Statistical analysis plan

Covid-Vaccine-Monitor

Rapid Safety Assessment of SARS-CoV-2 vaccines in EU Member States using electronic health care data sources

Specific Contract 01 implementing FWC EMA/2018/23/PE (Lot 3)

V1.2

Version	**Date**	**Authors**	**Comments**
0.1	January 19, 2022	Miriam Sturkenboom	First draft for readiness based on protocol v2.2
0.2/0.8		Susana Perez Gutthan, Sophie Bots, Carlos Duran, Anna Schultze, Svetlana Belitser, Bradley Layton, Xavier Garcia, Olaf Klungel	Comments and additions of WP4 sections in section 7
0.9	April 3, 2022	Miriam Sturkenboom	Inclusion of comments, adding section 8 and inclusion of appendices
		Anna Schultze, Svetlana Belitser, Ivonne Martin, Daniel Oberski	Update section 4.4 and 4.5
1.0	April 8, 2022	Miriam Sturkenboom	Finalization of comments
1.1	June 13, 2022	Miriam Sturkenboom, Susana Perez Gutthan, Sophie Bots, Carlos Duran, Anna Schultze, Svetlana Belitser, Bradley Layton, Xavier Garcia, Olaf Klungel, Ivonne Martin, Fabio Riefolo	Inclusion of EMA comments
	July 31, 2023	Rosa Gini	Updating code books
1.2	August 13, 2023	Miriam Sturkenboom	Cleaning for public release

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
CI	Confidence interval
DAP	Data Access Provider
DRE	Digital Research Environment
EMA	European Medicines Agency
EMR	Electronic Medical Records
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance.
ETL	Extract, Transform, and Load
EU PAS	The European Union electronic Register of Post-Authorisation Studies
GDPR	General Data Protection Regulation
GP	General Practitioner
GPP	Good Participatory Practice
HIV	Human Immunodeficiency Virus
ICD	International Classification of Diseases
ICMJE	International Committee of Medical Journal Editors
ICU	Intensive Care Unit
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
VAC4EU	Vaccine monitoring Collaboration for Europe

Table of Contents

1.Title

Rapid Safety Assessment of SARS-CoV-2 vaccines in EU Member States using electronic health care datasources

2. Marketing authorisation holder

Not applicable
This protocol has been developed by the EU PE&PV research network EMA/2017/09/PE (Lot 3, SC01): Rapid Safety Assessment of SARS-CoV-2 vaccines in EU Member States using electronic health care datas ources.

3. Responsible parties

University Medical Center Utrecht, The Netherlands
Prof. dr. Miriam Sturkenboom
Dr. C Duran
Drs. V Hoxhaj
Dr. Daniel Oberski
Prof. dr. Eijkemans
Drs. R Elbers
Drs. E Alsina
Dr. Fariba Ahmadizar

Key collaborators and roles in CVM study

Name	**Organization**
Prof. dr. Miriam Sturkenboom, Dr. D Weibel, Drs. V Hoxhaj, Dr. Daniel Oberski, Dr. Ivonne Martin, Prof. dr. Eijkemans, Dr. R van de Bor, Drs. R Elbers, Drs. E Alsina, Drs. MM Sisay, Dr. Fariba Ahmadizar	University Medical Center Utrecht, Utrecht, The Netherlands (UMCU)
Prof. dr. Olaf Klungel, Dr. Patrick Souverein, Dr. Sophie Bots, Drs. Svetlana Belitser, Dr. Satu Johanna Siiskonen	Universiteit Utrecht (UU), Utrecht, The Netherlands CPRD data
Dr. Daniel Weibel, Dr. Patrick Mahy, Dr. Angela Bastidas	VAC4EU
Dr. Rosa Gini, Claudia Bartolini, Davide Messina	Agenzia Regionale di Sanitá Toscana (ARS), Data Tuscany region
Dr. Ursula Kirchmayer, Valeria Belleudi	Department of Epidemiology, Lazio Regional Health Service, ASL Roma1 (DEP Lazio)
Dr. B. Layton, Dr. X. Garcia de Albeniz, Ms. Estel Plana Dr. A. Arana, , Dr. Alison Kawai, Ms Lia Gutierrez, Dr. J. Fortuny, Rachel Weinrib, Dr. S. Perez-Gutthann	RTI Health Solutions (RTI-HS), Spain & US
Prof. Dr. RMC Herings, Jetty Overbeek, Karin Swart	PHARMO Institute - PHARMO Database Network
Prof. Dr. Gianluca Trifirò, Ylenia Ingrasciotta, Valentina Ientile	INSPIRE srl, Messina; Caserta database
Prof. Hedvig Nordeng, Dr. Angela Lupattelli	University of Oslo (UiO), Norway; Norwegian data
Dr. Mar Martin, Dr. Patricia Garcia-Poza, Dr. Consuelo Huerta, Dr. Dolores Montero, Maria Martinez, Airam de Burgos	Spanish Agency on Medicines and Medical Devices (AEMPS) - BIFAP database
Prof. Ian Douglas, Dr. Anna Schultze	London School of Hygiene and Tropical Medicine (LSHTM): method
Eva Molero, Dr. Fabio Riefolo	Team-it Institute SL (TEAMIT), coordination
Felipe Villalobos, Boni Bolibar	IDIAP Jordi Gol, SIDIAP data
Ainara Mira-Iglesias, Juan José Carreras-Martínez, Jorge Arasa, Mónica López-Lacort, Elisa Correcher Martínez, Valle Morales-Cuenca, Javier Díez-Domingo	Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO) – Valencia health system Integrated Database (VID)

Sponsor:

This Statistical Analysis Plan has been developed as a deliverable of the framework contract No EMA/2018/28/PE (SC01, Lot 3) with the European Medicines Agency.

4. Introduction

EMA's mission is the protection and promotion of public and animal health, through the evaluation and supervision of medicines for human and veterinary use.

COVID-19 vaccines in the EU are evaluated by EMA via the centralised procedure, based on a rolling review. While a large number of COVID-19 vaccines are still progressing in clinical development, four vaccines (from Pfizer/BioNTech, Moderna, AstraZeneca, Janssen) have been granted conditional marketing authorisation. While more vaccines are expected to be authorised in 2021, large-scale vaccination campaigns are being rolled out across the EU, with tens and perhaps hundreds of millions of EU citizens expected to be vaccinated in 2021 and 2022.

Multiple vaccine products are being used at the national level, many of them based on novel technologies, with safety experience limited to pre-licensure clinical trials. Therefore, there is a public health need for comprehensive safety surveillance. Real-world safety monitoring of COVID-19 vaccines through observational studies should be implemented across Europe in a multi-layer approach by (i) Member States (ii) vaccine manufacturers and (iii) the Agency, to complement its routine pharmacovigilance activities.

5. Goal and objectives

5.1 Goal

In order to complement spontaneous reporting systems for signal detection (routine pharmacovigilance) and other initial safety monitoring activities such as pharmaco-epidemiological studies conducted or planned by different stakeholders, the Agency procured an early safety monitoring study through its framework contracts (Early-Covid-Vaccine-Monitor; EUPAS39798) which is conducted by the EU PE&PV research network and VAC4EU.

The **goal** of the COVID-Vaccine Monitor study is to rapidly assess **signals of** potential safety concerns emerging from active surveillance and identified by PRAC. Rapid signal assessment means the collection of additional information in order to further characterise the incidence of the safety concern in comparison to its expected incidence in **non-vaccinated populations or with suitable active comparator populations**.

5.2 Objectives

The objectives are divided into two phases, the first phase is the readiness phase. This will be conducted by all 10 DAPs, and be the basis for the selection to participate in real studies as well as the basis for the methods group (WP4) to assess the impact of methodological choices and assumptions using the study designs with negative controls.

Readiness phase conducted with all DAPs

The readiness phase will include the following objectives:

- To provide an overview of the methods and results for identification of COVID-19 vaccine exposure in the data sources

To monitor the number of individuals exposed to any COVID-19 vaccine and to compare this to COVID-19 vaccine exposure (benchmark):

https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html#uptake-tab).

- To quantitatively evaluate different algorithms to identify adverse events by provenance in electronic health care data

Methodological work

- To conduct time-to-onset analyses for the AESI with respect to time since vaccination
- To assess the association between the vaccines of interest and negative control using the SCRI to estimate systematic bias (unmeasured confounding), this will be performed by methods WP4
- To test the impact of (by WP4)
    - different comparators in the cohort design, by using the negative control outcomes
    - different censoring criteria in the cohort study
    - different control periods/duration for the SCRI
    - different algorithms to assess vaccine exposure (doses), events, and covariates

based on the analysis of negative control outcomes and quality checks by WP4.

Rapid assessment studies requested by EMA with selected number of DAPs

Primary objective

The primary objective for this rapid assessment study is to assess the potential association between the occurrence of specific AESIs and vaccination with COVID-19 vaccines within disease-specific risk periods in individuals exposed to the COVID-19 vaccines compared to other COVID-19 vaccine exposed individuals, or compared to a control window within the same individual.

Secondary objectives

The secondary objectives for rapid assessment studies are:

- To assess the potential association between the occurrence of specific AESIs and vaccination with COVID-19 vaccines in the following subgroups:

    - immunocompromised persons
    - persons with the presence of co-morbidities elevating the risk of serious COVID-19
    - persons with a history of diagnosed COVID-19 disease
    - age groups (<20 and 10-year age categories))
    - patients with a prior history (ever) of that event more than a year before.

Gender

- To conduct sensitivity analyses requested by methods group (WP4)

The following VAC4EU and/or EU PE&PV research network data access providers were invited to participate in the readiness proposal

Table 1 Participating data access providers and data sources

N	Country	Data Access Provider	Name Data source	Experience ConcePTION CDM v2.2	AESI experience	Active population	Type of data source
1	NL	PHARMO / UMCU	PHARMO	Yes	Yes (ACCESS)	6 million	Record linkage
2	ES	AEMPS	BIFAP	Yes	Yes (ACCESS)	10 million	GP medical records
3	ES	IDIAPJGol	SIDIAP	Yes	Yes (ACCESS)	5.7 million	Record linkage
4	ES	FISABIO	VID	Yes	Yes (ACCESS)	5 million	Record linkage
5	IT	SoSeTe	PEDIANET	Yes	Yes (ACCESS)	0.5 million	Pediatric medical record
6	IT	ARS Toscana	ARS data	Yes	Yes (ACCESS)	3.6 million	Record linkage
7	IT	Lazio	Lazio data	No	No	5.8 million	Record linkage
8	IT	INSPIRE srl	Caserta data	No	No	1 million	Record linkage
9	UK	Utrecht University	CPRD/HES GOLD	Yes	Yes (ACCESS)	16 million	GP & Hospital medical record
10	NO	University Oslo	Norwegian	Yes	No	5 million	Record linkage

6. Research methods

6.1 Study Design

The study comprises a readiness phase, to assess whether the data source is fit for the purpose of vaccine studies. The pool of data sources that are ready can then be utilized for specific rapid assessment studies when requested by EMA.

6.1.1 Readiness phase

The readiness phase study period starts follow-up on 1 January 2019. The primary design is a cohort study including all subjects with at least one day of follow-up after January 1, 2019, and at least 365 days of availability prior to that date, unless the date of birth occurred was during 2019-2021.

In the readiness phase, data sources

- Prepare the ETL design for the transformation of local data into the ConcePTION CDM (CCDM)

- Run level 1-3 quality checks on data required for all AESI and covariates, aiming at investigating the completeness (level 1), the logic of the converted data (level 2), and subsequently whether the data is fit for purpose, especially as regards vaccine and events data.

    - Level 3 checks include the generation of **incidence rates for the events** and covariates (2019-2020) by age and gender. This data may also be utilized to further understand misclassification of outcomes and exposure by the methods group in WP4 task 4.5.

    - Level 3 checks also include renewed verification of vaccine uptake and timing

    - Additional readiness assessment include: vaccine uptake, characteristics of vaccinated, and incidence rates of AESI during 2019 and 2020 (see shell tables section 10)

- Conduct the cohort and SCRI study designs with vaccine-negative control outcome pairs in collaboration with WP4 to:

    - develop and run analytical R-code that is used for the negative control outcomes and can be re-used for the rapid assessment studies and its sensitivity analyses

    - To assess systematic bias and generate information to assess the methodological developments by WP4 that may be incorporated in sensitivity analyses of rapid assessment studies

        - control window duration and timing in the SCRI (pre-post control window)
        - residual confounding assessment and impact of different comparators such as
            - contemporary unvaccinated comparators matched on calendar time and other factors
            - contemporary unvaccinated comparators, time zero randomly sampled
            - subjects vaccinated with a different COVID-19 vaccine.
        - outcome misclassification (using different algorithms for events)
        - censoring (left and right)

6.1.2 Overview of Study Design for rapid hypothesis testing (assessment) studies

Rapid assessment of safety concerns is conducted using a retrospective observational study using electronic health care databases that have gone through the readiness phase. Eligible individuals will be included in the study from the start of vaccination campaigns: 1 December 2020, and the study will end at the last date of data availability in each database.

For specific events of concern, the study design depends on whether the event is considered acute or non-acute and follows the decision framework described in the ACCESS template protocols (EUPAS 39361).

The primary study design for acute events (events expected to occur within 60 days of vaccination) will be a self-controlled risk interval (SCRI) design and for non-acute events (events expected to occur or be diagnosed with delay, within 180 days) a cohort design with contemporary exposed (vaccinated) comparators. Acute events may also be studied using the cohort design to address uncertainties around risk windows and limitations of the SCRI design. In the SCRI design, a risk window is compared to an unexposed pre-vaccination control window within each person. Subjects start follow-up at time zero (time of vaccination or the start of the pre-vaccination control window for the SCRI) and end follow-up at the earliest of occurrence of latest data availability of the databank, subject exit, the completion of the period, or death. At least one year of enrollment/ presence prior to time zero (cohort entry) will be required to determine whether individuals meet the study criteria and to define baseline characteristics, unless the persons is born during the study period.

6.1.2.1 Self-controlled Risk Interval Design

Self-controlled studies are commonly used to evaluate the safety of medicinal products, including vaccines, as they do not require an external control group to be identified and control for all time-fixed confounding by design. A range of self-controlled designs have been employed in vaccine epidemiology, including self-controlled case series (SCCS) and self-controlled risk interval (SCRI) designs. An SCRI is a special case of the SCCS, in which the baseline time is fixed for some short period of time in relation to the administration of the vaccine, for example, the 60 days preceding the vaccination, whereas in an SCCS all time which is not designated a risk window is used as control time. The definition of the control window is an important part of the design of self-controlled studies, and can impact findings for several reasons. A specific concern in the study of COVID-19 vaccines is that the use of a pre-vaccination control window may introduce bias, as many safety events significantly reduce the probability of a person being vaccinated. Such “event-dependency of the exposures” can lead to overestimation of any safety signal. Although short pre-exposure windows may be used to account for this when bias is not severe, this will not address the issue if the probability of receiving the vaccine is permanently altered after the outcome. In those scenarios, an alternative option is to use only post-vaccination time as the control time, though the trade-off here is that rapid studies are less feasible as they require longer accrued follow up in each database. In addition to potentially impacting the extent of bias in analyses, the choice of control window might also influence the efficiency of the method. This is both because a longer control window may allow for the inclusion of a greater number of cases, and because the power of a self-controlled case series depends in part on the ratio of risk to observation time.

The SCRI design compares the risk of the event of interest in post-vaccination risk windows to a pre-vaccination control window within the same individual. We use a pre-vaccination control window to allow for rapid hypothesis testing, since data lag times may occur, we do not want to wait too long after introduction of COVID-19 vaccines to be able to analyse.

The implications of using a pre-vaccination control period will be investigated in a simulation study and empirical evaluation for the methodological development (WP4) of the project (Section 7.3, and

Protocol of WP4, Evaluation of assumptions of SCRI in simulation studies), and we will adapt the SAP based on the findings of this study. Key issues are:

- length and timing of the buffer period (to account for any healthy vaccine effect),
- contra-indication,
- death
- the use of a post-vaccination control window

These assumptions will be investigated in an ongoing simulation study as part of WP4 [Protocol of WP4, Evaluation of assumptions of SCRI in simulation study]. The SCRI design includes only individuals who received at least one dose of a COVID-19 vaccine during the study period and who experience the specific event in the control period or after vaccination (starting date of vaccination). Study subjects enter the study at the time of the start of the control window, which starts 90 days (as a default) before the date of vaccination with a COVID-19 vaccine. The SCRI design compares the risk of each outcome during the risk window following dose 1 or dose 2 with the self-matched control interval, used to assess the baseline risk of the outcome. The control period is 60 days long and is followed by a 30-day pre-exposure (buffer) period, to account for healthy vaccinee effect and potential temporary event-dependency of the exposure, the length of the pre-exposure period may be adapted based on the assessment of the methods by WP4 and the specific event of interest. Cases with an event in either the risk or control window will contribute to the estimation of the incidence rate ratio of interest. If an event occurs in the pre-exposure period, it is kept in the study to enable sensitivity analyses.

The risk window post-vaccination starts at day 1 and is divided into dose-specific risk intervals following each dose of the COVID-19 vaccine, except for anaphylaxis for which the risk interval starts at day 0. If a second dose is given within the risk interval of the first dose, the period of follow-up for the first dose will be censored. Sensitivity analyses will be conducted that include day 0 in the risk interval. As calendar time is suspect to be a strong confounder, analyses will adjust for calendar time in 30-day intervals.

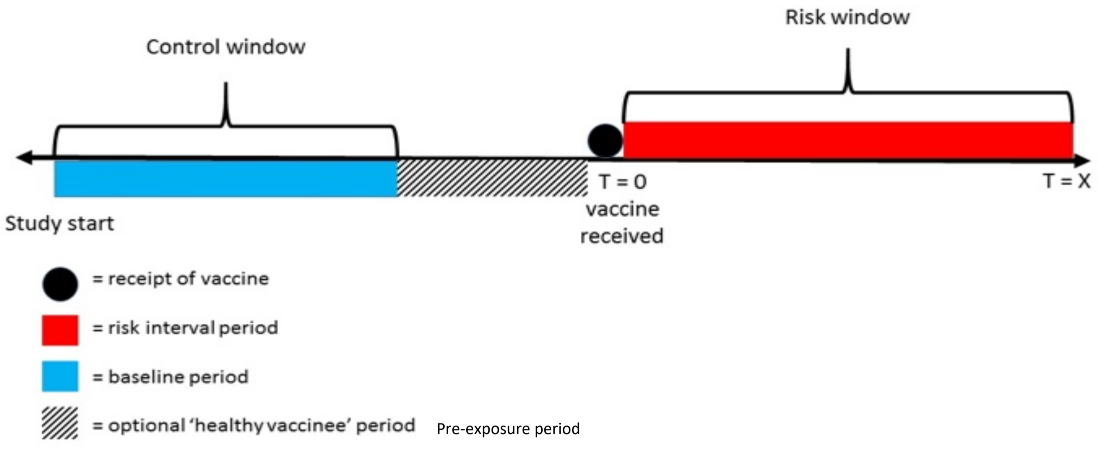


Figure 1: Self-Controlled Risk Interval Design.

