
	Chilblain – like lesions
Other system	Anosmia, ageusia
	Anaphylaxis
	Multisystem inflammatory syndrome in children
	Death (any causes)
	COVID-19 disease (by levels of severity): Level 1: any recorded diagnosis, level 2: hospitalization for COVID-19 (confirmed or suspected), level 3: ICU admission in those with COVID-19 related admission; level 4: Acute respiratory distress requiring ventilation (ARDS) during a hospitalization for COVID-19; level 5 death during a hospitalization for COVID-19 (any cause)
	Sudden death
Pregnancy outcome - Maternal	Gestational Diabetes
	Preeclampsia
	Maternal death
Pregnancy outcome - Neonates	Fetal growth restriction
	Spontaneous abortions
	Stillbirth
	Preterm birth
	Major congenital anomalies
	Microcephaly
	Neonatal death
	Termination Of Pregnancy for Fetal Anomaly

Two additional events: colonic diverticulitis and hypertension, will be included as control events. These events will serve as indicators to investigate potential changes in health care behaviours during the pandemic and associated lockdown periods. Colonic diverticulitis was chosen as a serious event necessitating urgent healthcare contact while hypertension was chosen as a less serious event for which healthcare contact may be delayed.

AESI are defined using event definition forms (see **Annex 4**) and identification in the data sources will make use of medical and/or procedure and/or prescription/dispensing codes. Using information contained in event definition forms together with data access provider experience, various algorithms for definition of each AESI may be explored, algorithm development will be part of the study.

9.3.2.2 At-Risk Medical Conditions to develop severe COVID-19

At risk medical conditions for developing severe COVID-19 have been defined based on scientific evidence available on Center Disease Control (CDC website, July 2020) and National Health Services (NHS website, July 2020) websites. Those websites are updated regularly and provide a classification of at-risk conditions for developing severe COVID-19 based on level of evidence.

The selected at-risk medical conditions are considered as at higher risk to develop severe COVID-19 (**Table 4**).

The following variables will be created:

- At-Risk groups: medical codes and associated dates for at-risk medical conditions characterizing at-risk groups for developing severe COVID-19 as well as prescription and/or dispensing records for drug exposures which may be used as proxies for their identification. At-risk groups will be created for each of the at-risk medical conditions listed in **Table 4**. Multimorbidity will not be considered (subjects may belong to more than one at-risk group).

- Pregnancy start and end dates: pregnancy start and end dates will be assessed from medical birth registers for those data sources with access to a registry, while existing algorithms for defining start and end of pregnancy will be utilized in those data sources with an existing algorithm.

Table 4. Comorbid conditions with evidence of increased COVID-19 severity

At-risk medical conditions	Medicinal product proxy(ies) (ATC code)
Cancer (with chemo/immuno/radio-therapy, cancer treatment, immunosuppressant; targeted cancer treatment (such as protein kinase inhibitors or PARP inhibitors); blood or bone marrow cancer (such as leukemia, lymphoma, myeloma))	Alkylating agents (L01A) Antimetabolites (L01B) Plant alkaloids and other natural products (L01C) Cytotoxic antibiotics and related substances (L01D) Other antineoplastic agents (L01X) Hormones and related agents (L02A) Hormone antagonists and related agents (L02B)

	Immunostimulants (L03) Immunosuppressants (L04)
Type 2 Diabetes	Blood glucose lowering drugs, excluding insulins(A10B)
Obesity (BMI > 30)	Peripherally acting antiobesity products (A08AB) Centrally acting antiobesity products (A08AA)
Cardiovascular disease/ Serious heart conditions including heart failure, coronary artery disease, cardiomyopathies	Antiarrhythmics, class I and III (C01B) Cardiac stimulants excl. Cardiac glycosides (C01C) Vasodilators used in cardiac diseases (C01D) Other cardiac preparations (C01E) Antithrombotic agents (B01A)
Chronic lung disease including COPD, cystic fibrosis, severe asthma	Drugs for obstructive airway diseases (R03) Lung surfactants (R07AA) Respiratory stimulants (R07AB)
Chronic kidney disease	Erythropoietin (B03XA01)
HIV	Protease inhibitors (J05AE) Combinations to treat HIV (J05AR) NRTI (J05AF) NNRTI (J05AG)
Immunosuppression	Immunosuppressants (L04A) Corticosteroids (H02)
Sickle Cell Disease	Hydroxyurea (L01XX05) Other hematological agents (B06AX)
Negative Control	
Negative Control Conditions	**Medicinal product proxy(ies) (ATC code)**
Colonic Diverticulitis	-
Hypertension	First anti-hypertensive drugs (C02, C03, C07, C08, C09)
Pregnancy	
Start date of pregnancy	
End date of pregnancy	

9.3.2.3 Operationalization

For each of the events of interest living event definition forms have been created (see **Annex 2, Annex 4**) comprising the following chapters:

- Event definition: using the Brighton Collaboration definitions if available and otherwise definitions from European learned societies
- Synonyms / lay terms used for the event: these show how an event may be described/called in free text
- Laboratory tests done specific for event (may be used as confirmation)
- Diagnostic tests done specific for event (may be used as confirmation in building algorithms)
- Drugs used to treat event (may be used as confirmation in building algorithms)
- Procedures used specific for event treatment (may be used as confirmation in building algorithms)
- Setting (outpatient specialist, in-hospital, GP, emergency room) where condition will be most frequently diagnosed
- Diagnosis codes or algorithms used in different papers to identify the events in Europe/USA
- Experience of participating data sources to identify or validate the events (to be completed by each data source)
- Proposed codes by Codemapper (Becker et al., 2017)
- Algorithm proposal for event identification; several algorithms will be built during the execution of the protocol using diagnosis codes, provenance, and confirmatory tests/drugs/procedures. (Roberto et al., 2016; Gini et al, 2019):
- Published background rates
- Extracted codes (upon characterization)
- Study design related information
    - Estimated lag time from onset to diagnosis
    - Is condition a contraindication to any vaccination?
    - Is this a chronic or potentially recurrent condition?
    - Does this condition cause increased fatality?
    - Time to onset (from vaccination and/or infection)
- References

The event definition form will be used throughout the project to transparently track how an event is defined and identified in each of the data sources. It will be the basis for the creation of study variables and algorithms and be an evolving document capturing which codes and algorithms were used (see **Annex 2**). Forms will be made available on the VAC4EU website after approval from EMA.

9.3.3 Other variables

- Demographic characteristics: dates of birth and death, sex, country and /or region, data source.

In those data sources in which full date of birth is not available, date of birth will be derived as follows:

- Date of birth will be defined as the 15th of the birth month and birth year. If the birth month is missing, the birth date will be defined as the 30th June of the birth year.

9.4 Data sources

9.4.1 Description of data sources participating in this protocol

9.4.1.1 Germany: GePaRD

GePaRD is based on claims data from four statutory health insurance providers in Germany and currently includes information on approximately 25 million persons who have been insured with one of the participating providers since 2004 or later. Per data year, there is information on approximately 20% of the general population and all geographical regions of Germany are represented. In addition to demographic data, GePaRD contains information on dispensations of reimbursable prescription drugs as well as outpatient (i.e., from general practitioners and specialists) and inpatient services and diagnoses. GePaRD also contains information on influenza vaccinations and routine childhood immunizations and there is experience with studies on utilization and risk of vaccination and on background incidence of adverse events of vaccinations (Hense et al., 2014; Schink et al., 2014). GePaRD data have been used for vaccine safety studies. GePaRD is listed under the ENCePP resources database.

9.4.1.2 Netherlands: PHARMO Database Network

The PHARMO Database Network is a population-based network of electronic healthcare databases and combines anonymous data from different primary and secondary healthcare settings in the Netherlands. These different data sources, including data from general practices, in- and out-patient pharmacies, clinical laboratories, hospitals, the cancer registry, pathology registry and perinatal registry, are linked on a patient level through validated algorithms. To ensure the privacy of the data in the PHARMO Database Network, the collection, processing, linkage and anonymization of the data is performed by STIZON. STIZON is an independent, ISO/IEC 27001 certified foundation, which acts as a Trusted Third Party between the data sources and the PHARMO Institute. The longitudinal nature of the PHARMO Database Network system enables to follow-up more than 9 million persons of a well-defined population in the Netherlands for an average of twelve years. Currently, the PHARMO Database Network covers over 6 million active persons out of 17 million inhabitants of the Netherlands. Data collection period, catchment area and overlap between data sources differ. Therefore, the final cohort size for any study will depend on the data sources included. As data sources are linked on an annual basis, the average lag time of the data is one year. All electronic patient records in the PHARMO Database Network include information on age, sex, socioeconomic status and mortality. Other available information depends on the data source. A detailed description of the different data sources is given below. PHARMO is always seeking new opportunities to link with healthcare databases. Furthermore, it is possible to link additional data collections, such as data from chart reviews, patient-reported outcomes or data from general practice trials.