6.1.2.2 Cohort Design with Concurrent comparator

A retrospective cohort design is used to estimate the rate of non-acute events of interest after receipt of COVID-19 vaccination dose and compare this incidence primarily with that

occurring in a COVID-19 vaccinated matched comparator group. Additional comparators may be tested for the methods work (e.g. concurrent individuals vaccinated with another COVID-19 vaccine, concurrent unvaccinated individuals, with time zero either randomly sampled or matching) using the negative control events in the readiness phase.
For sensitivity analysis of acute events that are analysed using an SCRI approach, the same cohort design may be chosen, if there is uncertainty about the risk window or direct comparison between different vaccines is needed.

Exposed cohort (index cohort): individuals who have received at least one dose of a specific COVID-19 vaccine.

Concurrently exposed cohort (reference cohort): individuals that have been vaccinated with another type of COVID-19 vaccine.

In this retrospective cohort design, time zero (cohort entry) is defined as the time at which the exposure status is assigned, when selection criteria are applied, and when study outcomes start to be counted. Time zero (ie, recipients of the vaccine) is the day the specific COVID-19 vaccination (index cohort) was received for anaphylaxis and date of vaccination +1 for other events of interest.

Concurrently unexposed cohort (reference cohort): individuals who have not received a COVID-19 on or before time zero.

In the concurrent unvaccinated comparator group with random time zero, all eligible individuals are included on a calendar date on which they are unvaccinated. Calendar time and risk factors will be balanced using propensity score methods (weighting) specific for each outcome when needed.

In the concurrent unvaccinated comparator group with matched time zero, persons in the index cohort are individually matched to one individual in the concurrent cohort on key clinical variables (exact age, sex, and presence of one or more risk factors for severe COVID-19 [e.g. cancer, sickle cell, obesity, chronic kidney disease, chronic respiratory disease, human immunodeficiency virus infection]) at time zero. In case there is no balance in several variables these may be included in a propensity score. The effect of propensity score weighting will be considered, if matching failed to achieve balance. Individuals will be classified into exposure groups that are compatible with their data at time zero.

Two concurrently unexposed cohorts will be considered. While daily matching on calendar time will balance calendar time between groups, accounting for outcome seasonality, pandemic health care access, and vaccination prioritization, there is considerable computational burden involved in the separate daily matching algorithms which each consider large numbers of individuals, often the full source population. This computational burden may impede the estimation of the variance via bootstrapping, which may be needed for certain effect estimates, (e.g., risk differences based on the Kaplan Meier estimator). Thus, the performance of a more streamlined approach which randomly selects a sample of the eligible time zero of unvaccinated individuals will also be evaluated; calendar time and personal characteristics will be subsequently balanced with propensity score weighting.

6.2 Setting, Study Population and follow-up

6.2.1 Study Setting

For the implementation of the readiness study, 10 electronic health care databases in Northern, Southern and Western Europe that have shown interest, are used. The data sources that were included are those who have been working in prior studies (EU PE&PV or VAC4EU) and were interested to participate

Italy
- ARS Toscana (Agenzia Regionale di Sanità della Toscana)
- Lazio region, Department of Epidemiology
- Pedianet (Societa Servizi Informatici)
- Caserta local health database (INSPIRE srl)

Netherlands
- PHARMO Database Network (PHARMO Institute for Drug Outcomes Research) (NL)

United Kingdom
- CPRD (Clinical Practice Research Datalink) & HES data (UK)

Norway
- The Norwegian health registers

Spain
- SIDIAP (Sistema d'Informació per el Desenvolupament de la Investigació en Atenció Primària)
- BIFAP (Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria)
- FISABIO (VID, Valencia health system Integrated Database)

Further information on the data sources used in this study can be found in Section 6.4. All data sources are participating in the readiness phase.

For actual rapid assessment studies, choices are made based on:

- Availability of fit for purpose data
- Sample size and resources
- Ability to commit to timelines

6.2.2 Source population

The source population comprises all individuals registered in each of the participating healthcare data sources.

6.2.3 Study Duration and Follow-Up

Readiness

For the readiness phase study, the study period starts on 1 January 2019 and ends on December 31st, 2021 or latest if possible. Subjects are followed from 1 January 2019 until the earliest of the following dates: death, end of data availability, subject exit, or the completion of the period. If persons have multiple periods within the same data source, we only use the period in which the first COVID-19 vaccine was provided as active follow-up.

Rapid hypothesis testing study

For the SCRI, the study period starts on 1 September 2020 and lasts until the end of the study period. For the cohort study the study period starts at December 1st, 2020.

SCRI: Follow-up ends at the earliest of the following: end of data availability, subject withdrawal of the data sources, end of the duration of the risk period. For death and fatal events, specific additional criteria may be posed.

Cohort: The cohort design follow-up ends at occurrence of each AESI, or censoring at death, end of data availability, subject exits the database or recommended end date (as per DAP decision, based on an assessment of the validity of the data). For unvaccinated groups, individuals will be censored when they receive a COVID-19 vaccine dose. For vaccinated groups, individuals will be censored when they receive a COVID-19 vaccine dose of a different brand than the one received on time zero.

6.2.4 Inclusion Criteria

6.2.4.1 Readiness study

For the readiness study, the person is included if there is at least one day of follow-up and the person has at least 12 months of data in the data source at the start of follow-up or is born during 2019-2020.

6.2.4.2 SCRI Design

For analyses of outcomes assessed with the SCRI design, the following criteria must be met. Note that the study population for each outcome-specific analysis may thus be different.

- Received at least one dose of COVID-19 vaccine during the study period.
- Has experienced the specific event of interest during the predefined observation period.
- Has at least 12 months of data/registration in the data sources at study entry (except when born during study period)

6.2.4.3 Cohort design

Individuals must meet all the following inclusion criteria to be eligible for inclusion in the cohort study:

- At time zero, being in the underlying population of the data source for at least 12 months; or, being born in the previous 12 months in the underlying population.
- No history of vaccination with a COVID-19 vaccine before time zero

6.2.5 Exclusion Criteria

For the readiness study, there are no exclusion criteria. Individuals are excluded from the hypothesis testing studies if:

- They have a recorded diagnosis for the specific event in the 365 days prior to cohort /SCRI entry (time zero). Persons with such acute diagnoses more than a year ago will be maintained to allow for subgroup analyses. Upon investigation of one event, we do not exclude any history or prevalence of other groups of events (AESIs).
- They have a contra-indication for one of the COVID-19 vaccines.

6.3 Variables

6.3.1 Exposure Assessment

Exposure will be based on available recorded prescription, dispensing, or administration of the COVID-19 vaccines. Vaccine receipt and date of vaccination will be obtained from all possible sources that capture COVID-19 vaccination, such as dispensing records, general practice records, immunisation registers, vaccination records or other data banks. The main exposure of interest for the rapid assessment studies is the receipt of COVID-19 vaccine.

- ARS Toscana (IT): ARS will identify vaccines from the regional immunization register using the nationally used product code, including batch number.

- Pedianet (IT): Information on COVID-19 vaccine includes the date of immunisation, type of vaccine, vaccine batches, dose.

- Lazio (IT): DEP Lazio identifies vaccines from the regional immunization register using the nationally used product code, including batch number.

- PHARMO (NL): Data on vaccination are obtained from PHARMO's GP database. Information on vaccines include ATC code, brand, batch, and date of administration/recording. Several COVID-19 vaccines have been administered through other routes and original immunization data are not yet linked with GPs, this may change in the future.

- Caserta LHU database (IT): Caserta LHU record linkage database contains information from all claims databases (e.g. hospitalizations, drug dispensing, etc.) of Caserta province catchment area (around 1 million population). In addition, those claims data can be linked to the local immunization registry which includes name and batch of the vaccine; manufacturing company; dose; administration route; administration location (eg, general practice); date of administration.

- CPRD (UK): The CPRD contains information recorded by National Health Service (NHS) primary care general practitioners (GPs); and information on the administration of COVID-19 vaccines to individuals is available. This includes, alongside an encrypted unique patient identifier; the name of the vaccine; manufacturing company; dose; stage of the vaccine schedule; administration route; administration location (eg, general practice); batch identifiers/numbers; date of administration; and GP prior to, on, or after the vaccination date. In addition, patient demographic, practice-level, and staff-level information will also be available.

- Norwegian health registers (NO): The national, electronic immunisation register (SYSVAK) was established in 1995 and records an individual's vaccination status and vaccination coverage in Norway. All vaccinations are subject to notification to SYSVAK and are registered without obtaining patient consent. This applies to all COVID-19 vaccines. In SYSVAK, the following data are registered: individual personal identifier, vaccine name and Anatomical Therapeutic Chemical (ATC) code, vaccine batch number, date of vaccination, reason for vaccination as health care professional versus risk-group patient, and the centre where the vaccine was administered.

- SIDIAP (ES): SIDIAP has available information on the administration of COVID-19 vaccines to individuals linked to a unique and anonymous identifier. The information will be originated from the electronic medical records. For each patient, SIDIAP will have date and centre of administration, health professional administering the vaccine, dose, brand, reasons for vaccination (eg, risk group), and other information related to vaccination.

- BIFAP (ES): 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). Data on vaccination with COVID-19 vaccines are obtained from the COVID-19 vaccination registries in the participating regions and linked to the primary care medical records in BIFAP. Date of vaccination, brand, batch, and dose are registered.

- FISABIO (ES): Data on vaccine exposure will be obtained from the Vaccine Information System (VIS), which includes information on vaccine type, manufacturer, batch number, number of doses, location and administration date.

The vaccination strategies for the different exposure groups are defined as follows:

- Subjects who receive a first dose of a specific COVID-19 vaccine are classified as exposed to D1 for that specific vaccine (if brand is unknown it will be unknown).

- In the SCRI design subjects who receive a second, third or fourth dose of COVID-19 vaccine will only contribute time to the prior dose/brand risk window, and move into the risk window of the next dose for a COVID-19 vaccine, by brand for both the cohort as well as the SCRI design, once this occurs.

In the cohort study the vaccination strategy for the matched reference cohort(s) are defined at time zero based on the Dose 1 as:

- Pfizer
- Moderna
- Janssen
- AstraZeneca
- Novavax
- Unknown

depending on the type of first COVID-19 vaccine. For the unknown category we will assess with the WP4 group whether imputation is possible based on the vaccination role out time, and vaccination group.

In the readiness study on negative control outcomes, different types of comparator cohorts will be tested for assessment of impact (contemporary unvaccinated comparators with time zero sampled at random, contemporary unvaccinated comparators, matched on calendar time, subjects vaccinated with a different COVID-19 vaccine.

For the SCRI design, person-time in the risk interval will be considered “exposed” while person-time in the control interval will be considered “unexposed.” Risk intervals are specific to the outcome of interest.

6.3.2 Study Outcomes

AESIs, as listed below (Table **2**) and in line with the definitions and code lists that have been created for the ACCESS and used and finetuned in ECVM & VAC4EU PASS projects (as stored in the VAC4EU Sharepoint), are used, with a date of diagnosis. Algorithms including combination of different events and/or medicines will be used for some of the covariates and outcomes. Zenodo provides the code lists for each of the different vocabularies that are used.

During the readiness phase, the impact of the provenance of information on outcomes, as well as different algorithms, is assessed by WP4. In case a new signal arises, the protocol may be amended and the new event may need to be included. For new events, level 3 checks (including incidence rates) will be conducted.

Table 2 List of AESI and the negative control events, design and primary risk period duration

Event	ACCESS	SCRI	cohort	Naïve period to estimate new onset	Primary Risk period*
Multisystem inflammatory syndrome	✓	✓	✓	365 days	28 days
Acute respiratory distress syndrome	✓	✓	✓	365 days	28 days
Acute cardiovascular injury	✓	✓	✓	365 days	
Microangiopathy	✓	✓	✓	365 days	28 days
Acute CAD	✓	✓	✓	365 days	28 days
Arrhythmia	✓	✓	✓	365 days	28 days
Myocarditis	✓	✓	✓	365 days	28 days
Pericarditis	✓	✓	✓	365 days	28 days
Coagulation disorders, including deep vein thrombosis, pulmonary embolus, cerebrovascular stroke, limb ischaemia, haemorrhagic disease	✓				
VTE (DVT & PE & Splanchnic)	✓	✓	✓	365 days	28 days
CVST	✓	✓	✓	365 days	28 days
Arterial thrombosis (AMI /Ischemic stroke)	✓	✓	✓	365 days	28 days
TTS (VTE, arterial thrombosis, or CVST with thrombocytopenia in 10 days)	✓	✓	✓	365 days	28 days
Hemorrhagic stroke	✓	✓	✓	365 days	28 days
DIC	✓	✓	✓	365 days	28 days
Generalised convulsion	✓	✓	✓	30 days	14 days
Guillain Barré Syndrome	✓	✓	✓	365 days	42 days
Diabetes (type 1)	✓		✓	365 days	180 days
Acute kidney injury	✓		✓	365 days	180 days
Acute liver injury	✓		✓	365 days	180 days
Anosmia, ageusia	✓	✓	✓	365 days	28 days
Chilblain-like lesions	✓	✓	✓	365 days	28 days
Single organ cutaneous vasculitis	✓	✓	✓	365 days	28 days
Erythema multiforme	✓	✓	✓	365 days	7 days
Anaphylaxis	✓	✓	✓	30 days	2 days
Death (any cause)** (postvaccination control window)	✓	✓	✓	365 days	7 days
Sudden death (by codes)** (postvaccination control window)	✓	✓	✓	365 days	7 days
Meningoencephalitis	✓	✓	✓	365 days	28 days
Acute disseminated encephalomyelitis (ADEM)	✓	✓	✓	365 days	28 days
Narcolepsy	✓		✓	365 days	180 days
Thrombocytopenia	✓	✓	✓	365 days	28 days
Transverse myelitis	✓	✓	✓	365 days	28 days
Bells' palsy		✓	✓	365 days	28 days
Haemophagocytic lymphohistiocytosis[1]		✓	✓	365 days	180 days
Kawasaki's disease		✓	✓	365 days	28 days
Pancreatitis		✓	✓	365 days	28 days
Rhabdomyolysis		✓	✓	365 days	28 days
SCARs		✓	✓	365 days	28 days
Sensorineural hearing loss			✓	365 days	180 days
Thyroiditis			✓	365 days	180 days
Negative control events					
Gout		✓	✓	365 days	28 days
Otitis externa		✓	✓	365 days	28 days
Trigeminal neuralgia		✓	✓	365 days	28 days
Acute kidney injury	✓	✓	✓	365 days	28 days
Anaphylaxis (not drug-induced)	✓	✓	✓	365 days	28 days
C. difficile infection		✓	✓	365 days	28 days
Conjunctivitis		✓	✓	365 days	28 days
COVID-19 unrelated mortality		✓	✓	365 days	28 days
COVID-19 within 12 days after vaccination	✓	✓	✓	365 days	28 days

[1] https://primaryimmune.org/disease/hemophagocytic-lymphohistiocytosis-hlh

Event	ACCESS	SCRI	cohort	Naïve period to estimate new onset	Primary Risk period*
Diverticulitis	✓	✓	✓	365 days	28 days
Fractures		✓	✓	365 days	28 days
Gall stones		✓	✓	365 days	28 days
Influenza		✓		365 days	28 days
Liver cirrhosis		✓	✓	365 days	28 days
Organic (secondary) psychosis		✓	✓	365 days	28 days
Osteoarthritis		✓	✓	365 days	28 days
Osteomyelitis		✓	✓	365 days	28 days
Reactive arthritis		✓	✓	365 days	28 days
Renovascular disease		✓	✓	365 days	28 days
Sjögren's syndrome		✓	✓	365 days	28 days
Urinary tract infections		✓	✓	365 days	28 days
Valvular heart disease (non-congenital, not rheumatic)		✓	✓	365 days	28 days

*For death we may conduct different SCRI analyses

Negative control outcomes must have two important features, which are (a) no association with the exposure of interest and (b) similar sources of bias as the true outcome. This second feature ensures that the negative control outcome tests the same mechanisms of potential confounding that could be present for the true outcome (1). Negative control outcomes that lack feature (b) are of little value in detecting unmeasured confounding, as illustrated by Groenwold et al. Table 2 lists the selected negative control outcomes.

6.3.3 Covariate Definition

Readiness study

In the readiness study covariates (as listed below for the rapid assessment study) are extracted and inspected for algorithms and for methodological analysis.

Rapid hypothesis testing study

Time-varying variables for the SCRI design will be measured at time of occurrence for time-varying factors (e.g. COVID-19). For the cohort design and SCRI, covariate status for stable factors will be measured at time zero. All covariates will be assessed in specific periods, default is during the one-year prior time zero.

Population characteristics are identified based on diagnoses, medicines, laboratory data, survey observation or medical observations, and observation period information.

Demographic characteristics (all measured at time zero)

- Age (0-11 months, 1-<5, 5-11, 12-17, 18-29, 30-59, 60-79, 80+)
- Sex

Pregnancy

- Pregnancy status at time zero (if available), using the pregnancy algorithm developed in the ConcePTION project (see https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies)

Comorbidities with conclusive and higher suggestive evidence for more severe COVID-19 disease[2], all measured at time zero and considered when recorded in year prior to time zero.