The General Practitioner database comprises data from electronic patient records registered by GPs. The records include information on diagnoses and symptoms, laboratory test results, referrals to specialists and healthcare product/drug prescriptions. The prescription records include information on type of product, prescription date, strength, dosage regimen, quantity and route of administration. Drug prescriptions are coded according to the WHO ATC Classification System [www.whocc.no]. Diagnoses and symptoms are coded according to the International Classification of Primary Care - ICPC [www.nhg.org], which can be mapped to the International Classification of Diseases - ICD codes, but can also be entered as free text. GP data cover a catchment area representing 3.2 million residents (~20% of the Dutch population).

The Out-patient Pharmacy Database comprises GP or specialist prescribed healthcare products dispensed by the out-patient pharmacy. The dispensing records include information on type of product, date, strength, dosage regimen, quantity, route of administration, prescriber specialty and costs. Drug dispensings are coded according to the WHO ATC Classification System. Out-patient pharmacy data cover a catchment area representing 4.2 million residents (~25% of the Dutch population). PHARMO is listed under the ENCePP resources database. PHARMO data capture influenza vaccine and may be linked to the PRAEVENTIS database that is held by RIVM, based on specific permissions.

9.4.1.3 Denmark: Danish Registries

Denmark has a tax-funded health care system ensuring easy and equal access to health care for all its citizens, and with this system all contacts are recorded in administrative and medical registers (Schmidt et al., 2019). The records carry a unique personal identification number, called the CPR-number, assigned to every Danish citizen. Linkage between registers at an individual level is possible because this CPR-number is used in all Danish registers and assigned by the Danish Civil Registration System (Schmidt et al., 2014). All registers have a nationwide coverage and an almost 100% capture of contacts covering information on currently 5.8 million inhabitants plus historical information. For the purpose of the study we will obtain information from the following registries. The Danish National Prescription Registry (DNPR) includes data on all outpatient dispensing of medications and vaccines at Danish pharmacies from 1995 and onwards, including dispensing date, ATC code, product code and amount (Pottegard et al., 2017). The Danish National Health Service Register includes data on primary care services, including general practitioner contacts, examinations, procedures, and vaccinations; psychologist or psychiatrist and other primary care provider visits; etc. From the Danish Civil Registration System, data on demographics (sex, date of birth) and censoring (migration, vital status). The Danish National Patient Registry contains diagnoses and procedures from all hospitalizations since 1977 and contacts to hospital outpatient clinics since 1995 (Schmidt et al., 2015). The Danish National Health Service Register contains information on referral for vaccine administration from GPs (Sahl Andersen et al., 2011). The Danish databases were characterized in the ADVANCE project and considered fit for purpose for vaccine coverage, benefits and risk assessment and could participate in near real-time monitoring (Sturkenboom et al., 2020; Bollaerts et al., 2019).

9.4.1.4 Spain: BIFAP

BIFAP (Base de Datos para la Investigación Farmacoepidemiológica en Atencion Primaria), a computerized database of medical records of primary care (www.bifap.aemps.es) is a non-profit research project funded by the Spanish Agency for Medicines and Medical Devices (AEMPS). The project started in 2001 and current version of the database with information until December 2019 includes clinical information of 6,419 GPs and 1,147 pediatricians. Ten participant autonomous regions send their data to BIFAP every year. BIFAP database currently includes anonymized clinical and prescription/dispensing data from around 14 million (8 active population) patients representing 85% of all patients of those regions participating in the database, and 25% of the Spanish population. Mean duration of follow-up in the database is 8.6 years. Diagnoses are classified according to the International Classification of Primary Care (ICPC)-2 and ICD-9 code system. Information on hospital outpatient diagnosis is being progressively included. The BIFAP database was characterized in the ADVANCE project and considered fit for purpose for vaccine coverage, benefits and risk assessment (Sturkenboom et al., 2020).

9.4.1.5 Spain: SIDIAP

The Information System for Research in Primary Care (Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària' - SIDIAP; www.sidiap.org) was created in 2010 by the Catalan Health Institute (CHI) and the IDIAPJGol Institute. It includes information collected since 01 January 2006 during routine visits at 278 primary care centers pertaining to the CHI in Catalonia (North-East Spain) with 3,414 participating GPs. SIDIAP has pseudo-anonymized records for 5.7 million people (80% of the Catalan population) being highly representative of the Catalan population.

The SIDIAP data comprises the clinical and referral events registered by primary care health professionals (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. It 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 ICD-10 codes, ATC codes and structured forms designed for the collection of variables relevant for primary care clinical management, such as country of origin, sex, age, height, weight, body mass index, tobacco and alcohol use, blood pressure measurements, blood and urine test results. In relation to vaccines, SIDIAP includes all routine childhood and adult immunizations, including the antigen and the number of administered doses. Encoding personal and clinic identifiers ensures the confidentiality of the information in the SIDIAP database. The SIDIAP database is updated annually at each start of the year.

But nowadays, with the COVID-19 pandemic, there is the possibility to have shorter term updates in order to monitor the evolution of the pandemic. Recent reports have shown the SIDIAP data to be useful for epidemiological research. SIDIAP is listed under the ENCePP resources database www.encepp.eu/encepp/resourcesDatabase.jsp). The SIDIAP database was characterized in the ADVANCE project and considered fit for purpose for vaccine coverage, benefits and risk assessment (Sturkenboom et al., 2020).

9.4.1.6 Spain: FISABIO

The region of Valencia, with 5 million inhabitants, is part of the Spanish National Health System, a universal public healthcare system. Information will be obtained from the population-based electronic information systems of the Valencia Health Agency (VHA) and the regional Government of Valencia: (1) The Population Information System (SIP) provides an identification number for each person under Valencian Health Service (VHS) coverage, and registers some demographic characteristics, and dates and causes of VHA discharge, including death. (2) The minimum basic dataset at hospital discharge is a synopsis of clinical and administrative information on all hospital discharges, including diagnoses and procedures (all electronic health systems in the VHS use the ICD-9-CM). (3) The Emergency Department module (ED) including ED dates of visit and discharge and reason for discharge. (4) The electronic medical record (EMR) for ambulatory care, available in all primary healthcare centers and other ambulatory settings. It has all the information on patients regarding diagnoses, their personal and family medical history, laboratory results, lifestyle, etc. (5) The pharmaceutical module (prescription information system), part of EMR, includes information about both physician prescriptions and dispensations from pharmacy claims. (6) The Corporate Resource Catalogue provides information about the geographical and functional organization of VHS, its health centers, health services provided and professionals in healthcare. Specific public health registries are available and linkable at an individual level (such as the perinatal registry and the congenital anomalies registry, from which pregnancy outcomes can be obtained) All the information in these systems can be linked at an individual level through the SIP number. The FISABIO database was not characterized in ADVANCE, but did provide important information evaluating the pandemic influenza vaccine and narcolepsy (Dodd CN et al., 2018; Weibel et al., 2019)

9.4.1.7 Italy: PEDIANET database

PEDIANET, a pediatric general practice research database, contains reason for accessing healthcare, health status (according to the Guidelines of Health Supervision of the American Academy of Pediatrics), demographic data, diagnosis and clinical details (free text or coded using the ICD-9 CM), prescriptions (pharmaceutical prescriptions identified by the ATC code), specialist appointments, diagnostic procedures, hospital admissions, growth parameters and outcome data of the children habitually seen by about 140 family pediatricians (FPs) distributed throughout Italy.

PEDIANET can link to other databases using unique patient identifiers. In the first database, information on routine childhood vaccination are captured including vaccine brand and dose. In the second database, information on patient hospitalization date, reason for hospitalization, days of hospitalizations and discharge diagnosis (up to six diagnosis) are captured. The FPs participation in the database is voluntary and patients and their parents provide consent for use of their data for research purposes. In Italy each child is assigned to a FP, who is the referral for any health visit or any drug prescription, thus the database contains a very detailed personal medical history. The data, generated during routine practice care using common software (JuniorBit®), are anonymized and sent monthly to a centralized database in Padua for validation. The PEDIANET database can be linked to regional vaccination data which was successfully tested in the ADVANCE project where it was characterized and deemed fit for purpose for pediatric routine vaccines (Sturkenboom et al., 2020).