- **Cancer** diagnosis or cancer medicines (L01A*, L01B*, L01C*, L01D*, L01X*, L02A*, L02B*, L03*, L04*)
- **Chronic kidney disease** diagnosis (exclusion criterium for assessment for acute kidney injury)
- **Chronic liver disease** diagnosis (cirrhosis, non-alcoholic fatty liver disease, alcoholic liver disease, autoimmune hepatitis)
- **Chronic respiratory disease** diagnosis (chronic obstructive pulmonary disease, bronchiectasis, asthma, interstitial lung disease, cystic fibrosis) or drug proxies (R03*, R07A*)
- **Cardio/Cerebrovascular** disease diagnosis (stroke, transient ischemic attack (TIA), aneurysm, and vascular malformation, coronary artery disease, heart failure or cardiomyopathies) or drug proxies for such disease (C01*, C03*, C07*, C08*, C09* ,B01AC*)
- **Obesity** diagnoses or anti-obesity medicines as proxy (A08AB*, A08AA*)
- **Down** syndrome diagnoses
- **Mental health disease** (depression, dementia, and schizophrenia spectrum disorders) or drug proxies (N05A*, N06A*, N06D*)
- **Sickle cell disease** diagnosis or drug proxies (L01XX05, B06AX01)
- **Diabetes** (type 1 or 2) or diabetes medicines as proxy (A10B*, A10A*)
- **Human immunodeficiency virus** diagnoses or drug proxies (J05AE*, J05AR*, J05AF*, J05AG*)
- **Immunosuppressants**: Use of corticosteroids or other immunosuppressive medications (H02*, L04*)

Covid-19 History

- COVID-19 infection: Covid-19 Dx diagnosis code or positive test further classified by severity:
  - Level 1: any recorded COVID-19 diagnosis, notification to a registry, or positive test
  - Level 2: hospitalization for COVID-19 (COVID-19 diagnosis in primary/secondary discharge diagnosis)
  - Level 3: ICU admission in those with COVID-19 related admission or Acute respiratory distress requiring ventilation during hospitalization for COVID-19
  - Level 4: death during hospitalization for COVID-19 (any cause)

Table 3 Retrieval of Covid-19 PCR/Antigen test

	Medical observations (labs)	**Survey Observations**
Italy, ARS Tuscany		*survey_meaning=`covid_registry´*
Italy, Lazio	NA	
Italy, Pedianet	*"mo_origin = TAMPONI_COVID19 AND mo_source_value = positive"*	
Spain, Valencia VID	*mo_meaning='covid19_pcr_test' AND mo_source_value=`positive´ or (mo_meaning=*'covid19_antigen_test′ *AND mo_source_value=*`positive*´)*	
Spain SIDIAP	mo_meaning='covid19_pcr_test' AND mo_source_value=`positive´	

[2] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html

	OR mo_meaning='covid19_antigen_test' AND mo_source_value=`positive´	
Spain, BIFAP	NA	
Norway, Norwegian Registers	mo_meaning = COVID-19 positive test AND mo_code = 713	
Netherlands, PHARMO	NA	
UK, CPRD	mo_meaning= "covid_lab_test " AND mo_unit= `positive´	

Prior history of events

- prior VTE (deep venous thromboembolism, Pulmonary embolism, splanchnic) or drug proxies (B01AB*)
- History of anaphylaxis diagnosis or use of injectable epinephrine (C01CA24)
- History of allergic reactions

Comedication that may be associated with any of the AESI, **assessed at start of follow-up and at time zero** (prescription/dispensing 90 days prior)

- Antithrombotic agents (B01A*)
- Sex hormones (G03*)
- Antibiotics (J01*)
- Antiviral medications (J05*)
- Lipid lowering drugs (C10*)
- Vaccines (J07 not J07BX03)

The AESI may have different sets of risk factors, and outcome-specific analyses may contain different covariate sets.

For subgroup analyses, we will use the following groups

- immunocompromised persons (yes/no)
- persons with the presence of co-morbidities elevating the risk of serious COVID-19 (yes/no)
- persons with a history of diagnosed COVID-19 disease (yes/no)
- pregnant women at time zero (yes/no)
- age groups
- gender

6.4 Data Sources

The study uses data from secondary electronic health record databases that are population-based. All data sources have the ability to provide data on COVID-19 vaccines, outcomes (diagnoses, procedures, and treatments), and important covariates. It is not currently known the extent to which COVID-19 vaccines, product types, and batch numbers are captured well in the data sources.

6.4.1 PHARMO (NL)

The PHARMO Database Network, which is maintained by the PHARMO Institute for Drug Outcomes Research, is a population-based network of electronic health record databases that combines anonymous data from different primary and secondary health care settings in the Netherlands. These different data banks—including data from general practices, in- and outpatient pharmacies, clinical laboratories, hospitals, the cancer register, the pathology register, and the perinatal register—are linked on a patient level through validated algorithms. To ensure data privacy in the PHARMO Database Network, the collection, processing, linkage, and anonymisation of the data are performed by STIZON, which is an independent, ISO/IEC 27001 certified foundation that acts as a trusted third party between the data sources and the PHARMO Institute. The General Practitioner databank comprises data from electronic patient records registered by GPs. The records include information on diagnoses and symptoms, laboratory test results, referrals to specialists, and health care 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 coding system. 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. General practitioner data cover a catchment area representing 3.2 million residents (~20% of the Dutch population). PHARMO GP databank captures vaccinations supplied by the GP (influenza, zoster, COVID-19).

6.4.2 Clinical Practice Research Datalink and Hospital Episode Statistics (UK)

The CPRD from the UK collates the computerised medical records of GPs in the UK who act as the gatekeepers of health care and maintain patients' life-long electronic health records. Accordingly, GPs are responsible for primary health care and specialist referrals, and they also store information about specialist referrals and hospitalisations. General practitioners act as the first point of contact for any non-emergency health-related issues, which may then be managed within primary care and/or referred to secondary care, as necessary. Secondary care teams also provide information to GPs about their patients, including key diagnoses. The data recorded in the CPRD include demographic information, prescription details, clinical events, preventive care, specialist referrals, hospital admissions, and major outcomes, including death. Most of the data are coded using Read or SNOMED codes. Data validation with original records (specialist letters) is also available. The population in the data bank is generalisable to the UK population based on age, sex, socioeconomic class, and national geographic coverage CPRD Aurum versions is used. There are currently approximately 59 million individuals (acceptable for research purposes) -16 million of whom are active (ie, still alive and registered with the GP practice)- in over 2,000 primary care practices (https://cprd.com/Data). Data include demographics, all GP/health care professional consultations (eg, phone calls, letters, e- mails, in surgery, at home), diagnoses and symptoms, laboratory test results, treatments (including all prescriptions), all data referrals to other care providers, hospital discharge summary (date and Read/SNOMED codes), hospital clinic summary, preventive treatment and immunisations, and death (date and cause). For a proportion of the CPRD panel practices(> 80%), the GPs have agreed to permit the CPRD to link at the patient level to HES data. The CPRD is listed under the ENCePP resources database, and access will be provided by University Utrecht). Other CPRD-linked COVID-19 data sets, which may provide further follow-up information on AESI, include the Public Health England (PHE) Second Generation Surveillance System (SGSS) COVID-19 positive virology test pillar 1 tests, PHE COVID-19 Hospitalisation in England Surveillance System, and the Intensive Care National Audit and Research Centre data on COVID-19 intensive care admissions.

6.4.3 Norwegian Health Registers (NO)

The Norwegian data sources in this project are several national health registers, ie, the Medical Birth Registry of Norway (MBRN), the National Patient Register (NPR), Norway Control and Payment of Health Reimbursement (KUHR), the Norwegian Immunisation Registry (SYSVAK), the National Prescription Registry, and Statistics Norway. The source population will be identified using the Norwegian Institute of Health's (NIPH) copy of the Norwegian population data file from the National Registry. The NPR and KUHR (and the MBRN for the pregnant population) provide data on inpatient and outpatient diagnostic codes. Information on population background data is derived from Statistics Norway (eg, education, occupation status, sex, age). Data on vaccination status are derived from SYSVAK and the Norwegian Prescription Database. The latter register includes data on filled prescriptions for possible co-medications and other prescription drug use.

Norwegian Immunisation Registry

The SYSVAK is the national electronic immunisation register that records an individual's vaccination status and vaccination coverage in Norway. It became nationwide in 1995, and includes information such as personal identity number, the vaccine code, disease vaccinated against, and vaccination date.

The Norwegian Patient Registry

The NPR is an administrative database of records reported by all government-owned hospitals and outpatient clinics and by all private health clinics that receive governmental reimbursement. The NPR contains information on admission to hospitals and specialist health care on an individual level from 2008. The data include date of admission and discharge as well as primary and secondary diagnosis. The NPR has included Norwegian national identification numbers since 2008. Consequently, person-specific data from 2008 onwards are available. Diagnostic codes in the NPR follow ICD-10.

Norway Control and Payment of Health Reimbursement

The KUHR is an administrative database based on electronically submitted reimbursement claims from physicians to the Norwegian Health Economics Administration. It contains information from primary health care, GP, and emergency services on morbidity, utilisation of health care services, and health care use. Person-specific data are available since 2006 . Diagnostic codes in the KUHR follow ICD-10, but the ICPC is more frequently used by GPs.

The Norwegian Prescription Database

Since January 2004, all pharmacies in Norway have been obliged to send data electronically to the Norwegian Institute of Public Health regarding all prescribed drugs (irrespective of reimbursement) dispensed to individuals in ambulatory care. Relevant variables for this project include detailed information on drugs dispensed and date of dispensing.

The Medical Birth Registry of Norway

The MBRN is a population-based register containing information on all births in Norway since 1967 (more than 2.3 million births). The MBRN is based on mandatory notification of all births or late abortions occurring at 12 weeks of gestation onwards. The MBRN includes identification of the mother and father, including national identification numbers, parental demographic information, the mother's health before and during pregnancy, complications during pregnancy and delivery, and length of pregnancy, as well as information on the infant, including congenital malformations and other perinatal outcomes.

Statistics Norway

Statistics Norway provides microdata for research projects and includes information on population characteristics, housing conditions, education, income, and welfare benefits. These data are potential important confounders.

The National Registry

The National Registry (Folkeregisteret) holds information about all inhabitants in Norway. The NIPH holds a copy of the Norwegian population data file from the National Registry that will be used to identify the source population in Norway.

Norwegian Surveillance System for Communicable Diseases

Notification of infectious diseases to the Norwegian Surveillance System for Communicable Diseases (MSIS) is an important part in the surveillance of infectious diseases in Norway. Microbiological laboratories analysing specimens from humans, and all doctors in Norway, are required by law to notify cases of certain diseases (71 in total including SARS-CoV-2) to the MSIS central unit at the Norwegian Institute of Public Health. The following variables are available since 1977: notifiable disease, month and year of diagnosis, age groups, county of residence, and place of infection. Data on positive COVID-19 tests are updated continuously.

6.4.4 SIDIAP (ES)

The Information System for the Improvement of 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 and the IDIAPJGol Institute. It includes information collected since 01 January 2006 during routine visits at 278 primary care centres pertaining to the Catalan Health Institute in Catalonia (North-East Spain) with 3,414 participating GPs. SIDIAP has pseudo-anonymised records for 5.7 million people (80% of the Catalan population) and is highly representative of the Catalan population. The SIDIAP data comprise the clinical and referral events registered by primary care health professionals (eg, 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. The SIDIAP data 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, and blood and urine test results. Regarding vaccinations, SIDIAP includes all routine childhood and adult immunisations, including the antigen and the number of administered doses. Encoding personal and clinic identifiers ensures the confidentiality of the information in the SIDIAP database. Currently, 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.

6.4.5 BIFAP database (ES)

BIFAP (Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria) is a longitudinal population-based database of EMRs from patients attended in primary care facilities of the SNS (Sistema Nacional de Salud), the Spanish National Health System, and located in one of the participating regions throughout Spain. Since 2001, this database has been progressively and increasingly collecting health data, with annual updates, and the current complete version of the database with information until December 2019 includes clinical data of 10.153 Primary Care Practitioners (PCPs) and pediatricians. Nine participant Autonomous Region send their data to BIFAP every year. BIFAP database currently includes anonymized clinical and prescription/dispensing data from more than 13.7 million (9.4 active population) patients representing 85% of all patients of those regions participating in the database, and 29% of the Spanish population. Mean duration of follow-up in the database is 8.7 years. Information collected by PCPs includes administrative data, socio-demographic data, lifestyle, and other general data, clinical diagnosis and health problems, results of diagnostic procedures, interventions, and prescriptions/dispensations. Diagnoses are classified according to the International Classification of Primary Care (ICPC)-2 and ICD-9 code system, and a variable proportion of clinical information is registered in "medical notes" in free text fields in the EMR. Additionally, information on hospital discharge diagnoses coded in ICD-10 terminology is linked to patients included in BIFAP for a subset of periods and regions participating in the database. All information on prescriptions of medicines by the PCP is incorporated and linked by the PCP to a health problem (episode of care), and information on the dispensation of medicines at pharmacies is extracted from the e-prescription system that is widely implemented in Spain.

The BIFAP database was characterized in the ADVANCE project and considered fit for purpose for vaccine coverage, benefits and risk assessment (Sturkenboom et al. 2020). The BIFAP program currently participates in several European projects financed by the EMA, the main objective of some of them is to contribute to the surveillance of vaccine safety against COVID-19: ACCESS ("VACcine Covid-19 tracking readinESS") and "Early-Covid-Vaccine-Monitoring".

6.4.6 FISABIO, VID database (ES)

The VID is a set of population-wide electronic databases covering residents of the Valencia region in Spain, representing approximately 5 million individuals (Garcia-Sempere et al 2020). All the information in the VID databases can be linked at the individual level through a single personal identification. The data sets in the VID are as follows:

The Population Information System (SIP) is a database that provides basic information on health system coverage (eg, dates and causes of Valencia health system entitlement or disentitlement, insurance modality, pharmaceutical copayment status, assigned Healthcare Department) as well as some sociodemographic data (eg, sex, date of birth, nationality, employment status, geographic location). Importantly, the SIP database includes the date of death captured from the Mortality Registry. The SIP database is paramount to the VID, as it is the source of the individual, exclusive, and permanent identifier number associated with each individual (the SIP number), which is then used throughout the rest of the databases, thereby allowing data linkage across the multiple databases in the network.

The Ambulatory Medical Record (ABUCASIS) is the electronic medical record for primary and specialised outpatient activity, with 96% population coverage since 2009. ABUCASIS is integrated by two main modules: the Ambulatory Information System (SIA) and the Pharmaceutical Module (GAIA), including paediatric and adult primary care, mental health care, prenatal care, and specialist outpatient services, as well as providing information about dates, visits, procedures, laboratory test results, diagnoses, and clinical and lifestyle information. It also includes information on several health programmes (eg, healthy children, vaccines, pregnancy, notifiable diseases), the primary care nurse clinical record, and the health-related social assistance record. The SIA module uses the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) for coding diagnoses (and, partially, ICD-10-ES from 2019). The SIA also uses the Clinical Risk Groups system to stratify the morbidity of the entire population.

The GAIA Pharmaceutical module stores data on all outpatient pharmaceutical prescriptions and dispensings, including both primary care and outpatient hospital departments, using the Anatomical Therapeutic Chemical (ATC) classification system and the National Pharmaceutical Catalogue, which allow the identification of the exact content of each dispensing. GAIA does not include in-hospital medication or medication administered in the Accident and Emergency Department (AED). GAIA provides detailed information on prescriptions issued by physicians, such as the duration of treatment and dosage.

The Hospital Medical Record (ORION) provides comprehensive information covering all areas of specialised care, from admission, outpatient consultations, hospitalisation, emergencies, diagnostic services (eg, laboratory tests, imaging, microbiology, pathology), pharmacy, surgical block including day surgery, critical care, prevention and safety, social work, at-home hospitalisation, and day hospitalisation. ORION is currently in the process of being integrated for the whole region, with several databases already fully integrated and available for all hospitals, including the Minimum Basic Data Set at Hospital Discharge (MBDS) and the AED clinical record.

The MBDS is a synopsis of clinical and administrative information on all hospital admissions and major ambulatory surgery in the Valencia health system hospitals, including public-private partnership hospitals (approximately 450,000 admissions per year in the region). The MBDS includes admission and discharge dates, age, sex, geographic area and zone of residence, main diagnosis at discharge, up to 30 secondary diagnoses (comorbidities or complications), clinical procedures performed during the hospital episode, and the diagnosis-related group(s) assigned at discharge. The MBDS used the ICD-9-CM system for coding through December 2015 and ICD-10-ES afterwards. The MBDS was extended in 2015 to include the "present on admission" diagnosis marker and information on tumour morphology. The AED clinical record was launched in 2008 and collects triage data, diagnoses, tests, and procedures performed in public emergency departments. As with the MBDS, the coding system used the ICD-9-CM until December 2015 and the ICD-10-ES thereafter. Diagnosis codification has been increasing from approximately 45% of all emergency department visits between 2008 and 2014 up to approximately 75% in 2017, largely due to the progressive incorporation of hospital coding.

Data on vaccine exposure is obtained from the Vaccine Information System (VIS), which includes information on vaccine type, manufacturer, batch number, number of doses, location and administration date, adverse reactions related to vaccines, and if applicable, risk groups. Information in the VIS is updated daily.
All databases included in the VID are updated frequently (every 1 to 3 months), except the MBDS database, which is updated every 6 months.

6.4.7 ARS Toscana Database (IT)

The Italian National Healthcare System is organised at the regional level: the national government sets standards of assistance and tax-based funding for each region, which regional governments are responsible for providing to all their inhabitants. Tuscany is an Italian region, with approximately 3.6 million inhabitants. The Agenzia Regionale di Sanità della Toscana (ARS Toscana) is a research institute of the Tuscany region. The ARS Toscana database comprises all information collected by the Tuscany region to account for the health care delivered to its inhabitants. Moreover, ARS Toscana collects data from regional initiatives. All data banks in the ARS Toscana data source can be linked at the individual level through a pseudo-anonymous identifier. Two data banks collect dispensings of reimbursed medicines from, respectively, community pharmacies and hospital pharmacies. In the latter data bank, dispensings for outpatient and ambulatory use are complete, and dispensings for inpatient use are partial. Other data banks include hospital discharges, emergency care admissions, records of exemptions from copayment, diagnostic tests and procedures, causes of death, the mental health services register, the birth register, the spontaneous abortion register, and the induced terminations register. A pathology register is available, mostly recorded in free text, but with morphology and topographic SNOMED codes. A COVID-19 registry including all positive cases with clinical follow up is also available. Mother-child linkage is possible through the birth register. Vaccination data are available for children since 2016 and for adults since 2019. All the data banks can be linked at the individual level through a pseudonymous identifier. Data banks are updated approximately every 2 months. Some of them are updated at the date of transmission (eg, vaccines, COVID-19 registry, access to emergency room), others (eg medicines dispensings and hospital discharge records) have a delay of approximately 4 months.

6.4.8 Lazio regional database (IT)

Lazio is an Italian region, with approximately 5.8 million inhabitants. The Department of Epidemiology, ASL Roma1 (DEP Lazio) is a department of the Local Health Authority ASL Roma1, recognised as a regional reference center for epidemiological services and research for27 the Lazio Regional Health Service.
DEP Lazio has access to data collected in the regional administrative healthcare databases referring to mortality, hospital discharge records, emergency room visits, co-payment exemptions, drug claims for outpatients from community and hospital pharmacies, and ambulatory specialist visits. A COVID-19 case registry and COVID-19 vaccine registry are also available.
All data are collected at patient level and can be linked between databases through a pseudo-anonymous identifier.

Data are updated with different lag times, and delays vary between 2 weeks and 6 months.

6.4.9 Caserta LHU database (IT)

The Caserta database is a claims database containing patient-level data from the city of Caserta, in the Campania region. The coverage of this database is very high: from 2005-2020 the catchment area population in Caserta consists of more than 1 million persons (15% of the Campania regional population). The Caserta linkage databases consists of several databases which are linked through a unique patient identifier: a demographic registry, pharmacy claims database with information on concerning all dispensed drugs reimbursed by the Italian NHS, a as well as hospital discharge diagnose databases, emergency department admissions database, claims for diagnostic and laboratory tests ordered, and a registry of patients exempt from reasons for healthcare service co-payment exemptions (e.g. diabetes mellitus, dementia, and other chronic diseases), emergency department visit diagnoses and diagnostic tests. Patient level data from these claims databases, including other drugs reimbursed by the NHS and dispensed by community pharmacies, can be linked together, using a unique patient identifier. The healthcare information in the databases is coded using international coding systems, such as International Classification of Diseases, 9th Edition (ICD 9 CM) for diagnoses and Anatomic Therapeutic and Chemical (ATC) classification for drugs.
A COVID-19 registry including all positive cases with clinical follow up is also available.

6.4.10 PEDIANET (IT)

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).

6.5 Study Size

The estimated size of the source population comprises 45 million individuals.

Cohort

Table 4 shows the statistical power that can be obtained for a range of relative risks and a range of population sizes for a matching ratio index/reference 1:1. For example, 100,000 individuals in the index cohort and 100,000 individuals in the reference cohort allows the detection of a relative risk equal to or greater than 2 with 80% statistical power for diseases with a background incidence rate of ≥100 per 100,000 person-years.

Table 4 Statistical Power for Cohort Design Based on Incidence and Relative Risk.

Number of exposed individuals*	Incidence rate in reference cohort (cases per 100,000 personyears)	Relative risk						
		1.5	**2**	**3**	**4**	**5**	**7**	**10**
50,000	1	3.89	4.88	6.21	7.17	7.99	9.63	12.57
	10	6.17	10.68	21.69	34.84	49.10	75.34	95.86
	50	12.78	30.99	71.98	93.98	99.34	100.00	100.00
	100	20.20	52.57	94.59	99.87	100.00	100.00	100.00
100,000	1	4.26	5.73	8.19	10.42	12.70	17.82	27.59
	10	7.99	16.03	37.24	60.29	79.26	97.06	99.98
	50	20.20	52.55	94.58	99.87	100.00	100.00	100.00
	100	34.22	80.42	99.89	100.00	100.00	100.00	100.00
200,000	1	4.84	7.14	11.74	16.68	22.18	35.07	57.36
	10	11.23	26.14	62.59	88.12	97.73	99.98	100.00
	50	34.20	80.40	99.89	100.00	100.00	100.00	100.00
	100	57.87	97.60	100.00	100.00	100.00	100.00	100.00

Source: Rothman, 2015
*Assuming each individual contributes a 60-day risk window

SCRI

Table 5 shows the statistical power that can be obtained for a range of relative risks and a range of sample sizes. For example, a sample size of 100 cases in the risk or control period will allow the detection of a relative risk equal to or greater than 2 with 93% statistical power. The methods group will assess the impact of the length of the control period on the power.

Table 5 Detectable Relative Risk and Statistical Power for SCRI Design.

Relative Risk	Sample Size*	Power
1.5	20	0.142
2	20	0.320
2.5	20	0.495
3	20	0.638
1.5	50	0.292
2	50	0.667
2.5	50	0.881
3	50	0.963
1.5	100	0.519
2	100	0.926

2.5	100	0.994
3	100	1.000
1.5	150	0.692
2	150	0.987
2.5	150	1.000
3	150	1.000
1.5	200	0.812
2	200	0.998
2.5	200	1.000
3	200	1.000

*sample size = number of events in risk and control period

Table 6 shows the number of vaccinated subjects assuming the same time in control and risk period to obtain an 80% statistical power for a range of relative risks. For example, a sample size of 69 vaccinated individuals with an AE of interest will allow the detection of a relative risk equal to or greater than 2 with 80% statistical power.

Table 6 Detectable Relative Risk and Sample Size for 80% Statistical Power for SCRI Design

Relative Risk	Subjects with AE of Interest	Power
1.5	195	0.802
2	69	0.805
2.5	41	0.809
3	29	0.804

6.6 Data Management

This study is conducted in a distributed manner using a common protocol, the ConcePTION common data model (CDM), and common analytics programs (Figure 2). The data pipeline has been developing from the EU-ADR project and was further improved in the IMI-ConcePTION project (https://www.imiconception.eu/) and used in multiple EMA-tendered and VAC4EU studies. The ConcepPTION CDM description has been described by Thurin *et al, 2022*. Our pipeline 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.

6.6.1 Data Extraction & ETL

Each database access provider (DAP) creates extraction, transform, and load (ETL) specifications using the standard ConcePTION ETL design template (accessible via this link: https://docs.google.com/document/d/1SWi31tnNJL7u5jJLbBHmoZa7AvfcVaqX7jiXgL9uAWg/edit) and upload it to the VAC4EU FAIR Catalogue. The version 2.2 of the ConcePTION CDM is used for this analysis. Following completion of this template and review with study statisticians and principal investigators, each DAP extracts the relevant study data locally using their software (eg, Stata, SAS, R, Oracle). This data is loaded into the CDM structure in csv format. These data remain local.

6.6.2 Data transformation

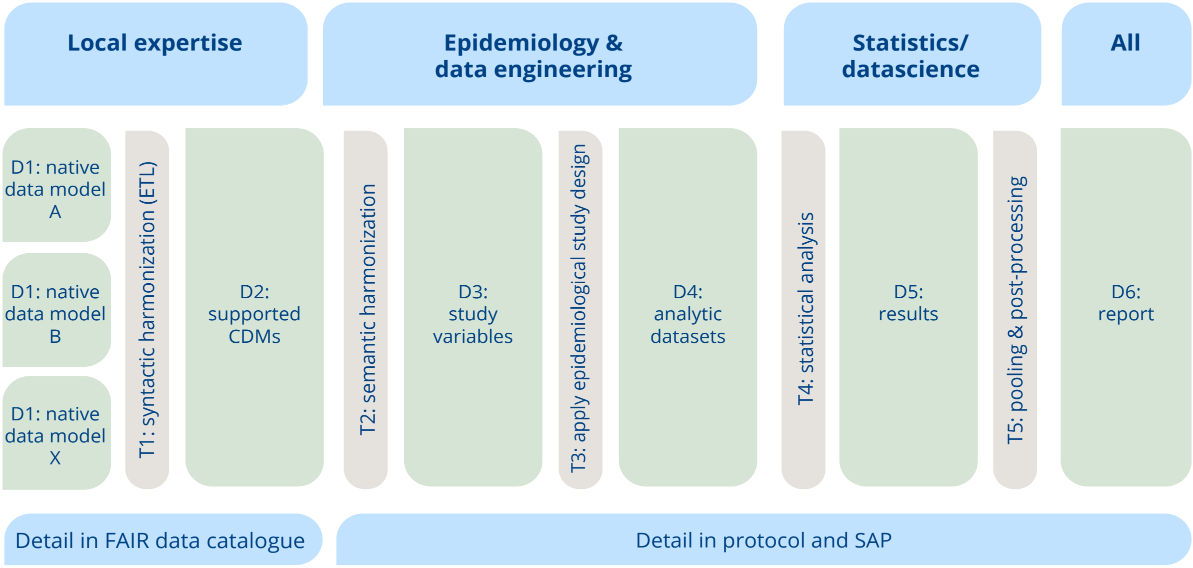


Figure 2 Analytics pipeline. D = Data set(s), T = Data transformation step(s)

6.6.2.1 Generic data analytics pipeline

The data analytics tools comprises a suite of open-source R-based scripts and functions that are hosted on the VAC4EU GitHub and are designed in the sequence provided in Figure 2. Briefly:

T1 = syntactic (structural) transformation of native data into the ConcePTION CDM tables and variables, this is done by the data access providers

The quality of T1 will be verified using level 1 (completeness) and 2 (consistency) data quality checks during onboarding of data partners, and upon every refresh of data (below in readiness phase).

T2 = Transformation of data to study variables for the requested units of analysis by creation of the study population, time anchoring, completion/cleaning missing features in data (e.g., treatment duration, vaccine doses), ordering records in time for one subject, applying algorithms to define events, recoding. The key input for this step is definitions and rules for ‘phenotypes’, algorithms and code sets (e.g., ICD9/10, ICPC, Read, SNOMED) (See Zenodo).

Data quality of study variables (D3) is benchmarked within (temporal trends) and between data-sources using level 3 checks for each study.

T3= Application of the epidemiological study design (cohort, case control, self-controlled), such as sampling from the study population, matching, censoring. We will re-use and tailor existing packages if possible.

T4= Statistical estimations: counting, rates, regression analyses, generalized models etc.

T5= Two-stage pooling of the results and the postprocessing to create overall tables and figures. T5 is conducted on a central environment.

Versioning control will use Git and scripts can be downloaded by the DAPs and run locally. Results (D5) are sent to the digital research environment (DRE, see figure 3) for pooling and post-processing.

6.6.2.2 Ensuring quality of R-scripts

A tight quality control process will be used in development of R-functions and study scripts. This is essential for the robust and transparent transformation of real-world data into evidence. Before launching a new script release the script is tested on a simulated test dataset and a real world dataset. QC is conducted by code review and testing.

6.6.3 Data Access

Within the DRE (see figure 3), each project-specific area consists of a separate secure folder called a "workspace." Each workspace is completely secure, and researchers are in full control of their data. Each workspace has its own list of users, which can be managed by its administrators. The DRE architecture allows researchers to use a solution within the boundaries of data management rules and regulations. Although General Data Protection Regulation and Good (Clinical) Research Practice still apply to researchers, the DRE offers tools to more easily control and monitor which activities take place within projects. All researchers who need access to the DRE are granted access to study specific secure workspaces.

Access to this workspace is only possible with double authentication using an identification code 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. The Download of files is only possible after requesting and receiving permission from a workspace member with an "owner" role.

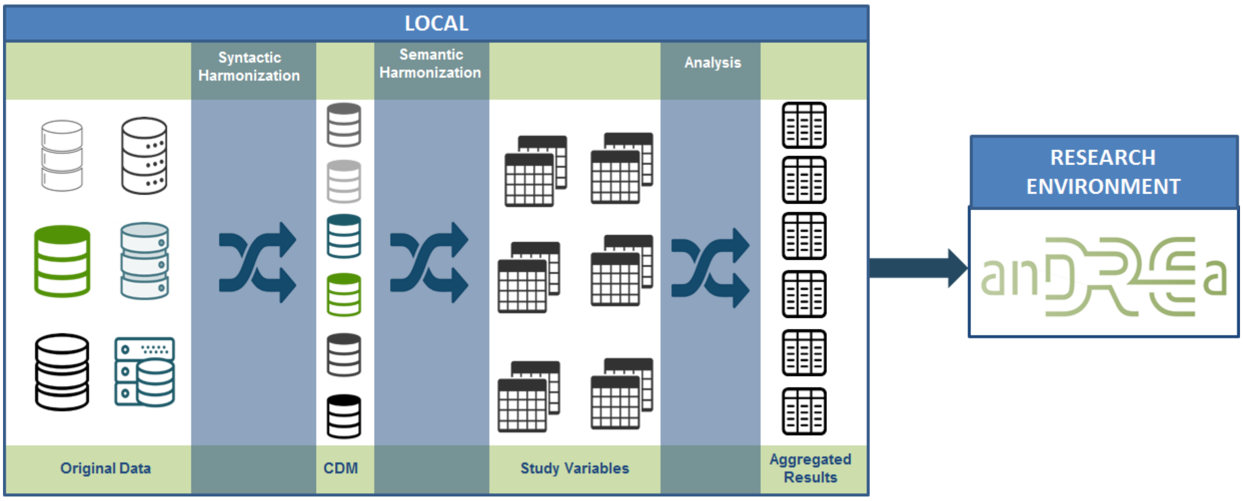


Figure 3 Data transformation and flow

6.6.4 Data Processing

Due to the nature of the study, a repeated data processing procedure is envisioned for readiness and for each novel study request, based on the pipeline described in the previous section. This allows optimising the data processing timelines and archiving procedures. The code for data processing will be documented and edited on the VAC4EU Github and be made publicly available for the CVM study.

For the readiness phase, a baseline data extraction is made by each of the DAPs. This creates a baseline instance of the data source. This is ETL'ed into the ConcePTION CDM and forms the baseline instance of the CDM. The data pipeline will be run for the first time on the baseline instance of the CDM of each DAP, and produce a baseline set of analytic datasets that will be centrally analysed for the baseline assessment.

The output datasets produced by these scripts are then be uploaded to the Digital Research Environment (DRE) for pooled analysis of incidence and visualization. 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/).

All final statistical computations are performed on the DRE using R/SAS or Stata. Data access providers have access to the project workspace for verification of the results.

6.6.5 Record Retention

DAPs are responsible locally to archive each data source instance that is used for the study. The meta-data table in the CDM allows for storing of details on the data source instance. The DAP has the obligation to archive the data source instances, the ETL scripts, the R-scripts that were used, and the results that were uploaded to the DRE, locally.
Aggregated results from DAPs, will be stored in the DRE for inspection by the study sponsor for at least five years.
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 Good Pharmacoepidemiology Practices (GPP) guidelines. 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.
All materials from the DRE will be retained for at least 15 years on a UMCU secure drive. The final study protocol and possible amendments, the final statistical report, statistical programs and output files will be archived on the UMCU secure drive according to Julius Clinical standard operating procedures.

7. Data Analysis

7.1 Readiness phase

During the Readiness phase DAP perform the ETL and the level ConcePTION CDM based 1-3 data quality checks. All checks are verified and assessed by the study team.

7.1.1 Level 1 quality checks (completeness of ETL)

Level 1 data checks review the completeness and content of each variable in each table of the D2 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.). Level 1 checks R-code and instructions are independent of any study and publicly available on the UMCU-RWE). They should be run on each new data instance that is ETL'ed.

Specific objectives of level 1 checks are:

1. To assess the integrity of the Extract-Transform-Load (ETL) process from the original data to the ConcePTION CDM for each Data Access Provider (DAP).
2. To provide feedback on the integrity of the ETL to the DAP iteratively for the refinement of the DAP's ETL procedure.
3. To produce high-level characterization of the data which has been ETL'd to the instance of the CDM in terms of presence/absence of CDM tables and columns, missingness in key variables, frequencies of categorical variables and distribution of dates and continuous variables.

The level 1 checks are divided in 5 major steps:

Step 1: Checking of ConcePTION CDM table formatting

1. Check if all rows of the CDM .csv files in the working directory contain the correct number of variables.
2. Check if all variables in the CDM table are present irrespective of their content.
3. Check if variable names in the csv are written in lowercase.
4. Check for presence of all mandatory variables according to the ConcePTION CDM.

5. Check for presence of non-mandatory variables by comparing between the table of interest and the information recorded in the METADATA table.
6. Check presence of vocabularies for specific variables.
7. Assess formats for all values and compare to a list of acceptable formats which have been filled out in the METADATA table.

Step 2: Missing data analysis

1. Tabulate missingness in all variables, overall and by calendar year (in the tables that contain a date variable).
2. Missing data are stratified by meaning (in the tables that contain a meaning variable).
3. Missing data are displayed using bar charts for each CDM table and reported as counts and percentages.
4. Missing data stratified by meaning or calendar year are displayed using line charts for each CDM table and reported as counts and percentages.
5. Missing data are stratified by meaning and calendar year are displayed using heat maps for each CDM table and reported as counts and percentages.

Step 3: Dates check

1. Check if dates are in the correct format (8 characters).
2. Check if date variables contain allowable values, e.g:
   - Year: 1995-present (exception for dates that represent end of follow up where years in the future will be allowed.)
   - Month: 01-12
   - Day: 01-31

Step 4: Check conventions and construct frequency tables of other and categorical variables.

1. Check if the table of interest contains any duplicate rows.
2. Check that all conventions for the table of interest have been adhered to.
3. Construct frequency tables of categorical variables, overall and by calendar year (when the table of interest contains a date variable).
4. All frequency tables are stratified by meaning when the table of interest contains a meaning variable.
5. Results are reported separately for variables with 2 or more categories.

The results are displayed graphically with bar charts or line charts.

Step 5: Distribution of continuous variables and date variables

1. For continuous variables mean, median, interquartile range, skewness and kurtosis is reported.
2. Distribution of date variables is reported as counts of dates overall and by calendar year.
3. All results are stratified by the meaning variable if the table of interest contains one.

Results are displayed graphically with bar charts or line charts.

Level 1 R-scripts output an R-markdown report that is submitted to the DRE and inspected and assessed by the study team and the DAP.

7.1.2 Level 2 quality checks (internal consistency of data in CDM)

Aims of Level 2 quality checks are to assess internal consistency of the data both within and between tables of the ConcePTION CDM instance for each DAP. Level 2 checks R-code is independent of any study and available on the UMCU-RWE Github

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. Examples of this type of check include: observations occurring before birth date, observations occurring after a recorded death date, parents aged 12 years old or younger etc.

The level 2 checks are divided in 8 major steps:

1. Detect event dates that occur before birth date.
2. Detect event dates after date of death.
3. Detect event dates outside observation periods.
4. Detect subjects included in a CDM table without a corresponding record in the PERSONS table.
5. Detect observations associated with a visit_occurrence_id which occur before the visit_start_date.
6. Detect observations associated with a visit_occurrence_id which occur after the visit_end_date.
7. Detect observations associated with a visit_occurrence_id for which the associated person_id differs from that in the VISIT_OCCURRENCE table.
8. Subjects indicated in PERSON_RELATIONSHIPS as the parent of a child with a birth_date less than 12 years prior to the recorded birth_date of the associated child.