9.4.1.8 Italy: ARS database

The Italian National Healthcare System is organized at regional level: the national government sets standards of assistance and a tax-based funding for each region, and regional governments are responsible to provide to all their inhabitants. Tuscany is an Italian region, with around 3.6 million inhabitants. The Agenzia Regionale di Sanita' della Toscana (ARS) is a research institute of the Tuscany Region. The ARS database comprises all information that are collected by the Tuscany Region to account for the healthcare delivered to its inhabitants. Moreover, ARS collects data from regional initiatives. All the data in the ARS data source can be linked with each other at the individual level, through a pseudo-anonymous identifier. The ARS database routinely collects primary care and secondary care prescriptions of drugs for outpatient use, and is able to link them at the individual level with hospital admissions, admissions to emergency care, records of exemptions from copayment, diagnostic tests and procedures, causes of death, mental health services registry, birth registry, spontaneous abortion registry, induced terminations registry. A pathology registry is available, mostly recorded in free text, but with morphology and topographic Snomed codes. Mother-child linkage is possible through the birth registry. Vaccine data is available since 2016 for children and since 2019 for adults. However, to date, 2019 vaccination data for adults may still be incomplete. The ARS database was characterized in the ADVANCE project and considered fit for purpose for vaccine coverage, benefits and risk assessment when using the new vaccine registry (from 2019) (Sturkenboom et al., 2020).

9.4.1.9 United Kingdom: CPRD & HES

The Clinical Practice Research Datalink (CPRD) from the UK collates the computerized medical records of general practitioners (GPs) in the UK who act as the gatekeepers of healthcare and maintain patients' life-long electronic health records. As such they are responsible for primary healthcare and specialist referrals, and they also store information stemming from specialist referrals, and hospitalizations. GPs 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 feedback 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. The majority of the data are coded in Read Codes. Validation of data with original records (specialist letters) is also available.

The dataset is generalizable to the UK population based upon age, sex, socioeconomic class and national geographic coverage when GOLD & Aurum versions are used.
There are currently approximately 42 million patients (acceptable for research purposes) – of which 13 million are active (still alive and registered with the GP practice) – in approximately 1,700 practices (https://cprd.com/Data). Data include demographics, all GP/healthcare professional consultations (phone, letter, email, in surgery, at home), diagnoses and symptoms, laboratory test results, treatments, including all prescriptions, all data referrals to other care, hospital discharge summary (date and Read codes), hospital clinic summary, preventive treatment and immunizations, death (date and cause). For a proportion of the CPRD panel practices (>80%), the GPs have agreed to permit CPRD to link at patient level to the Hospital Episode Statistics (HES) data. CPRD is listed under the ENCePP resources database, access will be provided by the Utrecht University. The CPRD was not yet characterized in the ADVANCE project, where the UK THIN and RCGP databases were used, but has been largely used in vaccine studies.

The HES database contains details of all admissions to National Health System (NHS) hospitals in England; approximately 60% of GP practices in the CPRD are linked to the HES database. Not all patients in the CPRD have linked data (e.g. if they live outside England or if their GP has not agreed that their data should be used in this way). As with standard CPRD patients, HES data are limited to research-standard patients. CPRD records are linked to the HES using a combination of the patient's NHS number, gender and date of birth (Williams et al., 2012).

9.4.1.10 France: Système National des Données de Santé (SNDS)

The SNDS (Système National des Données de Santé) 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, etc.
- Death: month, year and cause
- Long-Term Disease registration associated with an ICD-10 diagnostic codes
- 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.
- Inpatients 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, and the related costs. Drugs included in the diagnosis

related group cost are not captured. However, expansive 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 dispensings. 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 and data extraction need approval from the CESREES (*Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé*) in charge of assessing scientific quality of the project, and authorization from the CNIL (French data protection commission), and then contracts with the SNDS data holder (CNAM) for data extraction. Bordeaux PharmacoEpi (BPE), a research platform of the University of Bordeaux specialized in real world studies, will be in charge of requesting access to SNDS data. The SNIIRAM data were not yet characterized in the ADVANCE project but have been used for vaccine studies.

9.4.2 Data Availability

Analysis is scheduled to start in October 2020 using the most up-to-date data available for each data source. **Table 5** displays estimated end dates of available data for analysis in October 2020; data which is not available for a data source is indicated by NA.

Table 5. Estimated end date of data availability for each data source for inclusion in October 2020 analysis for EMA

Country	Data Access Provider	Name Data Source	End date of data availability/lag times						
			Diagnoses, Signs, and Symptoms						
			Hospital	Emergency	Primary care	Outpatient specialist care	Prescription and/or Dispensing	Procedures and/or Measurements	Birth Registry
Germany	BIPS	GePaRD	December 2017	December 2017	December 2017	December 2017	December 2017	December 2017	December 2017
Netherlands	PHARMO	PHARMO	December 2019	December 2019	December 2019	NA	December 2019	NA	December 2017
Denmark	Aarhus University	Danish Registries	December 2018	December 2018	NA	NA	December 2018	NA	December 2018
Spain	AEMPS	BIFAP	NA	NA	December 2019	NA	December 2019	December 2019	NA
Spain-Valencia	FISABIO	FISABIO	February 2020	February 2020	NA	NA	October 2019	NA	April 2020
Spain-Catalunya	IDIAP-Jordi Gol	SIDIAP	June 2020	June 2020	NA	June 2020	June 2020	June 2020	June 2020

Italy	SoSeTe	PEDIANET	April 2020	April 2020	June 2020	NA	June 2020	June 2020	June 2020
Italy	ARS	ARS database	June 2020	June 2020	NA	June 2020	May 2020	June 2020	June 2020
United Kingdom	Utrecht University	CPRD/HES	June 2019	June 2019	September 2020	June 2019	September 2020	September 2020	NA
France	BPE	SNDS	December 2019	December 2019	NA	December 2019	December 2019	December 2019	NA

* Updates may be possible in the future

9.5 Study size

The study population will include all individuals registered with at least one year of data prior to the start of the study period or follow-up from birth. Overall, the study population will comprise approximately 130.6 million individuals (see **Table 4**).

9.6 Data management

This study will be conducted in a distributed manner using a common protocol, common data model (CDM), and common analytics programs. This process was used successfully in several other European multi-database projects (Trifiro et al., 2014) (**Figure 3**). The data pipeline has been further specified in theoretical terms by Gini et al. (Gini et al., 2016; Gini et al., 2020) and further improved in the IMI-ConcePTION project (https://www.imi-conception.eu/). This process maximizes the involvement of the data providers in the study by utilizing their knowledge on the characteristics and the process underlying the data collection which makes analysis more efficient.

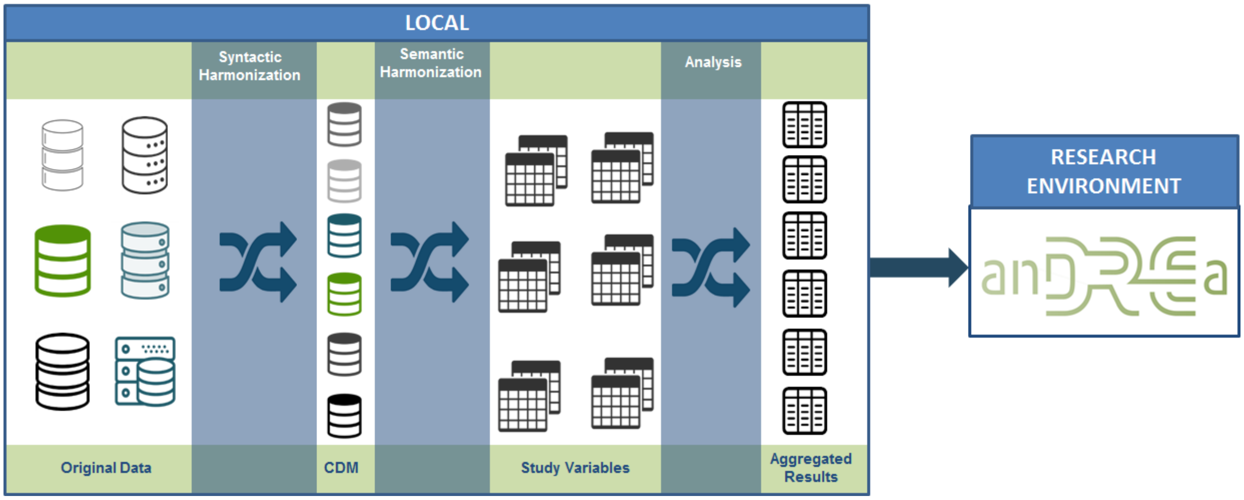
1. First, to harmonize the structure of the data sets held by each partner, a shared syntactic foundation is utilized. Syntactic foundation is described in **Annex 1** and refers to the syntactically harmonized CDM. In this common data model, data is represented in a common structure but the content of the data remain in their original format. The extraction, transform, and load (ETL) design will be shared on a searchable FAIR catalogue. The VAC4EU FAIR data catalogue is a meta-data management tool designed to contain searchable meta-data describing organizations that can provide access to specific data sources. FAIR means: findable, accessible, interoperable and re-usable. The VAC4EU catalogue will be a mirror of the ConcePTION catalogue but will be populated with new and additional data sources. Data quality checks will be conducted to measure the integrity of the ETL as well as internal consistency within the instance of the CDM (see **section 10. Quality Control**).

2. Second, to reconcile differences across terminologies a shared semantic foundation is built for the definition of events under study by collecting relevant concepts in a structured fashion using a standardized event definition template (see **Annex 2**). The Codemapper tool will be used to create diagnosis code lists based upon completed event definition templates for each AESI and comorbid

risk condition (Becker et al., 2017). Based on the relevant diagnostic medical codes and keywords, as well as other relevant concepts (e.g. medications), one or more algorithms will be constructed (typically one sensitive, or broad, algorithm and one specific, or narrow algorithm) to operationalize the identification and measurement of each event. These algorithms may differ per database, as the components that go into the study variable may differ (Roberto, 2016; Gini, 2019). No validation will be done for this study, as there are no resources for this within the budget of the EMA tender. Wherever possible the event definition sheet will specific prior validation of algorithms and codes. Specifications for both ETL and semantic harmonization will be shared on the catalogue. Scripts for semantic harmonization will be developed in R, distributed to data access providers for local deployment, and shared on the catalogue. The impact of choices of different algorithms will be assessed quantitatively. This will result in a set of study variables which are both semantically and syntactically harmonized.

3. Third, following conversion to harmonized study variable sets, R programs for calculation of incidence and prevalence will be distributed to data access providers for local deployment. The aggregated results produced by these scripts will then be uploaded to the Digital Research Environment (DRE) for pooled analysis and visualization (see **Figure 3**). The DRE is made available through UMCU/VAC4EU (https://www.andrea-consortium.org/). The DRE is a cloud based, globally available research environment where data is stored and organized securely and where researchers can collaborate (https://www.andrea-consortium.org/azure-dre/).

Figure 3 Data management plan



9.6.1 Data extraction

Each database access provider (DAP) will create ETL specifications using the standard ConcePTION ETL design template (accessible via this link: https://docs.google.com/document/d/1SWi31tnNJL7u5jJLbBHmoZa7AvfcVaqX7jiXgL9uAWg/edit.

Following completion of this template and review with study statisticians, each DAP will extract the relevant study data locally using their software (eg Stata, SAS, R, Oracle). This data will be loaded into the CDM structure (see **Annex 1**) in csv format. These data remain local (**Figure 3**).

9.6.2 Data Processing and transformation

Data processing and transformation will be conducted using R code against the syntactically harmonized CDM. The R scripts will first transform the data in the syntactically harmonized CDM to semantically harmonized study variables (see **Figure 3**). Following creation of study variables, the data will be characterized. This characterization will include calculation of code counts and incidence rates, as well as benchmarking within data source (over time), between data sources, and externally (against published estimates). Subsequently, R code to conduct analysis against semantically harmonized study variables will be distributed and run locally to produce aggregated results. The R scripts for these processing and analysis steps will be developed and tested centrally and sent to the DAPs. The R scripts will be structured in modular form in such a way that transparency is ensured. Functions to be used in the modules will be either standard R packages or packages designed, developed and tested on purpose for multi-database studies. As a result, scripts will be thoroughly documented and this will allow verification. The DAPs will run the R code locally and send aggregated analysis results to the anDREa digital research environment using a secure file transfer protocol. In the anDREa DRE, results will be further plotted, inspected (for quality assessment) and pooled (if needed) for final reporting.

9.6.3 Software and Hardware

All final statistical computations will be performed on the DRE using R and/or SAS. Data access providers will have access to the workspace for verification of the scripts.

9.6.4 Storage

Aggregated results, ETL specifications, and a repository of study scripts will be stored in the DRE.

9.6.5 Access

Within the DRE, each project-specific area consists of a separate, secure folder, called a 'workspace'. Each workspace is completely secure, so researchers are in full control of their data. Each workspace has its own list of users, which can be managed by its administrators.

The architecture of the DRE allows researchers to use a solution within the boundaries of data management rules and regulations. Although General Data Protection Regulation (GDPR) and Good (Clinical) Research Practice still rely on researchers, the DRE offers tools to more easily control and monitor which activities take place within projects.

All researchers who need access to DRE are granted access to study-specific secure workspaces. Access to this workspace is only possible with double authentication using an ID and password together with the user's mobile phone for authentication.

Upload of files is possible for all researchers with access to the workspace within the DRE. Download of files is only possible after requesting and receiving permission from a workspace member with an 'owner' role.

9.6.6 Archiving and record retention

The final study aggregated results sets and statistical programs will be archived and stored on the DRE and the VAC4EU Sharepoint. The validation of the quality control (QC) of the statistical analysis will be documented. The final study protocol and possible amendments, the final statistical report, statistical programs and output files will be archived on a specific and secured drive centrally.

Documents that individually and collectively permit evaluation of the conduct of a study and the quality of the data produced will be retained for a period of 5 years in accordance with GPP guidelines. These documents could be retained for a longer period, however, if required by the applicable regulatory requirements or by an agreement between study partners. It is the responsibility of the principal investigator to inform the other investigators/institutions as to when these documents no longer need to be retained. Study records or documents may also include the analyses files, syntaxes (usually stored at the site of the database), ETL specifications, and output of data quality checks.

9.7 Data analysis

All analyses will be detailed in a Statistical Analysis Plan that will be developed ahead of data extraction.

9.7.1 Analysis of Demographics and Baseline Characteristics

Demographic characteristics (age at study entry and sex) and baseline characteristics such as at-risk medical conditions and pregnancy will be summarized for each data source using descriptive statistics.

Frequency tables including numbers and percentages will be generated for categorical variables (age at study entry in categories, sex, pregnancy, and at-risk medical conditions).

Mean, standard error, median and range will be provided for continuous variables (age at study entry).

9.7.2 Hypotheses

Not applicable. This study is not hypothesis testing.

9.7.3 Statistical Methods

Incidence rates by calendar year will be calculated by dividing the number of incident cases (not in run-in year) (numerator) by the total person-time at risk (denominator). A 95%CI will be computed using an exact method (Ulm, 1990).

Prevalence rates by calendar year will be calculated by dividing the number of existing cases in a year (numerator) by the average of the total number of persons recorded monthly (denominator). A 95%CI will be computed using an exact method (Vollset, 1993).

Incidence rates will also be reported stratified by time prior to SARS-CoV2 circulation and during SARS-CoV2 circulation period to investigate potential changes in health care behaviours during the pandemic and associated lockdown periods on the incidence rates.

Incidence rates will also be provided among persons at higher risk for developing severe COVID-19 (as described in section 9.3.3).

The period during which SARS-CoV2 was circulating will be country-specific and based on the number of cases of COVID-19 reported by ECDC (see **Annex 3**). The start of the SARS-CoV2 circulation period will be defined as the date at which the first case in the country was reported. The end of the SARS-CoV2 circulation period will be the date of last data available in each data source.

9.7.4 Statistical Analysis

9.7.4.1 Analysis of co-primary objectives

- Incidence rate (and 95% CI) of AESI will be calculated for all individuals by calendar years and data sources: the numerator will be the number of incident cases (not in the run-in year) in each calendar year (2017, 2018, 2019, 2020 pre-SARS-CoV2 period and 2020 SARS-CoV2 period) and each data source. The denominator will be the total person-years at risk, i.e. from 1st January or birth until date of event, death, last data draw-down, or leaving the database, whichever occurs first, in each calendar year and each data source.

- Incidence rate (and 95% CI) of pregnancy outcomes will be calculated in women aged 12 to 55 years by calendar year and data sources: the numerator will be the number of pregnancy outcomes among women aged 12 to 55 years in each calendar year (2017, 2018, 2019, 2020 pre-SARS-CoV2 period and 2020 SARS-CoV2 period) and each data source. The denominator will be the total pregnancies at risk among pregnant women in each calendar year and each data source.

- Incidence rate (and 95% CI) of recorded COVID-19 disease (overall and by severity level) will be calculated by calendar months and calendar weeks for the year 2020 and data sources: the numerator will be the number of incident COVID-19 cases and the denominator will be the total person-months or person-weeks at risk, i.e. from 1st January 2020 or birth until date of event, death, last data draw-down or leaving the database whichever occurs first, in each calendar month or week and each data source. Each COVID-19 cases will be classified by severity level (Level 1: any recorded diagnosis, level 2: hospitalization for COVID-19 (confirmed or suspected), level 3: ICU admission in those with COVID-19 related admission; level 4: Acute respiratory distress requiring ventilation (ARDS); level 5 death (any cause)). Incidence rate (and 95% CI) of each severity level (1, 2, 3, 4 or 5) of COVID-19 will be estimated using the same approach as that used for COVID-19.