Level 2 check scripts output an R-markdown report that is submitted to the DRE and inspected and assessed by the study team and the DAP.

7.1.3 Level 3 quality checks (study variable check)

Level 3 checks focus on key study variables (population, medications, diagnoses, pregnancy algorithm, medical observations, survey observations and vaccines, life style) based on time anchoring of the population, exclusion criteria and semantic harmonization of outcomes, exposures and covariates, and are divided into different modules which may be included or not depending on the study questions. Level 3 checks allow for benchmarking within a data source over time, between data sources and with external benchmark data. Level 3 checks are in development to optimize detection of deviations. The level 3 script for the ROC20 study is available on the UMCU-RWE Github.

To tailor the level 3 data quality scripts to the study, information is required on the following aspects:

7.1.3.1 Time anchoring of the study population

For this study:

- study period (2018-2022) and run-in time (one year)
- age restrictions (none)

7.1.3.2 Code lists for diagnoses of interest (events tables)

Diagnoses codes lists for events and covariates are created using the VAC4EU process using the open source VAC4EU Codemapper . CodeMapper is open source software and licensed under the Affero

GPL 3. The source code is available on github, and it's concept and function is described by Becker BFH et al., 2017. CodeMapper currently uses the Unified Medical Language System (UMLS) issued by the U.S. National Library of Medicine in version 2020AB.
The Codemapper tool (https://vac4eu.org/codemapper/) is used to create potential diagnosis code lists for all the vocabularies used by the DAPs: ICD9CM, ICD10CM, SNOMEDCT-US, SCTPSA, ICPC, ICPC2.0ENG. MEDCODEID for CPRD is not in the UMLS and mapped using the CPRD browser.

Semantic harmonization of diagnostic codes is a multi-step process because of the different vocabularies that are utilized in Europe

1. Initial diagnosis code lists for each single events or covariates are created using the Codemapper by a clinical epidemiologist. Tagging of concepts is conducted on the Codemapper to 'narrow' or 'possible' to allow for sensitivity analyses with more specific (narrow) or sensitive definitions. Codemapper outputs .xls or csv formats
2. At the VAC4EU Sharepoint we have a phenotype (event/covariate definition toolbox) with a structured naming of phenotypes

Each event definition has a separate folder and be named with

- the system letter (column B),
- event abbreviation (column D),
- type of event (column F) and
- full name of event (column E).

The folder name contains always 3 underscores. All folders that do not contain 3 underscores are automatically excluded by the script. No empty folders are allowed.

e.g. B_ITP_AESI_Immune thrombocytopenia

Inside each event folder there is only one excel file that will contain the codes.

The excel file are named with the variable name which is formed by the system letter, event abbreviation and type of event. Meaning is the same as the folder name without the full name. In this case the name will contain 2 underscores.

e.g. B_ITP_AESI

All code sheets contain the following columns:

- Coding system
- Code
- Code name
- Concept
- Concept name
- Tags
- Comments and edits (this column is used for commenting and review

3. Review of the code lists by the DAPs: they can add codes and comment in comment field
4. Construction of the Extraction code list and the study code list using an automated R-program which ingests the event specific forms and outputs a concatenated code list.
5. Review of code lists is ongoing during study
6. The code list is incorporated in the study specific R-script, which allows then for semantic harmonization from the CDM during the T2 step.
7. Upon finalization an event definition form is created for each event (see Zenodo VAC4EU community: https://www.zenodo.org/communities/vac4eu/?page=1&size=20).

8. Wherever possible the event definition sheet specifies prior validation of algorithms and information for benchmarking.

Records whose code match exactly a code in a codelist are identified as the corresponding AESI.

When converting the data source to the ConcePTION CDM, DAPs label diagnostic codes with characteristics, named *meanings*, considered useful to characterize the context where the codes were recorded (for instance 'primary hospital diagnosis'). Two DAPs requested to remove from the AESIs records bearing some meanings that they considered incicated that it was not an AESI on that date. ARS requested to remove the meanings 'exemption from copayment' and 'Assessment of whether a person qualifies for access to home or residential care'; and BIFAP requested to remove the meaning 'secondary hospital diagnosis'.

7.1.3.3 Code lists for medicines of interest

Code lists for medicines are based on ATC codes, medicines may be used as exposures, covariates, or proxies for events. In the VAC4EU CDM Medicines Sharepoint table we keep track of the Drug proxies (DP), exposures (EXP) and covariates definitions (COV). Edits are tracked and columns for each study are indicated. ATC codes of interest for this study are listed in the section on outcomes and covariates

7.1.3.4 Code list for ConcePTION pregnancy algorithm

To identify pregnancies across databases, we use a pregnancy algorithm developed within the framework of the IMI-ConcePTION project. This algorithm was built on a published algorithm for detecting pregnancies in an electronic health record database by Matcho et al 2018. which used US-based and the CPRD databases. As part of the ConcePTION work diagnosis codes of the Matcho algorithm were mapped to other EU used terminologies (ICD9/10, RCD, SNOMED, ICPC) using the VAC4EU Codemapper, and tagged into categories of live births, stillbirths, abortions, ectopic pregnancies and start of pregnancies to identify beginning/end or outcomes of pregnancy by a medical doctor (CD).

7.1.3.5 Lists for Specific conditions required Survey observations or medical observations of interest

When laboratory data or registers are used for study variables, these may be in local language. In the ConcePTION CDM such information is captured in Medical Observations of Survey observations, values are Dap specific and are provided in dedicated tables. For this study this holds for COVID-19 (see table 3) and for pregnancy 'prompts'.

7.1.3.6 Lists specifying life style conditions of interest

When laboratory data or registers are used for study variables, these may be in local language. In the ConcePTION CDM such information is captured in Medical Observations of Survey observations, values are Dap specific and are provided in dedicated tables. For this study we do not need life style conditions from survey or medical observations.

7.1.3.7 List for vaccines

Vaccines are recorded very heterogeneously as they may not always be prescribed/dispensed as other medicinal products. In the ConcePTION CDM the Vaccine table allows to specify the ATC code of the vaccine, the type of vaccine (voacabulary) and the brand of the vaccine. For the COVID-19 vaccine we use the manufacturer in the brand variable (free text field), the dose and the ATC code (J07BX03)

Level 3 data check scripts output a R-markdown report that is submitted to the DRE and inspected and assessed by the study team and the DAP.

7.1.4 Readiness output based on specific study scripts

Dedicated ROC20 readiness scripts will be generated to produce the tables with the following information (section 10).

Attrition table
This diagram describes the reasons for exclusion and the final study population based on all in-and exclusion criteria for the ROC20 study.

Code counts
This table describes per study variable (AESI and covariates), the count by code/meaning for first AESI occurring during follow-up, and covariates at time zero

Demographic and Baseline Characteristics
The distributions of baseline characteristics at the start of COVID-19 vaccination for each COVID-19 vaccine exposure group at dose 1 (t=0) (all vaccinated) are calculated to describe differences between the groups. For continuous variables, means, standard deviations, medians, and ~~other~~ quartiles are estimated. For categorical variables, counts and proportions plus their confidence intervals are estimated. To describe the relative imbalance of characteristics between different exposed groups, absolute standardized differences are calculated for each baseline characteristic using Pfizer as baseline. Multilevel categorical variables are used to calculate an overall standardized difference across different categorical levels.

Vaccine uptake
For every data source, the number of administered doses per vaccine brand (dose 1, dose 2, booster) by calendar time (in months) over the follow-up period stratified by age groups.

Coverage for first dose of COVID-19 vaccine in the population, calculated as person-time with vaccina as a percentage of the total person-time in every month in the study, is compared with the coverage rates at the ECDC COVID-19 vaccine tracker for the same month, in each data source, per age band (5-11, 12-17, 18-60, 60-79, 80+)

Distance between different doses and brands of covid-19 vaccines are calculated and describe so-called heterologous vaccine schedules whereby patients receive different vaccine types for their first, second or booster dose (Table 5.10-5.10).

Background Incidence of AESI/NCO
Incidence of an AESI /NCO is plotted prior to COVID-19 disease and after COVID-19 disease but prior to COVID-19 vaccine.

7.2 Methods: Evaluation of comparators (4.2)

7.2.1 Choice of comparator (task 4.2)

In this section, we address considerations related to task WP4.2, specifically regarding selecting appropriate comparators for potential use in sensitivity analyses of rapid hypothesis testing (assessment) studies. The primary rapid cycle analyses evaluated AESI following COVID-19 vaccination using a non-causal inference cohort analysis and a SCRI design. As a sensitivity analysis, a causal inference cohort approach will use contemporary comparators (either unvaccinated or vaccinated with another COVID-19 vaccine) with time zero alignment to avoid selection bias and addressing confounding. The negative control outcomes and AESIs of myocarditis and pericarditis will be used as a test case for these sensitivity analyses.

This causal inference cohort approach will better be able to estimate measures of incidence in the population, and the cohort approach does not require assumptions about timing of the risk and control windows relative to vaccination like the SCRI which limits the possibility for adjustment for time-varying confounding. The cohort approach will align the vaccinated group and the comparator group at a time zero (either the date of receipt of Dose 1 of a vaccine, or a matched/selected unvaccinated date) upon which the eligibility criteria will be evaluated, covariates assessed, and follow-up will begin to avoid the introduction of selection bias or immortal person-time bias.

Contemporary or historical recipients of influenza vaccination were recommended early in the pandemic, and in earlier versions of the CVM protocol as a potential comparison group to account for access to healthcare, healthcare seeking behavior, and adherence to recommendations. However, these groups are no longer recommended as a comparator group for the following reasons: 1) indications for influenza vaccination are narrower than those for COVID-19 vaccination, and thus an influenza vaccine recipient group may include only older individuals and those with chronic illnesses, introducing substantial confounding; 2) influenza vaccination is distributed only seasonally, and the time periods of influenza vaccination are much narrower than those for COVID-19 vaccination, potentially subjecting the comparisons to confounding by seasonality. Historic comparators before the COVID-19 pandemic matched on seasonality or calendar month were also proposed initially, however, healthcare utilization and diagnosis patterns have changed during the pandemic because of access issues or changes in health-seeking behaviors resulting in concerns of noncomparability. Thus, we propose focusing on contemporary comparisons of unvaccinated groups or active comparisons with recipients of other COVID-19 vaccines.

7.2.2 Design for Evaluating Potential Comparator Groups (4.2)

A retrospective cohort design will be used to estimate the effect of receiving at least one dose of a COVID-19 vaccine from a specific brand on the incidence rates of myocarditis (and a relevant negative control outcome) compared with either no vaccination or with vaccination with another brand, aligning time zero in both exposure groups.

Brand-specific analyses will each include analyses with both of the following comparison groups:

- Contemporary unvaccinated individuals
    - Contemporary unvaccinated individuals, time zero assigned by individual-level matching
    - Contemporary unvaccinated individuals, time zero sampled at random
- Individuals vaccinated with another COVID-19 vaccine brand (head-to-head, active comparator; time zero is the date of vaccination with the first dose)

In this retrospective cohort design and in both comparator situations, time zero will be defined as the time at which the exposure status is assigned, when inclusion and exclusion criteria are applied and when study outcomes start to be counted. The causal contrast of interest will be the observational analogue of a per-protocol effect, that is, the event rate difference that would be observed if all individuals received at least one dose of the vaccine brand of interest vs. if no individuals received it (for unvaccinated comparators), or if all individuals received at least one dose of the vaccine brand of interest vs. receiving at least one dose of a different vaccine brand (Pfizer to be used as comparator group). Individuals will be classified into exposure groups that are compatible with their vaccination status at time zero.

7.2.2.1 Source Population

The source population will be made up of all individuals registered in each of the participating healthcare data sources, as defined in the primary analysis (Section 6.2.2 Source population). We propose two approaches for the assignment of time zero to the unvaccinated. In the first approach time zero is assigned to the unvaccinated via individual level matching to the date of vaccination, and a second approach where time zero is assigned at random. The first approach guarantees that any time trend in COVID-19-related outcomes is equally distributed in both exposure groups by design. The disadvantage of this approach is that it is computationally intensive because it is the equivalent of creating a series of cohorts, each starting the calendar time when a vaccination occurred. This can be a limiting factor in situations when the sample size is large, as is usually the case in COVID-19 vaccine studies, which can include most of the population in a data source, or when iterative analyses based on resampling are used, as is the case in the estimation of the variance via bootstrapping. The second approach is less computationally intensive as it creates a single cohort. This efficiency is gained at the cost of additional assumptions, which are that we can equal in both exposure groups the distribution of calendar time zero and baseline characteristics via modelling. Given the large sample size of these studies, model misspecification is unlikely, and this assumption can be a weak one.

By comparing both approaches we will be able to inform the design of future studies facing challenges of sample size and resampling.

7.2.2.2 Study Population: Contemporary Unvaccinated Comparator, Unvaccinated Time Zero Assigned With Individual-level Matching

Time zero in the exposed groups (i.e., recipients of the vaccine brand being considered) will be the calendar date on which the first vaccination dose of that brand was received per person. Eligibility criteria will be evaluated on that date.

Time zero for the individuals in the unexposed group will be a matched calendar date on which they did not receive a COVID-19 vaccine dose. The unvaccinated group will not be restricted to "never vaccinated" individuals. This date will be chosen by calendar matching to the time zero of the corresponding exposed group; at each calendar day when an individual is vaccinated, those individuals who were not vaccinated on or before that same calendar day (time zero) will be considered for the unexposed group. On each calendar date, unvaccinated individuals will be matched to the vaccinated individual by important clinical variables (e.g., age, indicated and recommended characteristics to be vaccinated at the time, stratification variables) at time zero. Other studies comparing vaccinated and unvaccinated individuals have demonstrated confounding control after matching on similar demographic and clinical characteristics (*Dagan et al. 2021, Barda et al. 2021.*)

A 1:1 matching approach with replacement using a daily sequential cohort design will be used. Starting on the first day of the study period and chronologically on each date thereafter, we will attempt to 1:1 match with replacement individuals who were vaccinated on that date who meet the eligibility criteria that day to eligible unvaccinated individuals who meet the eligibility criteria that day.

We will explore the impact of exclusion of individuals who had contact with the health care system in the 7 days before time zero (as an indicator of a health event not related to subsequent vaccination that could reduce the probability of receiving the vaccine).

The following matching variables will be used

- Age (year of birth, extending to 2 years)
- Sex (exact matching)
- Prior recorded COVID-19 infection by severity (none, non-hospitalized, hospitalized) (exact matching)
- Geographic area, as available in each data source (exact matching)
- Pregnancy status at time zero (if this can be measured in the data sources)
- Immunocompromising conditions (yes/no exact matching)
- Number of comorbid conditions with evidence of increased COVID-19 severity (0, 1, 2, 3, ≥4) (exact match, see list in covariates)

The set of matching variables and variable levels may be adapted to the availability of variables in each data source. Furthermore, it will depend on computational limitations whether a large number of variables can be included as matching variable. If matching on the a priori defined set of matching variables results on dropping too many individuals because of the lack of matches, alternative sets of matching variables will be explored to maintain an informative study size.

Unvaccinated individuals will be considered eligible for matching in the unvaccinated group on every unique day they are unvaccinated and meet eligibility criteria (Section 6.2.4.3, Section 6.2.5), even if they had previously been included as an unvaccinated control (controls may have been matched previously). Individuals may be included in the unvaccinated group multiple times with different time zeros.

Additionally, an unvaccinated individual may be selected and matched at a particular time zero and then later become vaccinated (i.e., the comparison group will not be restricted to "never vaccinated" individuals). Thus, a single individual may contribute to both exposed and unexposed groups at different time points.

The final analytic cohort will consist of all individuals who successfully matched. Unmatched vaccinated individuals will not be included in the retrospective cohort analysis.

7.2.2.3 Contemporary Unvaccinated Comparator, Unvaccinated Time Zero Sampled At Random

This approach will use an unvaccinated comparator group but does not attempt to match on calendar time or individual-level characteristics. Calendar time and other patient-level characteristics will be adjustedfor analytically. While some outcomes may have seasonality requiring accounting for calendar time analytically, this approach does not require matching algorithms and may be computational less burdensome.

Time zero in the exposed groups (i.e., recipients of the vaccine brand being considered) will be the **calendar date on which the first vaccination** dose of that brand was received per person. Eligibility criteria will be evaluated on that day. All eligible individuals in the data source who received a COVID-19 vaccine during the study period will be included in the exposure group.

Time zero for the eligible individuals in the unexposed group will be a day when they did not receive a vaccine. All individuals may be considered for the unvaccinated group if they have ≥ 1 day of enrolment in the data source after December 1, 2020 meeting eligibility criteria when an individual has not previously received the COVID-19 vaccine. For individuals who are never vaccinated, all their eligible calendar days during the period will be considered as unvaccinated time and potential unvaccinated time zeros. However, this comparison group will not be restricted to “never vaccinated” individuals, and for those who are vaccinated, all eligible calendar days during the study period before their first dose will be considered as potential unvaccinated calendar days.

Because including all the candidate time zeros can be computationally challenging, we will select a 20% random sample of the candidate time zeros. Such random selection will be done with replacement. Therefore, a single individual will be able to contribute more than once to the unexposed group (with the same or with different times zero), and both to the unexposed and exposed groups. If feasible, efficiency changes derived from varying the size of the random sample of times zero (e.g., 20%, 40%, 60%) will be explored, as the size of the sample can have substantial impact on the evaluation of very rare events.

The final analytic sample for this comparison consists of all vaccinated individuals and selected unvaccinated individuals with their respective time zeros.

In addition to the overall cohort inclusion/exclusion criteria, individuals will be excluded if they have contact with the health care system in the 7 days before time zero (as an indicator of a health event not related to subsequent vaccination that could reduce the probability of receiving the vaccine).

7.2.2.4 Head-to-Head Active Comparator

Time zero in the exposed group (i.e., recipients of the vaccine brand of interest) will be the day the first vaccination dose was received. time zero in the comparator group with be a day when the first dose of the comparator vaccine brand was received.

An individual may receive doses from multiple vaccine brands, but only the first COVID-19 vaccine observed per individual will be eligible for inclusion in a brand-specific cohort. For each brand-specific cohort, all individuals meeting eligibility criteria will be included in the overall cohort.