- Incidence rate (and 95% CI) of MIS-C will be calculated in children aged 0 to 19 years by calendar month for the year 2020 and data sources: the numerator will be the number of incident cases among children aged 0 to 19 years in each data source. The denominator will be the total person-months at risk in those up to 19 years old, i.e. from 1st January 2020 or birth until date of event, death, last data draw-down, leaving the database, end of the month or 19th birthday, whichever occurs first, in each calendar month and each data source.

9.7.4.2 Analysis of secondary objectives

- Incidence rates (and 95% CI) of AESI using further stratifications will be estimated using the same approach as described for AESI in section 9.7.4.1. Rates will be stratified by calendar year, sex, age group (Year of age in subjects <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and older) and data source.

- Monthly incidence rates (and 95% CI) of AESI will be estimated for all individuals by month, sex, age group and data source: the numerator will be the number of incident cases (not in the run-in year) by months from 01 January 2017 until last data available (e.g. October 2020), sex, age group (Year of age in subjects <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and older) and each data source. The denominator will be the total person-months at risk, i.e. from 1st January or birth until date of event, death, last data draw-down, leaving the database or end of the month whichever occurs first, in each month and each data source. Monthly incidence rates of AESI will be presented graphically and will help to interpret potential seasonality patterns among selected AESI.

- Monthly incidence rates (and 95% CI) of MIS-C using further stratifications will be estimated using the same approach as described for MIS-C in section 9.7.4.1. Rates will be stratified by calendar month, sex, year of age and data source.

- Prevalence rates (and 95% CI) of at-risk medical conditions for developing severe COVID-19 and prevalence of the use of immunosuppressants will be calculated by dividing the number of individuals identified with an at-risk medical condition by the average of the total number of individuals recorded in a month. Prevalence rates will be estimated for each calendar year (2017, 2018, 2019, 2020 pre-SARS-CoV2 period and 2020 SARS-CoV2 period), by sex, age groups (Year of age in subjects <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and older) and data source.

Subjects identified as having an at-risk condition in the run-in period will be considered prevalent cases and at-risk at study start (01 January 2017).

- Incidence rates (and 95% CI) of AESI in each at-risk population will be estimated by calendar year, sex, age group (Year of age in subjects <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and older) and data source using the same approach as described for AESI in section 9.7.4.2.

Table 6 Incidence rates and prevalence rates calculations for the main analyses

Analysis	**Numerator**	**Denominator**	**Stratification factors**
Incidence rate of AESI	# of new cases of any AESI	Total person-years (person-months) at risk of all subjects	Data sources Calendar time (in years and months) Sex Age group #1
Incidence rate of pregnancy outcomes	# of new events of any pregnancy outcomes	Total pregnancies in women aged 12 to 55 years	Data sources Calendar time (in years) Age group #2
Incidence rate of recorded COVID-19	# of new cases of recorded COVID-19 split by severity	Total person-months or person-weeks at risk of all subjects	Data sources Calendar time (in months and weeks in 2020) Sex Age group #1 Disease severity
Incidence rate of MIS-C	# of new cases of MIS-C	Total person-months at risk of subjects aged 0 to 19 years	Data sources Calendar time (in month in 2020) Sex Age group #3
Proportion of subjects with each at-risk medical condition	# of existing individuals with at-risk medical conditions	Average of total # of individuals registered monthly	Data sources Calendar time (in years) Sex Age group #1
Incidence rate of AESI in each at-risk population	# of new cases of any AESI	Total person-years of existing individuals with at-risk medical conditions	Data sources Calendar time (in years) Sex Age group #1

AESI: Adverse Event of Special Interest; MIS-C: multisystem inflammatory syndrome
Age group #1: Year of age for subjects <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and older
Age group #2: 12-19, 20-29, 30-39, 40-55
Age group #3: 0-4, 5-9, 10-14, 15-19

9.7.5 Missing data

Since the underlying data represent attended medical care we generally assume that absence of information of clinical events means absence of that condition. No imputation will be done for missing data.

9.7.6 Sensitivity analysis

To investigate potential changes in health care behaviours during the pandemic and associated lockdown periods on the incidence rates, sensitivity analyses will be conducted on:

- Co-primary objectives (AESI and pregnancy outcomes as described in section 9.7.4.1) by considering:
  - The peak of the COVID-19 pandemic: A threshold number of cases will be applied to identify the beginning and the end of the pandemic peak in each of the countries. This threshold value will be fixed at 1/10 of the highest daily number of cases reported in a country (eg. in Germany the highest daily number of cases reported was 6294, therefore the beginning of the pandemic peak will begin on the date when at least 629 cases were reported and the end of the pandemic peak will be the date when no more than 629 cases are reported).
  - A pre-SARS-CoV2 period (overall period starting from 2017 until the start of SARS-CoV2 circulation period) and SARS-CoV2 circulation period. Periods will be defined for each country (see **Annex 3**).
  - A pre-lockdown (overall period starting from 2017 until the imposition of lockdown policies limiting face-to-face healthcare encounters), lockdown, and post-lockdown (dependent upon availability of data) periods. Periods will be defined for each country (see **Annex 3**).

- Control events: Two control events will be included in the study. Colonic diverticulitis will be included as a control event for which rapid medical care or surgery are required. Hypertension (defined as first prescription or dispensing of antihypertensive medication) will be included as a control event for which medical care may be delayed. Incidence rates of colonic diverticulitis and hypertension will be computed as described for AESI in section 9.4.7.1 and section 9.4.7.2. It is assumed that the reporting of colonic diverticulitis should not be affected by the pandemic, while it is expected that healthcare contact for hypertension may be delayed due to the pandemic and associated lockdowns.

10. Quality control

10.1 Quality management

The study 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 data access providers have experience in conducting pharmacoepidemiological research and research is done by researchers trained in pharmacoepidemiology. All programs will be developed according to agreed coding standards and will be validated by double programming or source code review with second programmer involvement. Only validated software (Stata, R and/or SAS version 9.4, SAS Institute Inc., Cary, NC) will be used for statistical analyses.

10.2 Data Quality

Data quality will be characterized in a transparent manner according to the procedures developed in the IMI-ConcePTION project on the syntactically harmonized data. This process will proceed iteratively and in collaboration with each data access provider.

Level 1 data checks review the completeness and content of each variable in each table of the CDM to ensure that the required variables contain data and conform to the formats specified by the CDM specifications (e.g., data types, variable lengths, formats, acceptable values, etc.).

This is a check conducted in collaboration with DAPs to verify that the extract, transform, and load (ETL) procedure to convert from source data to the syntactically harmonized CDM has been completed as expected. Formats for all values will be assessed and compared to a list of acceptable formats. Frequency tables of variables with finite allowable values will be created to identify unacceptable values. Distributions of date variables to assess any rounding will be constructed.
The Level 1 checks proceed as follows for each table of interest in the CDM:

1. Within the METADATA table of the CDM, check for presence of the table of interest in the instance.
2. Verify that the table is present in the directory specified by the DAP. If the table is not present, print a notification of its absence to the report.
3. Verify that mandatory variables are present and contain data. If a mandatory variable is absent or contains only missing data, print a notification of this to the report.
4. Check that all conventions for the table of interest have been adhered to. If a convention is not adhered to, print a notification of this to the report.
5. Check consistency between listed allowable values in the METADATA table and data in the table of interest.
6. Tabulate missingness in all variables, overall and by calendar year.
7. Construct distributions of date variables.
8. Construct frequency tables of categorical variables, overall and by calendar year.

Each DAP will be responsible for running the script to complete the Level 1 checks. An R Markdown report describing results of the checks for each table of the CDM will be produced. After addressing any issues identified in Level 1 checks, DAPs may rerun the script and inspect the results. This may proceed iteratively until the DAP declares the ETL to be sufficiently complete and correct. An example R Markdown report produced using simulated data will be included as an annex to the study report.

Level 2 data checks assess the logical relationship and integrity of data values within a variable or between two or more variables within and between tables.

In the level 2 checks, we assess records occurring outside of recorded person time (i.e. before birth, after death, or outside of recorded observation periods). We will identify persons listed in the PERSONS table who do not have any associated records in the other tables of the CDM and verify that persons identified as the mother of an infant in the PERSON_RELATIONSHIPS table of the CDM have a birth date at least twelve years prior to the birth date of their identified child.

Each DAP will be responsible for running the script to complete the Level 2 checks. An R Markdown report describing results of the checks for each table of the CDM will be produced. After addressing any issues identified in Level 2 checks, DAPs may rerun the script and inspect the results. This may proceed iteratively until the DAP declares the ETL to be sufficiently complete and correct. An example R Markdown report produced using simulated data will be included as an annex to the study report.