Each vaccine brand will be compared with Pfizer, as Pfizer is the most widely used vaccine in Europe and was among the first COVID-19 vaccines authorized in Europe. The following brand-specific comparisons will be made:

- ≥1 dose of AstraZeneca vaccine vs. ≥1 dose of Pfizer
- ≥1 dose of Janssen vaccine vs. ≥1 dose of Pfizer
- ≥1 dose of Moderna vs. ≥1 dose of Pfizer

For each brand-specific comparison, cohorts will be restricted to time periods where both vaccine brands were authorized and used in the respective countries (i.e., the brand-specific cohorts used in the comparison will be restricted to those with time zero dates occurring after the later authorization date of either vaccine).

7.2.2.5 Follow-up

Individuals will be followed from the date of time zero (inclusive) until the occurrence of the study outcome (myocarditis, pericarditis, COVID-19, MIS or negative control outcome, or other outcomes requested by EMA as part of ROC20) or censoring at the first occurrence of one of the following events:

- Death
- Censoring at one of the following:
    - Administrative end of follow-up (end of study period)
    - Individual exits the database
    - Receipt of COVID-19 vaccine dose of a different brand than that received on time zero

7.2.3 Statistical Analysis of Cohorts (4.2)

The analyses will be performed separately for each vaccine brand and for each comparison, as the vaccinated individuals included in each comparison may differ due to matching or time restrictions.

7.2.3.1 Descriptive Characteristics: Contemporary Unvaccinated Comparisons

The attrition of the study population will be described by reporting the counts of individuals excluded from the study cohorts by application of exclusion criteria. As unvaccinated individuals (in both the matched and randomly selected comparison groups) will be considered eligible to match on every day they meet eligibility criteria and may be excluded for multiple reasons on different days, the attrition of the unvaccinated with describe the attrition of opportunities to match rather than unique individuals; an individual will contribute an entry on the attrition chart on every day she/he is considered.

The distributions of baseline characteristics of the matched cohort at time zero by exposure group will be calculated to describe the study cohort and illustrate differences between the exposure groups.

- For continuous variables, means, standard deviations, medians, and first and third quartiles be estimated.
- For categorical variables, counts and proportions will be estimated.
- To describe the relative imbalance of characteristics between exposed groups, absolute standardised differences (ASD) will be calculated for each baseline characteristic. The larger the absolute standardised difference values, the greater the imbalance between baseline characteristics.
- Multilevel categorical variables will calculate an overall standardised difference across all levels.

For both of the contemporary unvaccinated comparator analyses, the distribution of characteristics between the matched vaccinated and unvaccinated groups will be displayed). If residual imbalances are noted and inverse probability of treatment (IPT) weighting is implemented the covariate distributions after IPT weighting will be repeated to observe if covariate balance has been improved.

7.2.3.2 Descriptive Characteristics: Head-to-Head Active Comparators

For the head-to-head active comparator analysis, the distributions of characteristics of both the vaccinated exposure groups and the ASD will be estimated both in the cohort before IPT weighting (crude) and after IPT weighting to evaluate the improvement in covariate balance by IPT weighting.

7.2.3.3 Measures of Occurrence and Association

Negative control outcomes will be evaluated first in each cohort comparison followed by myocarditis and pericarditis.

Outcome analyses will be performed separately for both contemporary unvaccinated comparison and the head-to-head active comparison in both unweighted (crude) and IPT weighted analyses. If the IPT-weighted analysis is not performed in the matched contemporary unvaccinated analysis, only the matched, unweighted results will be performed.

The cumulative incidence of the outcomes in each treatment group will be estimated as 1 minus the Kaplan-Meier survival curve, and the 95% confidence interval will be estimated as 1 minus the 95%

confidence interval limits of the Kaplan-Meier estimators, estimated with a robust variance estimator. The cumulative incidence curves and 95% confidence intervals will be plotted. Individuals without an event will be censored at one of the events described in Section 7.2.2.5 Follow-up.

To describe the rates of the outcomes, incidence rates for the outcomes will be calculated by dividing the number of cases by the follow-up person-time. 95% CIs will be estimated using the exact method. IRs will be expressed as events per person-years with appropriate scaling of the rate (e.g., per 100 person-years).

Cox proportional hazards models will estimate hazard ratios (HR) and 95% CIs for each outcome. CIs will be estimated with robust variance estimator to account for an individual's possibility to be included in the cohort multiple times.

The Cox proportional hazards model for each AESI will include the binary indicator for the occurrence of the event as the dependent variable, the exposure status as the binary independent variable, and the time since time zero in days as the time scale.

The cumulative incidence plots and HRs for AESI will be evaluated overall for all available follow-up time. If shorter follow-up times are of interest for specific outcomes (i.e., length of risk periods shown in Table 2, or extended risk windows, if specified), HR's can be estimated for the time period of interest.

7.2.3.4 Adjustment for Baseline Imbalances

There may be differences in characteristics between exposure groups that may determine their risk of outcomes. The analytic techniques employed to control for confounding will differ based on the comparison.

The negative control outcomes will be used to evaluate the extent of remaining confounding after adjustment for baseline imbalances. The observed HR estimate and 95% CI for the negative control outcomes should be consistent will a null effect to provide assurance of confounding control. Comparisons which demonstrate effective control of confounding will be used for evaluation of myocarditis/pericarditis.

Contemporary Unvaccinated Comparator Analysis, Matched

The vaccinated and unvaccinated exposure groups will already be matched on several key factors such as calendar time, demographics, and clinical factors (Section 7.2.2.2) which may address most of the major confounding. If there are not imbalances in the remaining measured covariates (Section 7.2.3.1), the analysis of the outcomes will proceed in the matched exposure groups without additional adjustment. A baseline variable will be considered imbalanced if the standardized mean difference between groups is >0.1

If imbalances in other measured covariates remain after matching, stabilized inverse probability of treatment weights (sIPTW)will be estimated. To build the weights, we will first compute a propensity score (PS), defined as the probability of vaccination conditional on the matching variables and on those variables found to be disbalanced after matching. The PS will be computed using a logistic regression. The PS will be used to estimate sIPTW as the following, p_e is the marginal probability of vaccination:

For vaccinated individuals: $\text{sIPTW} = p_e / \text{PS}$

For unvaccinated individuals: $\text{sIPTW} = (1 - p_e) / (1 - \text{PS})$

The stabilized weights will be applied to the exposure groups, and the covariate balance in the weighted groups will be evaluated. If balance has been achieved, the negative control outcomes will be evaluated; the analysis of myocarditis and pericarditis will proceed if the negative control outcomes provide sufficient evidence of confounding control.

Contemporary Unvaccinated Comparator Analysis, Random Time Zero

In this analysis sIPTW will be used to balance covariates between exposure groups, as defined above. The variables used to compute the PS will be the baseline calendar time, and those showing disbalance between exposure groups and/or those considered potential confounders.
As above, the stabilized weights will be applied to the exposure groups, and the covariate balance in the weighted groups will be evaluated. If balance has been achieved, the negative control outcomes will be evaluated; the analysis of myocarditis and pericarditis will proceed if the negative control outcomes provide sufficient evidence of confounding control.

Head-to-Head Active Comparator

In the head-to-head comparison of different vaccine brands, no baseline matching will be employed. Comparison-specific PS models will be constructed with the vaccine brand of interest (versus the comparator vaccine brand) as the dependent variable, and the comparison-specific sIPTWs will be estimated following all the same processes described in Section 7.2.3.4 with the exposure vaccinate of interest versus the Pfizer comparison group.

The comparison-specific sIPTWs will be applied to the exposure groups, and the covariate balance will be evaluated in the weighted groups. If balance has been achieved, the negative control outcomes will be evaluated; the analysis of myocarditis and pericarditis if the negative control outcomes provide sufficient evidence of confounding control.

7.3 Methods: Sensitivity analysis for Self-controlled Risk Interval Design (task 4.3)

The aim of the study related to work in task 4.3 will be to investigate a number of assumptions of the SCRI to evaluate the robustness of this design for the purposes of evaluating COVID-19 vaccine safety. Specifically, the following work will be undertaken:

- Sensitivity analyses (note: incorporated as part of any empirical work and not considered a stand-alone output)
- Simulation study of core assumptions of the SCRI
- Simulation and empirical evaluation of the control period definition in the SCRI

On the first aspect of the task 4.3 work, several additional sensitivity analyses are recommended as part of any self-controlled study, and these will be implemented in all core WP3 analyses:

- Vary the length of the pre-exposure window
- Create exposure centred interval plots
- Tabulate deaths after the outcome
- Plot time between event and end of observation, by censoring status
- Vary the definition of the risk windows

In addition, there may be sensitivity analyses that are appropriate depending on the characteristics of a specific safety concern. For example, if there has been significant media coverage following the first report of a safety concern, this can make clinicians less likely to administer the vaccine to those they perceive to be at risk of such an event. Therefore, it may be appropriate to censor analyses at the first report of a safety concern. Such sensitivity analyses will be implemented as appropriate. Because these sensitivity analyses are primarily of interest as they apply to a specific clinical question, they will be added to any WP3 report rather than produced as a separate output.

The methodology and initial results from the second task, the simulation study of the core assumptions of an SCRI, have been described in a separate protocol (protocol of WP4, Evaluation of assumptions of SCRI in simulation studies) and again in the D.4.1_CTM_Interim_Report. This will therefore not be described further in this SAP. To address the third task, evaluating the performance of different specifications of control periods in self-controlled case series, we will adapt this simulation framework, as well as conduct an empirical investigation using a case study of COVID-19 vaccination and myocarditis. The rationale, objectives, and methods for this are presented below.

7.3.1. Rationale

A range of self-controlled designs have been employed in vaccine epidemiology, including self-controlled case series (SCCS) and self-controlled risk interval (SCRI) designs. An SCRI is a special case of the SCCS, in which the baseline time is fixed for some short period of time in relation to the administration of the vaccine, for example, the 60 days preceding the vaccination, whereas in an SCCS all time which is not designated a risk window is used as control time. The definition of the control window is an important part of the design of self-controlled studies, and can impact findings for several reasons. A specific concern in the study of COVID-19 vaccines is that the use of a pre-vaccination control window may introduce bias, as many safety events significantly reduce the probability of a person being vaccinated. Such "event-dependency of the exposures" can lead to overestimation of any safety signal. Although short pre-exposure windows may be used to account for this when bias is not severe, this will not address the issue if the probability of receiving the vaccine is permanently altered after the outcome. In those scenarios, an alternative option is to use only post-vaccination time as the control time, though the tradeoff here is that rapid studies are less feasible as they require longer accrued follow up in each database. In addition to potentially impacting the extent of bias in analyses, the choice of control window might also influence the efficiency of the method. This is both because a longer control window may allow for the inclusion of a greater number of cases, and because the power of a self-controlled case series depends in part on the ratio of risk to observation time. The aim of the current study will be to investigate the performance of different specifications of control periods in self-controlled case series, and to evaluate the impact of these design choices in a case study of COVID-19 vaccination and myocarditis.

7.3.2. Objectives

1. To evaluate the performance of an SCCS using a range of control period specifications in a simulation study whilst varying the extent of event-dependency of the exposures. This will be integrated the WP4 task of simulation study.
2. To evaluate the impact of choice of control period in a case study using an SCCS to investigate the association between COVID-19 vaccines and myocarditis.

7.3.3. Methods

The simulation study of the core assumptions will use the methods outlined in D.4.1_CTM_Interim_Report. For this investigation, we will adapt this framework to investigate the performance of different control periods. Briefly, over 1,000 replications, we will evaluate three different study designs using different control periods (Table 7) under a number of data generating mechanisms, varying in turn the strength of the association between the exposure and outcome (RR 1, 1.5, 2 and 5) and the strength of the event-dependency of the exposures. The latter will be introduced as a "delay" of the exposure following occurrence of the outcome, and we will consider the presence of no, moderate and strong bias (defined as no delay, a mean delay of 45 days and a mean delay of 180 days). These delays were chosen to enable an investigation of the impact of event-dependent exposures not handled by the 30-day pre-exposure window. We will look at multiple performance measures considering both bias and efficiency.

Table 7 Overview of study designs to be evaluated in WP4 Task 4.3

Study Type	**Control Period Definition**
SCCS	Start of Observation Period (1/9/2020) - End of Observation Period min(end of data availability, death) minus the 28-day risk period after each vaccine dose, and a 30 day pre-exposure window
SCRI	(-90, -30] relative to the first vaccine dose (**primary analysis)**
	(+29 after last vaccine dose, +90]

To complement the theoretical simulation results, we will also evaluate the three different designs described in table 7 in the data sources described above, as a sensitivity analysis to the core SCRI analyses of myocarditis. Our base case will use the same analysis strategy as in the core SCRI, as described in section 6.1.2 in this document. Briefly, the SCRI with a 90-day pre-vaccination window will correspond to our primary analysis. We will then implement the two additional design strategies outlined in table 7 as sensitivity analyses. When constructing the post-vaccination period, this will be considered post dose 2 vaccination time for those who received dose 2, and post dose 1 vaccination time for those without a second dose. As the biasing processes may vary depending on country-specific characteristics, results will be presented per country and not meta-analysed as in the primary analyses. The only aspect which will be varied in these analyses is the control window: all other analytical aspects will be the same as in the core SCRI – that is, we will use 28 day risk windows, allow second dose risk windows to take precedence over first dose risk windows, and we will adjusted for calendar time in 30 day increments.

7.4 Methods: Unmeasured confounding (task 4.4)

Observational (pharmaco-)epidemiologic studies aiming to distinguish causal effects from simple association must deal with confounding, among other factors (1). Measured confounders can be adjusted for in the analysis, but poorly measured or unmeasured confounders are more difficult to deal with. Several methods to detect or control for unmeasured confounding in both the design and the analysis phase exist and have been described in detail (2). This task will explore and compare several methods to detect and quantify unmeasured confounding in the analysis phase. These include negative control outcomes, quantitative bias analyses, and instrumental variable (IV) analyses.

7.4.1 Negative control outcomes

This task aims to provide a negative control outcome analysis framework for all adverse events of special interest (AESIs) included in the original Covid Vaccine Monitoring protocol. This comprises both the selection of a negative control outcome and outlining how the negative control outcome can be incorporated within the existing statistical analysis plan.

7.4.1.1 Identifying negative control outcomes for each AESI

Negative control outcomes were defined based on a combination of the *U-comparability* approach of *Lipsitch et al.* and the approach used by *Ryan et al* to create a reference set of negative control exposures for methodological research in drug safety.

The AESIs were grouped into categories based on the organ system they affected and divided based on whether the main origin of the AESI is infectious or non-infectious in nature (excluding Covid-19 as origin, which is a risk factor for all AESIs). This was done because AESIs within the same organ system can be expected to share risk factors and therefore may be assigned the same negative control outcome, whereas infectious and non-infectious diseases cannot be expected to share many risk factors and thus may not be assigned the same negative control outcome. The risk factor profile for each AESI group was based on a wide internet search including mainly webpages from medical institutions (Mayo Clinic, John Hopkins). The risk profile forms the basis of matching AESIs to potential negative control outcomes. These were identified based on expert medical knowledge within the consortium, a wide internet search based on the risk factor profiles, and by creating a list of potential outcomes from organ systems not affected by COVID-19 disease according to available published literature. We tried to identify at least one negative control outcome per AESI group.

Information on risk factors was compared between all sources and combined to create a complete risk factor profile. Risk factor profiles are presented per AESI group, specifying which risk factors apply to all AESIs in that group and which ones are AESI-specific. Based on this risk factor profile, we identified at least one candidate negative control outcome that is as *U-comparable* to the AESI as possible.

Subsequently, we checked published literature for any evidence that links the candidate negative control outcome(s) to Covid-19 vaccines. We performed a wide internet search for each potential negative control outcome combined with mention of COVID-19 vaccination. All hits returned by this search that suggest such a link exist, were traced back to the manuscript published in literature to determine whether the link can be considered likely or whether it may be unfounded. Following *Ryan et al*. we excluded candidate negative control outcomes if we find at least one published randomised trial or population-based observational study that reports a positive association (point estimate > 1) between this candidate outcome and Covid-19 vaccination. We also excluded candidate negative control outcomes if we find case reports that strongly suggest a causal association between Covid-19 vaccination and the candidate outcome to avoid potential misclassification as suggested by *Hauben et al.* In this process, we also considered the evidence relative to vaccine brand. For example, if we found published evidence for vaccine brands that were not distributed in the countries included in the consortium, we still consider the negative control outcome appropriate. Or if the evidence only concerns a subtype of vaccines (mRNA or vector based, or even only one brand), we consider the negative control appropriate for other vaccine types.

We also included some candidate negative control outcomes specifically for self-controlled designs. These adjust for time-fixed confounders by design but are sensitivity to time-varying confounders such as seasonality. A *U-comparable* negative control outcome therefore only has to be associated with seasonality and not necessarily with time-fixed confounders that are part of the true outcome's risk profile.

7.4.1.2 Use of negative controls to estimate confounding

Negative control outcomes can be implemented in both the cohort studies and the SCRI studies as a sensitivity analysis for real and readiness studies.

7.4.1.3 Quantitative bias analyses

Quantitative bias analyses provide information on how strong the relationship between an unmeasured confounder and the exposure and outcome of interest should be to explain away the observed association.

One of the most accessible approaches to quantitative bias analysis is the *E-value,* which was introduced by VanderWeele and Ding (Vanderweele, 2017). The E-value is one value (on the risk ratio scale) that represents how strongly an unmeasured confounder should be associated with both the vaccine exposure and the AE of interest for the observed vaccine-AE association not to be causal, conditional on measured confounders (Vanderweele, 2017). Its strengths lies in the ease of computation and ability to be compared between studies. It can therefore be calculated for any vaccine-AE association of interest as a first step towards quantitative bias analyses. If the E-value is considered so large that unmeasured confounding is unlikely (e.g., there is unlikely to exist an unmeasured confounder this strongly related to both vaccine and AE), further bias analyse may be considered unnecessary.