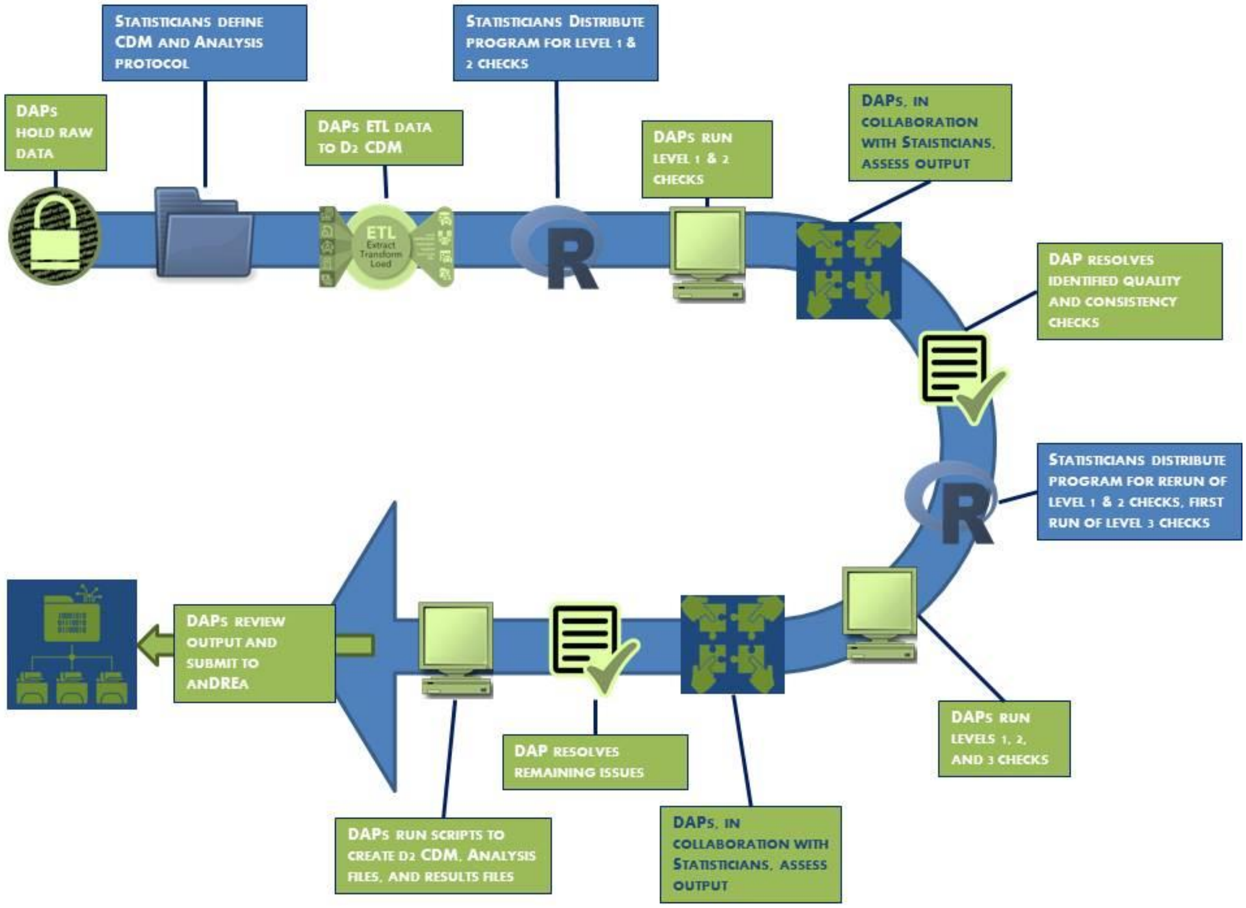
Level 3 data checks produces incidence and prevalence rates or proportions and trends over time within a data source (by examining output by age and year) for benchmarking between data sources and against external sources.

For the current study, Level 3 checks will quantify person time in each data source for the study population as a whole as well as for subpopulations of interest. These will be calculated overall and by calendar year. Additionally, counts of codes extracted to identify each event and exposure of interest will be calculated overall and by calendar year. Finally, codes will be grouped into concept sets based upon Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) as identified using the Codemapper tool (Becker et al., 2017). Counts and rates of each concept set will be calculated overall and by calendar year. Characterization summaries based upon level 3 checks will be included as an annex to the study report.

External benchmark data will be incidence rates of disease that have been obtained from the literature and are listed in the event definition form. Incidence rates from literature will be presented together with incidence rates estimated for the current study in the final study report. Discrepancies will be identified and interpreted based upon descriptions of the data source(s), algorithms for identification of events, and design choices including in and exclusion criteria in published studies vs. those employed for this protocol.

In order to identify data capture and recording issues associated with the SARS-CoV2 pandemic period, counts and rates of codes and concept sets will be calculated for the pandemic and lockdown periods of 2020 for each data source and compared graphically against the same period in the three prior years (2017-2019).

Figure 2 Data Quality Pipeline


DAPs hold raw data
Statisticians define CDM and Analysis protocol
DAPs ETL data to D2 CDM
ETL Extract Transform Load
Statisticians Distribute program for level 1 & 2 checks
R
DAPs run level 1 & 2 checks
DAPs, in collaboration with Staisticians, assess output
DAP resolves identified quality and consistency checks
Statisticians distribute program for rerun of level 1 & 2 checks, first run of level 3 checks
R
DAPs run levels 1, 2, and 3 checks
DAPs, in collaboration with Statisticians, assess output
DAP resolves remaining issues
DAPs run scripts to create D2 CDM, Analysis files, and results files
DAPs review output and submit to anDREa

11. Limitations of the research methods

11.1 Limitations related to the data sources

This study will include 10 different data sources in 7 countries which will be used to compute incidence rates of AESI, based on available and permissions in October 2020. These data sources were chosen based on availability, ability to run multisite studies and experience in using common data models plus ability to join the consortium quickly in May 2020 during a very short tender period. These data sources contain various type of data which are either representative of the national population (eg. CPRD, Nordic registries), or have a regional/multiregional scope (eg. BIFAP, SIDIAP, PEDIANET). Some data are collected at hospital level including or not emergency department or at GPs level only, others are collected at both hospital and GPs level. Given the heterogeneity in the type of encounters recorded, our analyses will be computed per data sources and no pooled estimates will be generated. The ACCESS consortium is now describing the data sources in this protocol, but a wide assessment of additional data sources and their capacity in conducting near real time monitoring, and vaccine registries is being collected and delivered in December 2020.

Some of the participating data sources in this protocol have long lag times, which means that they cannot contribute to all calendar years for the estimation of the background rates in the first analysis. Five data sources will contribute 2020 data: three will contribute hospitalization data only (ARS, FISABIO, and SIDIAP) in adults and children while PEDIANET will contribute hospitalization and GP data in paediatrics only, and CPRD will contribute GP data.

To have background rate data available prior to COVID-19 introduction the ACCESS consortium will make available as first analysis what can be obtained in October 2020, and anticipates that, if desired, updates can be made, and additional data sources can participate. Funding is currently limited to the data sources described.

Some of the data sources do not encompass to a birth registry, many do not encompass information on induced terminations and/or spontaneous abortions. Quality of information on the pregnancy start and end dates and pregnancy outcome is conditional on this availability.
Most of the data sources were characterized in the ADVANCE project and considered fit for purpose for benefits and risk assessment (Sturkenboom et al., 2020). However, no reporting of medical events in a database does not imply an absence of the event.

A broad set of AESI that are known for being related to vaccination or associated with COVID-19 will be included in this study. Some of them have a well-established clinical definition but for events such as MIS-C, ARDS, Coagulation disorders the Brighton Collaboration definition is under development by the CEPI funded SPEAC project at the time of this protocol development and may not be available at the time of data extraction and analysis. ACCESS has created a memorandum of understanding with SPEAC to exchange information rapidly. For each of the events we will use broad and narrow definitions which will be specified in the event definition forms. Additionally, several clinical case definitions for MIS-C are available which may result in heterogeneity in the reporting of this event across geographical areas. For this reason, multiple algorithms for MIS-C will be considered, ranging from sensitive (not requiring a COVID-19 diagnosis) to more specific. More generally, case ascertainment will not be conducted to confirm disease diagnosis, therefore misclassification of outcomes cannot be excluded.

Recorded disease diagnosis will be used as date to classify a case as incident. For long latency diseases (e.g., autoimmune diseases), the disease onset may have started months prior to the recorded diagnosis, however this cannot be estimated without review of records, which is not resources in this study.

Enhanced COVID-19 diseases following vaccination is a theoretical concern at the moment, and not yet shown in any of the studies. Since this event is conditioned on vaccination we cannot assess background rates during the pre-licensure vaccination period. To have some standard to measure against, we will assess COVID-19 according to five levels that are defined as: Level 1/ any recorded COVID-19 diagnosis Level 2/ hospitalisation for COVID-19 disease (suspected or confirmed); Level 3/ ICU admission related to COVID-19 disease; 4/ ARDS; Level 5/ death. The analyses described here are not intended to ascertain the incidence of COVID-19 which is not feasible as not all subjects are tested or diagnosed, but to assess time trends where possible and at least estimate the incidence of severe COVID-19 (Levels 2-4) in preparation for monitoring of enhanced disease following vaccination.

11.2 Limitations in the methodology

It is expected that healthcare behaviours will be impacted during the SARS-CoV2 circulation period due to lockdown situations in most countries. To better take into account this period, sensitivity analyses will be conducted. First, the year 2020 will be divided in two distinct periods based upon SARS-CoV2 circulation: the pre-circulation period and the circulation period of SARS-CoV2. In addition, we will consider the start of the epidemic peak at 1/10 of the highest daily number of cases reported in a country. Given that the epidemic peaks vary across countries, the peak period will be country-specific. This method is deemed appropriate knowing the high transmission rate of the infection (Zhao et al., 2020), it will detect the start of the peak period a few days after a first case was reported in a country and before the inflection point of the epidemic curve. Data to identify SARS-CoV2 circulation periods and peak periods for each of the country will be extracted from the ECDC website (https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases) and detailed in a Statistical Analysis Plan that will complement this study protocol. In addition, it is anticipated that the end of the epidemic peak may not be reached at the time of data extraction. For this reason, last data collected in 2020 will be considered in the SARS-CoV2 circulation period. Second, we will additionally conduct sensitivity analyses dividing the year 2020 into the periods prior to lockdown measures limiting face-to-face healthcare contacts, during the lockdown, and post-lockdown (dependent upon data availability) (see **Annex 3** for dates of circulation and lockdowns).

The study will also include control events to investigate potential changes in healthcare behaviours. The first control event, colonic diverticulitis, has been chosen because it may result in emergency department visit, and no biological plausibility has been identified so far with COVID-19. In addition, the occurrence of colonic diverticulitis increases with age affecting more frequently elderly (Barucha et al., 2015). This will allow assessment of trends in healthcare behaviours for a population similar to the one affected by COVID-19. In addition, several studies on the epidemiology of colonic diverticulitis have been conducted using EMR allowing for benchmarking with the use of identical medical codes (Talabani et al., 2014; Maguire et al., 2015). However, it cannot be excluded that people affected by an uncomplicated form of the disease did not seek for appropriate medical care during the pandemic. The second control event, hypertension, has been chosen because healthcare seeking for this condition or its symptoms may be delayed, and no biological plausibility has been identified with COVID-19, although hypertensive patients may be at greater risk of severe COVID-19 disease (Richardson et al, 2020).

It is acknowledged that Brighton Collaboration case definitions, which are utilized in many case definitions for the current study, are developed primarily for prospective studies, limiting their utility in observational data. However, they are a well-recognized standard and reference to develop code lists. For events without definition definitions from learned societies as well as experience of DAPs in identifying events are recorded in event definition forms and considered in algorithm development.

10. Protection of human subjects

The study will be conducted in accordance with all applicable regulatory requirements, including all applicable subject privacy requirements and the guiding principles of the Declaration of Helsinki.

The study is part of the ACCESS project which follows the framework contract stipulations of the EMA and EU PV&PE network. It also follows principles of the Vaccine monitoring Collaboration for Europe (VAC4EU) acting under a well-defined governance with articles of association and bylaws (https://vac4eu.org/governance/).
Governance approval will be obtained from DAPs to conduct the study prior to data extraction.

11. Management and reporting of adverse events/adverse reactions

The study will not evaluate medicinal products. Therefore, reporting of adverse events/adverse reactions is not applicable.

12. Plans for disseminating and communicating study results

The study protocol will be posted on the EU PAS register. Upon study completion and finalization of the study report, the results of this non-interventional study will be submitted for publication and posted in the EU PAS publicly accessible database of results. Publications will comply with the International Committee of Medical Journal Editors (ICMJE) guidelines. Following submission of the study report, a the incidence rates plus confidence intervals will be made available through a searchable, interactive Shiny dashboard. This dashboard will be made publicly available; data from individual data sources will be made visible pending DAP permission. Case counts will not be displayed to retain privacy. Results will be searchable and selectable by data source and by all strata (given sufficient cases in the strata) for primary and secondary analyses defined in this protocol. Further details will be specified in the SAP.

13. References

1. Le, T. Thanh, et al. "The COVID-19 vaccine development landscape." *Nat Rev Drug Discov* 19.5 (2020): 305-6.
2. StatNews March 2020. Available on: (https://www.statnews.com/2020/03/11/researchers-rush-to-start-moderna-coronavirus-vaccine-trial-without-usual-animal-testing/. Accessed on 24 June 2020.
3. Reuters June 2020. Available on: https://www.reuters.com/article/us-health-coronavirus-china-vaccine/cansinos-covid-19-vaccine-candidate-approved-for-military-use-in-china-idUSKBN2400DZ. Accessed on 29 July 2020.
4. SPEAC, March 2020. Available on https://media.tghn.org/articles/COVID-19_AESIs_SPEAC_V1.1_5Mar2020.pdf. Accessed on 24 June 2020.
5. Black, Steven, et al. "Importance of background rates of disease in assessment of vaccine safety during mass immunisation with pandemic H1N1 influenza vaccines." *The Lancet* 374.9707 (2009): 2115-2122.
6. Miller, E, et al. Risk of narcolepsy in children and young people receiving AS03 adjuvanted pandemic A/H1N1 2009 influenza vaccine: retrospective analysis. BMJ 346 (2013).
7. Williams, R et al. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf*. 2012. 3(2): 89–99.
8. Becker, Benedikt FH, et al. "CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project." *Pharmacoepidemiology and drug safety* 26.8 (2017): 998-1005.
9. Hence, et al. Estimation of Background Incidence Rates of Guillain-Barré Syndrome in Germany - A Retrospective Cohort Study with Electronic Healthcare Data. *Neuroepidemiology* 43:244-252 (2014).
10. Schink, Tania, et al. Risk of febrile convulsions after MMRV vaccination in comparison to MMR or MMR+V vaccination. *Vaccine* 32, issue 32: 645-650 (2014).
11. Schmidt, Morten, et al. The Danish health care system and epidemiological research: from health care contacts to database records. Clinical Epidemiology 11: 563–591 (2019).
12. Schmidt, Morten, et al. The Danish Civil Registration System as a Tool in Epidemiology. *European Journal of Epidemiology* 29, 541–549(2014).
13. Andersen, Sahl, et al. The Danish National Health Service Register. *Scandinavian Journal of Public Health* 39(Suppl 7): 34–37 (2011).
14. Sturkenboom M et al. ADVANCE database characterisation and fit for purpose assessment for multi-country studies on the coverage, benefits and risks of pertussis vaccinations. Vaccine (2020).
15. Bollaerts, Kaatje, et al. Benefit–Risk Monitoring of Vaccines Using an Interactive Dashboard: A Methodological Proposal from the ADVANCE Project. *Drug Safety* 41(8): 775–786 (2018).
16. Dodd, Caitlin, et al. Incidence rates of narcolepsy diagnoses in Taiwan, Canada, and Europe: The use of statistical simulation to evaluate methods for the rapid assessment of potential safety issues on a population level in the SOMNIA study. Plos One (2018).
17. Weibel, Daniel, et al. Narcolepsy and adjuvanted pandemic influenza A (H1N1) 2009 vaccines – Multi-country assessment. *Vaccine* 36(41): 6202–6211 (2019).
18. Zhao Shi et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCOV) in CHina, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International Journal of Infectious Diseases 92 (2020): 214-217.
19. Gini, Rosa, et al. "Data extraction and management in networks of observational health care databases for scientific research: a comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE strategies." *Egems* 4.1 (2016).
20. Gini, Rosa, et al. "Different Strategies to Execute Multi - Database Studies for Medicines Surveillance in Real - World Setting: A Reflection on the European Model." *Clinical Pharmacology & Therapeutics* (2020).
21. Barucha A et al. Temporal trends in the incidence and natual history of diverticulitis: a population-based study. *American Journal of Gastroenterology* 110(11): 1589-1596 (2015).