However, the E-value is mainly useful in situations with one known unmeasured potential confounder, as it can be misleading in other more complex situations (Sjölander, 2022). As part of hypothesis testing studies, we will also perform more elaborate quantitative bias analyses such as the array approach, as we will often expect more than one unmeasured confounder. In the array approach, we will calculate the adjusted risk estimate of interest across a range of values for the association between the confounder and the exposure, the association between the confounder and the outcome, and the prevalence of the confounder (Schneeweiss, 2006). Based on the resulting array of risk estimates, we can consider whether proposed potential unmeasured confounders would be strong enough to fully explain away the observed vaccine-AE association (Schneeweiss, 2006). This gives a better impression of the characteristics potential confounders should have, and allows researchers to compare these to data or literature on the associations between potential confounders and the outcome and exposure of interest. Although more laborious than calculating an E-value, array approaches are still considered computationally light as long as they model only one bias (Lash, 2014). To make quantitative bias sensitivity analyses feasible for any vaccine-AE association of interest, we will restrict ourselves to simple (one fixed value per bias parameter) or multidimensional (range of values per parameter) quantitative bias analyses as described in Lash (Lash, 2014). More elaborate analyses, such as those that not only address unmeasured confounding but also misclassification, can be performed in collaboration with the other WP4 sub-packages (see 7.5). in cases where multiple biases are expected to strongly affect the observed estimate and performing a computationally intensive sensitivity analysis if feasible.

7.4.1.4 Instrumental variable analyses

An instrumental variable (IV) is described as a variable that mimics the treatment assignment process in a randomised study, which (almost) perfectly coincides with the treatment received and only affects the outcome through the received treatment (Uddin, 2015). Formally, this translates to three assumptions a variable should meet to be a valid IV (Brookhart, 2010; Hernán, 2006; Labrecque, 2018):

1. The variable is either causally related to the exposure or shares a common cause (*relevance*)
2. The variable affects the outcome only through the exposure (*exclusion restriction*)
3. The variable does not share common causes with the outcome (*independence or exchangeability assumption*)

There are two additional assumptions for valid IV analysis, which are (a) *monotonicity* and (b) *homogeneity.* This entails that there (a) are no so-called *defiers* in the study which would deliberately choose the opposite treatment of what is suggested by their IV (in a randomised trial, these would be patients that choose whichever option they were not randomised to), and (b) that there is no effect modification by the IV on an additive scale (Hernán, 2006; Brookhart, 2010; Labrecque, 2018). Although these assumptions, with the exception of *relevance,* cannot be verified, several falsification strategies to assess them have been proposed (Labrecque, 2018). We will apply these strategies in the process of identifying and validating potential IVs.

As a valid IV mimics randomised treatment assignment, it can be a powerful tool to calculate effect estimates unaffected by potential confounders, similar to an intention-to-treat analysis (Hernán, 2006). However, an IV that is only weakly associated with the exposure (*weak instrument bias*) can exaggerate both any biases present in the association between the IV and the outcome and small sample bias in the association between the IV and the exposure, which might lead to an IV estimate that is more biased than the unadjusted estimate of the main analysis (Hernán, 2006). In addition, IVs are poorly equipped to address time-varying exposures (Hernán, 2006) and are often underpowered for drug safety studies of very rare outcomes (Brookhart, 2010). Therefore, we will consider whether IV analysis is feasible and of additional value on a case-to-case basis. Earlier work on identifying IVs for influenza vaccines has suggested it is difficult to find proper IVs in a vaccination setting (Groenwold, 2010). We will therefore first test whether potential IVs can be considered valid using the falsification strategies mentioned before. If these suggest none of the potential IVs can be considered a valid IV, we will not pursue IV analyses. We propose country and vaccine provider (who administered the vaccine) as potential IVs and will use the myocarditis case to test whether these IVs are valid. We will test whether the IVs are associated with the exposure of interest (*relevance* assumption) using simple regression. The *exclusion restriction* assumption will be tested using the instrumental inequalities as described in Labrecque and Wang (Labrecque, 2018; Wang, 2017). We will perform a covariate balance check to assess whether the *exchangeability* assumption holds, for example by creating a baseline table of all confounders across the levels of the IV.

When a good IV can be identified, we will next consider which analysis method to use. The best method to calculate the IV estimator depends on the number of IVs, measured confounders that need to be adjusted for, the nature of the association between the exposure and the outcome (linear, nonlinear), and the format of the exposure, outcome, and IV (continuous, binary) (Uddin, 2015). We will choose the best method based on these considerations, using the overview provided by Uddin and colleagues as a guide (Uddin, 2015).

7.5 Misclassification of outcomes (task 4.5)

This task aims to quantify the effect that misclassifications in adverse events have on estimates of relative risk, risk differences, and log-odds ratios. Where such effects can be quantified, we aim to mitigate them using measurement error models. Where they cannot be quantified precisely, we aim to indicate the sensitivity of estimates of relative risk, risk differences, and log-odds ratios to misclassification. As a convenient method to incorporate the additional uncertainty, Bayesian models will be used.

The task will proceed in the following steps:

1. Conduct a literature review of misclassification in adverse events of special interest (AESIs);
2. Conduct a simulation study quantifying the effect likely misclassification will have on estimates of relative risk, risk difference, and log-odds ratios in AESIs;
3. Develop a statistical measurement error model allowing for correction of relative risk, risk difference, and log-odds ratios in AESIs for known misclassification rates, and evaluate its performance in the simulation study;
4. For the case in which misclassification rates are not known precisely, extend the measurement error model using Bayesian priors.

All code developed will be made available under a suitable open-source license for further use.

7.5.1. Literature overview

To get a sense of the plausible degree of misclassification for different AESI, we will conduct a literature review of studies that quantify AESI misclassification indices (e.g., PPV, sensitivity, specificity). Note that an exhaustive systematic review is outside of the scope of this task; instead, the goal will be to provide an overview of plausible values of misclassification, based on earlier studies, inso far as these exist.

As a search strategy, we will use the following search term: validity index combined with the name of adverse event of interest and the codelist to obtain it from ACCESS (see table 8 as an example). The term validity index could be replaced by PPV, NPV, false positive, false negative, or misclassification. For specific adverse event where validation studies have not been conducted, we will use a proxy to obtain the estimate the misclassification probabilities. For example, in the myocarditis case, we may use another cardiac disease. Note that we do not expect to obtain an exhaustive overview; for example, sensitivity is rarely quantified in the literature.

Table 8 The search term used for literature study and its synonyms. Here, myocarditis is used as an example

AESI	Codelist to obtain the AESI (from ACCESS)	Validity index
Myocarditis acute myocarditis myocardial inflammation cardiomyopathy	ICD9 CM 422 ICD10 CM 140	false positive false negative sensitivity specificity PPV NPV misclassification

7.5.2 Simulation study

To quantify the effects that likely misclassification has on estimates of interest, we will conduct a simulation study. In this study, we will examine a setup with a binary true vaccination status and a binary AESI outcome,$y \in \{0,1\}$.

The following factors will be varied:

- Sample size n , ranging from $n = 50$ to $n = 15 \times 10^6$. This range is chosen to account for the existence of small subgroups and a large combined database;
- The proportion of vaccinated people,$p \in \{0.2, 0.5,0.7,0.8\}$;
- The true relative risk, parameterized in the simulation study as a log-odds ratio $\beta \in \{0, 0.05, 0.2, 0.5, 1.0, 2.0\}$;
- The base adverse event rate, parameterized as an log-odds $\alpha \in \{-9, -7, -4.5, -2\}$;
- The sensitivity (recall) of the event registration $p_{sens} = \{0.99, 0; .9, 0.8, 0.7, 0.6\}$,
- The specificity of the event registration, $p_{spec} = \{0.99, 0.9, 0.8, 0.7, 0.6\}$.
- The uniform differential misclassification coefficient $\delta \in \pm\{0, 0.1, 0.2, 0.5, 1\}$
- The non-uniform differential misclassification coefficient $\gamma \in \pm\{0, 0.1, 0.2, 0.5, 1\}$

Denote the observed (registered) AESI outcome as $y^* \in \{0,1\}$. Then we consider this observation to be the realization of a Bernoulli random variable whose success is determined by the true AESI status, and possibly (in the case of differential misclassification) the vaccination status,

$$y^* \sim \text{Bernoulli}\big(\pi(Y * = 1|\, Y = y, Z = z|)\big)$$

Where Y^*, Y and Z are the random variables, and we will use the convention that $\pi(Y^* = 1|Y = y,\, Z = z|)$ is a probability parameter for the conditional probability indicated in the superscript, with the corresponding values given in the subscript. This parameter is defined using the logistic function,

$$\pi(Y^* = 1|Y = y,\, Z = z|) = \big(1 + \exp\big(-(\tau + \lambda y + \delta z + \gamma yz)\big)\big)^{-1}$$

Note that $\delta = \gamma = 0$ gives non-differential misclassification, and λ , the logistic coefficient of the true AESI status, can be thought of as a quality parameter of the measurement, in the sense that both sensitivity and specificity will attain their “perfect” values 1 as $\lambda \rightarrow \infty$.

Similarly, to the “measurement” model above, the “structural” model is
$y \sim$ Bernoulli$\big(\pi(Y = 1|Z = z|)\big)$
With again the conditional probability $\big(\pi(Y = 1|Z = z|)\big)$ given by a logistic equation,

$$\pi(Y = 1|\, Z = z|) = \big(1 + \exp\big(-(\alpha + \beta z)\big)\big)^{-1}$$

Note that the parameter α determines the true event rate, and the logistic coefficient β is the log-odds ratio between vaccinated and unvaccinated groups, i.e. $\beta = 0$ corresponds to no true effect.

We will generate data from the model above using the conditions specified, and apply the following procedures to each replicate:

1. The standard analysis of relative risk, risk differences, and log-odds (without any modelling);
2. The above estimates derived from a maximum-likelihood based EM estimator of the parameters α and β of the above model, for given values of τ, λ, δ and γ (see following subsection)

We will summarize the resulting estimates by comparing them to the known true values. We will examine the following across conditions, for all estimators and target quantities:

- Bias (average difference between truth and estimate);
- Variance (average squared difference between estimate and average estimate);
- Mean squared error (average squared deviation between truth and estimate);
- Median absolute deviation;
- Computational issues (e.g. nonconvergence, inadmissible estimates).

In the first step, we will only evaluate performance of the standard analysis (1) above. In the second step we will develop an estimator for the model above and include (2) in the simulation study.

7.5.3 Model development and application to data

In this step, we develop an estimator for the model described above, extended with covariates,x. Since y is an unobserved ("latent", "hidden") variable, we will view the model as a mixture model and employ expectation-maximization. To obtain accurate standard errors, we will use analytical second derivatives.

A proof of concept of the EM estimator for this model can be found at https://github.com/daob/vaccine-misclassification/blob/main/R/estimators.R.

Note that in this model, it will be necessary to specify a value for the measurement parameters τ, λ, δ and γ , as these will not be identifiable. In a following step, we will allow the user to specific uncertainty about these parameters using priors in a Bayesian framework.

After the model is evaluated in the simulation study, we will assess based on these results whether it can sensibly be applied to the data at hand, and, if so, for which AESIs this might hold. We will then apply the developed modelling procedure to those situations, using input obtained from the literature overview where possible. Any possibilities of data analysis will be conducted in collaboration with work package 3.

In addition, we will perform sensitivity analysis varying the measurement parameters and γ to evaluate the robustness of relative risk, risk difference, and log-odds ratio estimates to (differential) misclassification.

7.5.4 Uncertainty propagation using Bayesian modelling.

Sensitivity analysis is a laborious process, and we are unlikely to obtain unequivocally trustworthy estimates of every AESI's misclassification properties. Therefore, our final aim will be to formulate the above model (including covariates) in the probabilistic programming language, allowing the user to specify uncertainty about the measurement parameters as probability distributions.

Standard R or Python libraries will then allow for estimation of this Bayesian version of the model using Hamiltonian Monte Carlo, or, if necessary, variational inference.

This will allow users to propagate both the effects of putative misclassification, as well as their uncertainty about the amount of such misclassification, into the final estimates of target quantities relative risk, risk difference, and log-odds ratios.

7.5.5 Misclassification: components analysis

The data sources participating in the readiness study are heterogeneous. In the readiness study, all the AESIs except DEATH are retrieved from the data banks that contribute diagnostic codes. When converting the data source to the ConcePTION CDM, DAPs label diagnostic codes with characteristics, named *meanings*, considered useful to characterize the context where the codes were recorded (for instance 'primary hospital diagnosis'). AESIs are retrieved based on specific lists of codes in various coding systems, possibly restricting records to those labelled with specific meanings, as agreed with DAPs.

In this analysis the background rate of each AESI is broken down per meaning of retrieved record, to assess the contribution of each component to the rate, and to provide a quantification of underestimation of the background rate itself. The methodology was developed in the ADVANCE study (Gini et al, 2020).

Study population

The study population for each AESI is composed by the persons who were in the data source at 1 January 2019 or entered during 2019, and who had no component recorded in the 365 days before study entry, or entered the data source at birth. Each person is followed from study entry until 31 December 2019.

Study variables

A component algorithm for an AESI is the algorithm extracting records based on a code list (either narrow, or possible) and restricting to a group of meanings.

The meanings used by each data source will be retrieved from all the AESIs, described and classified. A list of component algorithms will be created.

Analysis

The number of persons identified, and the cumulative incidence of each component and composite will be measured.

8. From CDM to Study variable tables (D3)

Operations using D2 CDM tables to define start and end of follow-up for each subject. Required tables and variables in the CDM

8.1 Required data in the v2.2 Conception CDM structure

DAPs may convert entire data bank to CDM tables, selections are done in the script. However, if a minimal data set is needed, please look at the requirements in the table below.

CDMv2.2:

Objective (describe)	**Required CDM tables to create study variables**	**Key requirements for ETL (to be filled by PI)** **Mandatory variables**
all	○ *Instance*	*source_table_name* *source_column_name* *included_in_instance* *date_when_data_last_updated* *since_when_data_complete* *up_to_when_data_complete* *restriction_in_values*
	○ *CDM source*	*data_access_provider_code* *data_access_provider_name* *data_source_name* *data_dictionary_link* *etl_link* *cdm_vocabulary_version* *cdm_version* *instance_number* *date_creation* *"recommended_end_date"*
	○ *Metadata*	*type_of_metadata* *tablename* *columnname* *other* *value*
	○ *Persons*	***persons to be included in the CDM instance: at least those persons in datasource with op_end_date > 1/1/2019 and most recent data*** *person_id* *day_of_birth* *month_of_birth* *year_of_birth* *day_of_death* *month_of_death* *year_of_death* *sex_at_instance_creation*
	○ *Observation periods*	***observation periods to be included in the CDM instance: at least all those with op_end_date>1/1/2018*** *person_id* *op_start_date* *op_end_date* *op_meaning* *op_origin*
	○ *Products*	*not needed*

Objective (describe)	Required CDM tables to create study variables	Key requirements for ETL (to be filled by PI) Mandatory variables
	o *Medicines*	***medicinal_product_atc_code to be included in the CDM instance all ATC codes in the Medicines code list from 2018 onwards*** *Variables to be provided for each record* *person_id* *medicinal_product_atc_code* *date_dispensing* *date_prescription* *disp_number_medicinal_product* *presc_quantity_per_day (when available)* *presc_quantity_unit* *presc_duration_days (when available)* *meaning_of_drug_record* *origin_of_drug_record*
	o *Vaccines*	***vx_atc to be included in the CDM instance: all J07*** *Fields to be completed* *person_id* *vx_record_date* *vx_admin_date* *vx_atc* *vx_type* *vx_text* *vx_dose* *vx_manufacturer* *vx_lot_num* *meaning_of_vx_record* *origin_of_vx_record*
	o *Events*	**event codes *to be included in the CDM instance*: at least those in *codesheet*** *Fields to be completed in table* *person_id* *start_date_record* *event_code* *event_record_vocabulary* *text_linked_to_event_code* *meaning_of_event* *origin_of_event*
	o *Visit occurrence*	*Not needed*
	o *Procedures*	
	o *Medical observations*	*Only if it contains diagnoses of interest*
	o *Survey observations* o *Survey_id*	*COVID-19 registers, Birth registers*
	o *Eurocat*	*When available*

CDM tables products, metadata, instance, cdm_source are always required

8.2 Creation of single component concept sets from CDM tables with codes

As a first step we will create the 'concepts' that are required for the studies using study code lists, these include (named as described in section 7.1.3 above)

- _AESI (Tagging Narrow) from the EVENTS table
- _AESI broad (Tagging Narrow & possible) from the EVENTS table

- Drugs (DP_) from the MEDICINES table
- _COV (covariates, tagging can be narrow, possible or empty) from the EVENTS table
- VACCINES from the VACCINES table (by vx_ATC or Vx_type)

In this step the function Create concept sets is used. The set of input tables are inspected and a group of datasets is created, each corresponding to a concept set. Each dataset contains the records of the input tables that match the corresponding concept set codes.

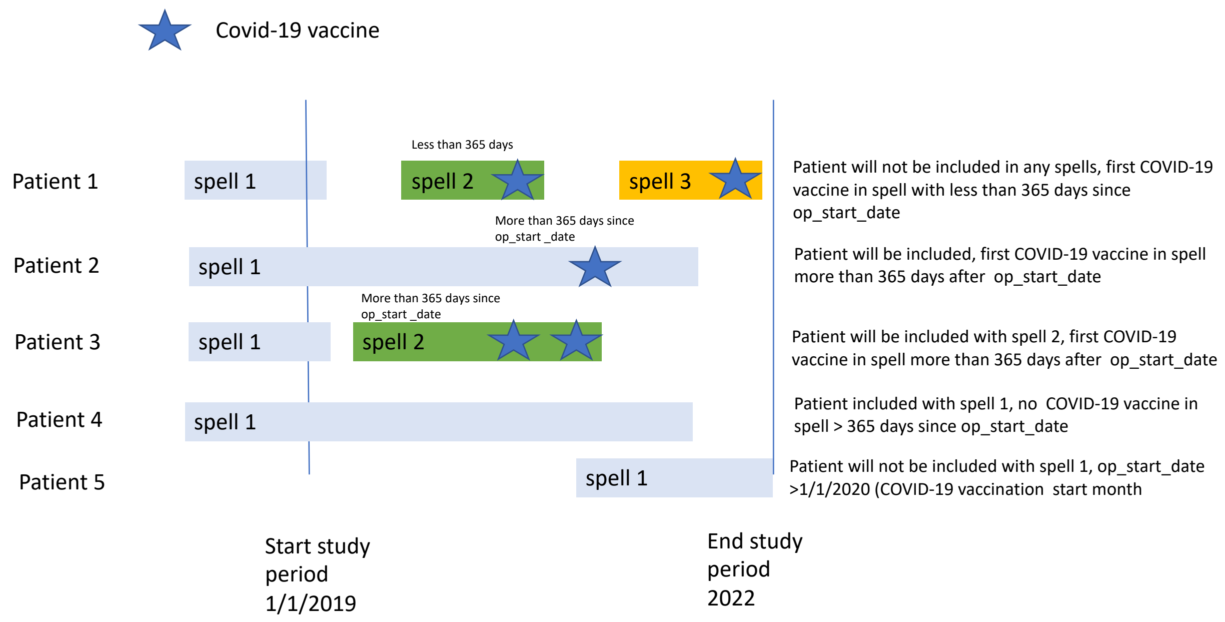
8.3 Spells (observation periods)

Step 1: Processing of observation periods table

To create the date of entry and date of exit into follow-up of each person, the table OBSERVATION_PERIODS is processed using the function CreateSpells. The table may contain multiple records per person, possibly overlapping. Gaps between spells are allowed for Tuscany (180 days). In many data sources persons have only one observation period, but in regional NHS data sources in Tuscany and Lazio, observation periods may differ if people transfer from paediatrician to GP, or between GPs, as the persons identifier is maintained but it is not clear whether the person has left or not.
Quality control: count persontime (op_end_date-op_start_date) prior to CreateSpells and after wards

Step 2: Choice of the proper spell in case there are multiple

To avoid gaps in information, we will use only one spell per person and this should be the spell **containing the first COVID-19 vaccination**, within that spell the person must have at least 365 days of follow-up, prior to time zero and op_start_date should be before 1/1/2020.