22. Talabani A et al. Major increase in admission – and incidence rates of acute colonic diverticulitis. International Journal of Colorectal Diseases 29:937-945 (2014).
23. Maguire L et al. Geographic and seasonal variation is associated with diverticulitis admissions. JAMA Surgery 150(1): 74-77 (2015).
24. Ulm, Kurt. "Simple method to calculate the confidence interval of a standardized mortality ratio (SMR)." *American journal of epidemiology* 131.2 (1990): 373-375.
25. Vollset, Stein Emil. "Confidence intervals for a binomial proportion." *Statistics in medicine* 12.9 (1993): 809-824.
26. Roberto, Giuseppe, et al. "Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project." *PloS one* 11.8 (2016): e0160648.
27. Gini, Rosa, et al. "Quantifying outcome misclassification in multi-database studies: the case study of pertussis in the ADVANCE project." *Vaccine* (2019).
28. Richardson, Safiya, et al. "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area." *Jama* (2020).

14. Annexes

Annex 1. Syntactically Harmonized Common Data Model

METADATA TABLES
The metadata tables contain data in a machine readable format which allows for processing of the data in the CDM.

PRODUCTS
Listing of national product codes for medicinal products. Contains a product ID foreign key to the DRUGS and VACCINES table. The PRODUCT_CODE table contains detailed data on products at the package level.

METADATA
The metadata table contains indicators which can act as machine readable guides for code written against the CDM. For instance, whether data in the drug table represents prescription or dispensing.

INSTANCE
The instance table contains data on the specific instance of the ConcePTION CDM, such as tables and columns from source data which have been included.

CDM_SOURCE
Contains high-level meta data describing the source data for the current instance such as the name of the source, data access provider, and date of last update.

CURATED TABLES
Curated tables differ from the other tables of the CDM in that data access providers are asked to create these tables using rule-based algorithms. These tables therefore represent a *syntactic* and *semantic* harmonization.

PERSON
One row of data per subject present in the data and meeting inclusion criteria for the CDM instance at any point during the study period. Data on each subject includes sex at the date of the instance creation, one date of birth, and one date of death (these may be derived using DAP-specific rules)

OBSERVATION_PERIODS
One row per period during which a subject is present in the data source. This may be based upon registration in a geographical area, registration in a GP practice, presence in a registry, etc.

PERSON_RELATIONSHIPS
Contains one row of data for each child present in the data and meeting inclusion for the CDM instance at any point during the study period, together with an identifier for the mother of the child and the father of the child if available.

ROUTINE HEALTH DATA TABLES
Routine health care data tables capture data observed in the course of routine health care in hospitals, GP offices, pharmacies, outpatient clinics, etc.

VISIT_OCCURRENCE
Contains an identifier of a visit to allow for linkage of diagnoses, procedures, dispensings, etc in the same visit if this information is available in a data source.

EVENTS
Contains data on events indicated by a diagnosis code or free text. It contains one row per diagnosed event.

MEDICINES
One record per prescription or dispensing. Contains data required to estimate duration of exposure.
Linkage to PRODUCT_CODE table to access data on drugs at the package level.

PROCEDURES
Contains data on procedures ordered or completed. For those procedures with an associated result, results and units are recorded. It contains one row per procedure.

MEDICAL_OBSERVATIONS
Contains observations recorded during routine healthcar. Can be a result from a laboratory test, or physical measurement, but also level of education, or sex, or a pathology report.

SURVEILLANCE TABLES
Surveillance tables contain data collected for purposes beyond routine health care either for surveillance of specific events or for recording of detailed information related to a unit of observation such as a pregnancy or chronic illness.

SURVEY_ID
Contains metadata on observations contained in the SURVEY_OBSERVATION table and allows for linkage between mothers and infants captured in a medical birth registry.

SURVEY_OBSERVATION
Contains one row per observation in any survey or registry data table – such as a medical birth registry, well child program database, cancer registry, etc.

Full CDM specifications can be accessed here:
https://drive.google.com/file/d/1hc-TBOfEzRBthGP78ZWIa13C0RdhU7bK/view?usp=sharing

Associated CDM vocabularies can be accessed here:
https://docs.google.com/spreadsheets/d/1idAEKC440rkIYIxCSRmEVgEPj_UouUI-I3kxNCpJt3U/edit?usp=sharing

Annex 2. Event definition form template

ACCESS
vACCine covid-19 monitoring readinESS

EVENT DEFINITION FORM

Event:
Outcome/covariate:
Version:
Status:

Contributing authors

authors	Role

Contents

1. Event definition
2. Synonyms / lay terms for the event
3. Laboratory tests that are specific for event
4. Diagnostic tests that are specific for event
5. Drugs that are used to treat event
6. Procedures used specific for event treatment
7. Setting (outpatient specialist, in-hospital, GP, emergency room) where condition will be most frequently /reliably diagnosed.
8. Diagnosis codes or algorithms used in different papers to extract the events in Europe/USA: seek literature for papers that have studied this event, and see how they extracted/measured the event.
9. Experience of participating datasources to extract the events prior to ACCESS.
10. Proposed codes by Codemapper
11. Algorithm proposal(s)
12. Background rates
13. References

Annex 3. Dates of COVID-19 pandemic response measures per country

Country	Data Access Provider	Name Data Source	Date of country's first detected case of COVID-19 in	Date at which 1/10 of the highest daily number of cases was	Start date of lockdown measures limiting face-to-face	End date of lockdown measures limiting face-to-face healthcare encounters
Germany	BIPS	GePaRD	28 January	13 March	March (exact date varies between federal states)	June (exact date varies between federal states)
Netherlands	PHARMO	PHARMO	28 February	14 March	23 March	11 May
Denmark	Aarhus University	Danish Registries	27 February	10 March	12 March	Healthcare encounters are still limited for people with respiratory tract infection sympotms. Patients with symptoms of possible COVID-19 have to call the GP asking for authorization to go to the medical centre to be tested.
Spain	AEMPS	BIFAP	01 February	11 March	14 March	Goverment lockdown measures were eased on 21 June. Healthcare encounters are still limited. Patients have to call the GP asking for authorization to go to the medical centre (as of 11 August 2020)
Spain-Valencia	FISABIO	FISABIO	01 February	11 March	14 March	Goverment lockdown measures were eased on 21 June. Healthcare encounters are still limited. Patients have to call the GP asking for authorization to go to the medical centre (as of 11 August 2020)
Spain-Catalunya	IDIAP-Jordi Gol	SIDIAP	01 February	11 March	14 March	Goverment lockdown measures were eased on 21 June. Healthcare encounters are still limited. Patients have to call the GP asking for authorization to go to the medical centre (as of 11 August 2020)
Italy	SoSeTe	PEDIANET	31 January	06 March	07 March	18 May

Italy	ARS	ARS data	31 January	06 March	07 March	18 May
United Kingdom	Utrecht University	CPRD	31 January	22 March	26 March	01 June
Norway	University Oslo*	Norwegian registries	27 February	11 March	12 March	20 April
France	BPE**	SNDS	24 January	14 March	17 March	11 May

[$] Data extracted from ECDC website (https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases)

* Norway will participate in providing information but cannot do background rates in time

**BPE is a subcontractor of VAC4EU

Annex 4. Event Definition Forms

Event definition forms have been drafted for each of the AESI. These are living documents in a google drive. https://drive.google.com/drive/folders/1Y_3cuGRN1g-jBv2ec1fC0aYcpxEjtrY9?usp=sharing

We provide links to allow for continuous updates.

Auto-immune diseases

1.Guillain-Barré Syndrome
2.Acute disseminated encephalomyelitis
3.Narcolepsy
4.Acute aseptic arthritis
5. Diabetes
6.Idiopathic Thrombocytopenia

Cardiovascular system

Acute cardiovascular injury including:

7.Microangiopathy,
8.Heart failure,
9.Stress cardiomyopathy,
10.Coronary artery disease,
11.Arrhythmia,
12.Myocarditis

Circulatory system

13.Coagulation disorders: Thromboembolism, Haemorrhage disease
14.Single Organ Cutaneous Vasculitis

Hepato-gastrointestinal and renal system

15.Acute liver injury
16.Acute kidney injury

Nerves and central nervous system

17.Generalized convulsion
18.Meningoencephalitis

Respiratory system

19. Acute respiratory distress syndrome

Skin and mucous membrane, bone and joints system

20.Erythema multiforme
21.Chilblain – like lesions

Other

22. Anosmia, ageusia
23.Anaphylaxis
24.Multisystem inflammatory syndrome in children
25.Death (any causes)

26. COVID-19 Enhancement of disease
27. Sudden death

Pregnancy outcome - Maternal

28. Gestational Diabetes
29. Preeclampsia
30. Maternal death

Pregnancy outcome - Neonates

31. Fetal growth restriction
32. Spontaneous abortions & Stillbirth
33. Preterm birth
34. Major congenital anomalies
35. Microcephaly
36. Neonatal death
37. Termination Of Pregnancy for Fetal Anomaly

Newly added Sept 21 :

38. transverse myelitis