Patient will not be included in any spells, first COVID-19 vaccine in spell with less than 365 days since op_start_date

Patient will be included, first COVID-19 vaccine in spell more than 365 days after op_start_date

Patient will be included with spell 2, first COVID-19 vaccine in spell more than 365 days after op_start_date

Patient included with spell 1, no COVID-19 vaccine in spell > 365 days since op_start_date

Patient will not be included with spell 1, op_start_date >1/1/2020 (COVID-19 vaccination start month

Figure 4: Examples how to choose spells.

The input for the CreateSpells function is the data in D2 Observation Periods person_id op_start_date, op_end_date (in case there subpopulations this may be more):

- if op_end_date is empty, replace it with the minimum of the following dates: recommended_end_date, as retrieved from CDM_SOURCE,
- date of death (PERSONS),
- end of study period (31-12-2022)
- date_creationD (from CDM_SOURCE)
- or recommended_end_date (from CDM_SOURCE)

output of CreateSpells will provide the following variables

- id: variable containing the identifier of the unit of observation
- spell_start_date: variable containing the start date
- spell_end_date: variable containing the end date
- spell_num: number of spell (the first has number 1)

As per figure 4: establish link the output of CreateSpells with the Vaccines concepts, and establish which spell needs to be selected.

8.4 Create prompt and item sets

Medical observations and survey observations contain information of relevance for some of the required variables: e.g. COVID-19 and pregnancies.

For COVID-19 we will create a D3_ TD_variable_COVID table (Table 9)

Table 9 D3_TD_variable_COVID Unit of observation: COVID-19 episode

VarName	**Description**	**Format**	**Vocabulary**
person_id	unique person identifier	character	
date	Date of the covid episode		

value_of_variable		categorical	severity1 = not hospitalised, no ICU; no death
			severity2 = hospitalised, no ICU, no death
			severity3 = ICU, no death
			severity4 = death

For pregnancies we will use the IMI-ConcePTION pregnancy algorithm as described in the following repository https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies

We will use the *D3_Included pregnancies* which is an output of the algorithm (see table 8) for structure.

Table 10: D3_included pregnancies Unit of observation: pregnancies that were completed or ongoing during the period of time included in the datasource instance.

Variable name	Description	Type	Vocabulary	Notes
pregnancy_id	unique identifier of a pregnancy	string	-	unit of observation. created with this table.
person_id		as in the CDM	as in the CDM	there may be multiple records per person
age_at_start_of_pregnancy				
pregnancy_start_date	best estimate of the date of pregnancy start			
pregnancy_end_date	best estimate of the date of pregnancy end			
meaning_start_date	method by which pregnancy_start_date was obtained		*to be decided*	
meaning_end_date	method by which pregnancy_end_date was obtained		*to be decided*	

type_of_pregnancy_end			LB = livebirth
			SB = stillbirth
			SA = spontaneous abortion
			T = termination
			ECT = ectopic pregnancy
			MD = maternal death
			UNK = unknown
				LB means that there is at least a life birth
				T means that there is a medical intervention, irrespective of the cause

date_of_principal_record			date when the principal record was recorded	
date_of_oldest_record			oldest record date of the group	this tells when the pregnancy could have been first identified in the data source
date_of_oldest_record_with_recorded_start_of_pregnancy				
nature_of_principal_record			description of the principal record of the group of records	
nature_of_oldest_record			description of the record with the oldest date in the group of records	
algorithm_for_reconciliation	Which choices had to be made to reconcile the overlapping pregnancies		see section 3.4.3, 'merge streams of the same person', Box 2	
detected_while_ongoing	Whether the pregnancy was detected while it was still ongoing		'yes' = yes 'no' = no	
time_since_start_when_detected	if detected_while_ongoing =='yes', this variable captures the number of days between pregnancy_start_date and date_of_oldest_record		integer	
error_of_start_date_at_date_of_oldest_record	difference between pregnancy_start_date and the pregnancy_start_date of the oldest record			this is the error that is done when the pregnancy is first detected; if it is > 0, the start of pregnancy was too early; if it is < 0 the start of

Variable name	Description	Type	Vocabulary	Notes
				pregnancy was too late
PROMPT	whether the pregnancy was included by the PROMPT stream		yes’ = yes ‘no’ = no	
CONCEPTSET	whether the pregnancy was included by the CONCEPTSET stream		yes’ = yes ‘no’ = no	
EUROCAT	whether the pregnancy was included by the EUROCAT stream		yes’ = yes ‘no’ = no	
ITEMSET	whether the pregnancy was included by the ITEMSET stream		yes’ = yes ‘no’ = no	
number_of_records_in_the_group	number of records in the group that originated the pregnancy			
reconciliation_type	Type of group that generated the pregnancy			
number_green	number of green records in the group that originated the pregnancy			
number_yellow	number of yellow records in the group that originated the pregnancy			
number_blue	number of blue records in the group that originated the pregnancy			
number_red	number of red records in the group that originated the pregnancy			
survey_id_1 …	survey identifier of the inclusion criteria for the pregnancy, if any			
visit_occurrence_id_1 …	visit occurrence of the inclusion criteria for the pregnancy, if any			

8.5 Study population

From the retained spells and the CDM_PERSONS and Concepts-Vaccines each subject gets assigned dates in the **D3_Total_Study_ population, based on the concept sets variables and cleaning/completion thereof the study population will be completed.**

- **Vaccines table cleaning**: order chronologically and assign an imputed dose if vx_dose is empty, distance between dose 1 and 2 should be 15 days, between dose 3 and 2 at least 60 days and between dose 4 and 3 at least 60 days.
- **Events table**: provide a code count for the events in the study population
- Combination of concepts sets and item sets or prompts, into **algorithms (e.g. TTS)**

Subpopulations

In case that organizations have data banks with different recommended end dates,or creation dates ,or different underlying populations, these will be run separately, this may be the case for BIFAP and SIDIAP.

D3_Total_study _population

The unit of observation of this table is unique *persons*, this table will be the basis for all analyses that are needed for WP3 and WP4.

Table 11 Structure of D3_Total_study _population (upper) and of the time-dependent datasetD3_TD_*condition (lower)*, for each time-dependent condition of interest *condition*

VarName	Description	Format	Vocabulary
person_id	person identifier	string	
spell_start_date	start of the observation period where the person was observed and that is used to identify the person study period	date	
study_exit_date	exit from the study	date	
start_followup_study	entry in the study (same as D4_study_population/study_entry_date)	date	
date_vax_1	date of vaccination 1	date	
date_vax_2	date of vaccination 2	date	
date_vax_3	date of vaccination 3	date	
date_vax_4	date of vaccination 4	date	
type_vax_1	manufacturer of vaccination 1	string	
type_vax_2	manufacturer of vaccination 2	string	
type_vax_3	manufacturer of vaccination 3	string	
type_vax_4	manufacturer of vaccination 4	string	
sex	gender at the moment when the instance was created	string	
date_of_birth	date of birth	date	
date_of_death	date of death	date	

VarName	Description	Format	Vocabulary
person_id	unique person identifier	character	from D4_study_population

date			date when the condition changes; the components of the condition last 365 days if they are diagnosis, and 90 days if they are drug proxies; unique spells are created when the algorithm is 1 (if either a dianosis or a drug proxy is active), and the algorithm is reverted to values 0 whenever no component is active
value_of_variable		binary	1 = at least one of the components of the algorithm that define the condition is active 0 = otherwise

Covariates, drug proxies are listed in the methods section and comprise: Pregnant, Cancer, Chronic kidney disease, Chronic liver disease, Chronic respiratory disease, Cardio-cerebrovascular disease, Obesity, Down syndrome, Mental disorders, Sickle cell disease, Diabetes (1&2), HIV, Immunosuppressants, Any risk factor. Use of medicines< 90 days Anti-thrombotics, Sex hormones, Antibiotics, Antivirals, Lipid lowering drugs, Other Vaccines. They are all stored in a separate dataset, to allow the value to be time-dependency

From the D3 study population matching, sampling and other designs can be applied for all the proposed analyses in T3. This will be created by the responsible groups.

8.6 SCRI dataset (D3 & D4)

The dataset is created in the T3 step from the D3 dataset for use in conditional Poisson regression analysis. The dataset is created inside the 'scri' function and can eventually be saved as a dataset.

A separate D4 dataset can be stored for each SCRI model.

An example of such a dataset for various risk/control windows and myocarditis as an event with additional variables: the first, second, third brands, sex, age categories, use of medication A, distance between doses, categorized time in case of calendar time adjustment.

In this dataset there are multiple rows per person. All variables are constant for each row/interval from rw_start to rw_end.

These variables can be used in SCRI analysis to evaluate the effects of these variable or interactions by combining the corresponding categorical variables on myocarditis (or other events).

Table 12 Structure of D4_SCRI_data

Variable	Description	Format	Vocabulary	Rules
person_id	unique person identifier	character	from cdm persons	from cdm persons
DAP	DAP name	character		From CDM_Instance
rw_start	Start of period	Numeric		
rw_end	End of period	Numeric		
interval	The period length	Numeric		
lab	Risk/control windows names	Character	recoded	Example: pre-exposure[-90;-30] buffer[-29;-1] dose 1 [0;0] dose 1 [1;7]

Variable	Description	Format	Vocabulary	Rules
				dose 1 [8;14] dose 1 [15;28] dose 1 [29;60] dose 1 >60 dose 2 [0;0] dose 2 [1;7] dose 2 [8;14] dose 2 [15;28] dose 2 [29;60] dose 2 >60
br1	COVID-19 vaccine manufacturer 1	Character		Example: no_vax Pfizer no_Pfizer
br2	COVID-19 vaccine manufacturer 2	Character		Example: Moder no_vax no_Moder
br3	COVID-19 vaccine manufacturer 3	Character		Example: Astra no_vax no_Astra
br4	COVID-19 vaccine manufacturer 4	Character		Example: J&J no_vax no_J&J
…	…	…		…
Age30_50	Categorized age variable			Example: age(-1,30] age(30,50] age(50,Inf]
age30	Categorized age variable			Example: age(-1,30] age(30,Inf]
sex	sex at instance creation	Character	from cdm persons	M- Male F-Female O- other sex U- Unknown
dose12_diff	Number of days between the first and the second doses	Numeric		
dose12_diff_cat	Categorized dose12_diff variable			Example: dist12:(-Inf,-1] dist12:(-1,21] dist12:(21,35] dist12:(35,Inf]
dose23_diff	Number of days between the third and the second doses	Numeric		
dose23_diff_cat	Categorized dose23_diff variable			Example: dist12:(-Inf,-1] dist12:(-1,84] dist12:(84,Inf]
Med_A	Use of medication during this period, i.e., from rw_start to rw_end	Numeric		0 – no use 1 - use
cal_time_cat	Splitting the observation period into intervals of, for example, 10 [days] to adjust for calendar time	Character		Example: [11; 20] [21; 30] [31; 40] [41; 50] ….

Variable	Description	Format	Vocabulary	Rules
myocarditis	Number of myocarditises in this period, i.e., from rw_start to rw_end			
myocarditis_date	Date of myocarditis	yyyymmdd		

9. References

Barda N, Dagan N, Cohen C, Hernán MA, Lipsitch M, Kohane IS, Reis BY, Balicer RD. Lancet. 2021 Dec 4;398(10316):2093-2100. doi: 10.1016/S0140-6736(21)02249-2.

Becker BFH, Avillach P, Romio S, van Mulligen EM, Weibel D, Sturkenboom MCJM, Kors JA; ADVANCE consortium. CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project. Pharmacoepidemiol Drug Saf. 2017 Aug;26(8):998-1005. doi: 10.1002/pds.4245. Epub 2017 Jun 28. PMID: 28657162; PMCID: PMC5575526.

Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf. 2010 Jun;19(6):537-54. doi: 10.1002/pds.1908. PMID: 20354968; PMCID: PMC2886161.

Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, Hernán MA, Lipsitch M, Reis B, Balicer RD. N Engl J Med. 2021 Apr 15;384(15):1412-1423. doi: 10.1056/NEJMoa2101765.

Gini R, Dodd CN, Bollaerts K, Bartolini C, Roberto G, Huerta-Alvarez C, Martín-Merino E, Duarte-Salles T, Picelli G, Tramontan L, Danieli G, Correa A, McGee C, Becker BFH, Switzer C, Gandhi-Banga S, Bauwens J, van der Maas NAT, Spiteri G, Sdona E, Weibel D, Sturkenboom M. Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project. Vaccine. 2020 Dec 22;38 Suppl 2:B56-B64. doi: 10.1016/j.vaccine.2019.07.045. Epub 2019 Oct 31. PMID: 31677950.

Groenwold RH, Hak E, Klungel OH, Hoes AW. Instrumental variables in influenza vaccination studies: mission impossible?! Value Health. 2010 Jan-Feb;13(1):132-7. doi: 10.1111/j.1524-4733.2009.00584.x. Epub 2009 Aug 20. PMID: 19695007.

Groenwold RH. Falsification end points for observational studies. Jama. 2013;309(17):1769-70.
Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? Jama. 2013;309(3):241-2.

Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006 Jul;17(4):360-72. doi: 10.1097/01.ede.0000222409.00878.37. Erratum in: Epidemiology. 2014 Jan;25(1):164. PMID: 16755261.

Labrecque J, Swanson SA. Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. Curr Epidemiol Rep. 2018;5(3):214-220. doi: 10.1007/s40471-018-0152-1. Epub 2018 Jun 22. PMID: 30148040; PMCID: PMC6096851.

Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014 Dec;43(6):1969-85. doi: 10.1093/ije/dyu149. Epub 2014 Jul 30. PMID: 25080530.

Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. PLoS One. 2018 Feb 1;13(2):e0192033. doi: 10.1371/journal.pone.0192033. PMID: 29389968; PMCID: PMC5794136.

Thurin NH, Pajouheshnia R, Roberto G, Dodd C, Hyeraci G, Bartolini C, Paoletti O, Nordeng H, Wallach-Kildemoes H, Ehrenstein V, Dudukina E, MacDonald T, De Paoli G, Loane M, Damase-Michel C, Beau AB, Droz-Perroteau C, Lassalle R, Bergman J, Swart K, Schink T, Cavero-Carbonell C, Barrachina-Bonet L, Gomez-Lumbreras A, Giner-Soriano M, Aragón M, Neville AJ, Puccini A, Pierini A, Ientile V, Trifirò G, Rissmann A, Leinonen MK, Martikainen V, Jordan S, Thayer D, Scanlon I, Georgiou ME, Cunnington M, Swertz M, Sturkenboom M, Gini R. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. Clin Pharmacol Ther. 2022 Jan;111(1):321-331. doi: 10.1002/cpt.2476. Epub 2021 Nov 26. PMID: 34826340.

Sturkenboom M, Braeye T, van der Aa L, Danieli G, Dodd C, Duarte-Salles T, Emborg HD, Gheorghe M, Kahlert J, Gini R, Huerta-Alvarez C, Martín-Merino E, McGee C, de Lusignan S, Picelli G, Roberto G, Tramontan L, Villa M, Weibel D, Titievsky L. ADVANCE database characterisation and fit for purpose assessment for multi-country studies on the coverage, benefits and risks of pertussis vaccinations. Vaccine. 2020 Dec 22;38 Suppl 2:B8-B21. doi: 10.1016/j.vaccine.2020.01.100. Epub 2020 Feb 12. PMID: 32061385.

Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383-8.

Uddin MJ, Groenwold RH, Ali MS, de Boer A, Roes KC, Chowdhury MA, et al. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. Int J Clin Pharm. 2016;38(3):714-23.

Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013;36 Suppl 1:S33-47.

Hauben M, Aronson JK, Ferner RE. Evidence of Misclassification of Drug-Event Associations Classified as Gold Standard 'Negative Controls' by the Observational Medical Outcomes Partnership (OMOP). Drug Saf. 2016;39(5):421-32.

Uddin MJ, Groenwold RH, Ton de Boer, Belitser SV, Roes KC, et al. (2015) Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods. Pharm Anal Acta 6:353.

Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf. 2006 May;15(5):291-303. doi: 10.1002/pds.1200. PMID: 16447304.

Sjölander A, Greenland S. Are E-values too optimistic or too pessimistic? Both and neither! Int J Epidemiol. 2022 Mar 1:dyac018. doi: 10.1093/ije/dyac018. Epub ahead of print. PMID: 35229872.

VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med. 2017 Aug 15;167(4):268-274. doi: 10.7326/M16-2607. Epub 2017 Jul 11. PMID: 28693043.

10. Table shells readiness phase

Table 1: Attrition table

Condition	DAp1	DAp2	Dap3	Dap4	Dap5	Dap6	Dap7	Dap8	Dap9	Dap10
Type of datasource										
Recommended end date										
Coding systems for diagnoses										
Attrition										
Persons in the instance of the data source, N										
Sex or birth date missing or absurd, no dates of entry or exit, N										
Death before 1/1/2019, N										
Exit from the data source before 1/1/2019, N										
Persons in the data source at or after 1/1/2019, N										
Less than 365 days history at any point in time after 1.1.2019 (and not born after 1.1.2019), N										
Final study population, N										

Table 2: Code counts

DAP	Name event	Code	Meaning	Count

Table 3: population characteristics at entry

		DAP1	DAP2	Etc.							
Study population	N										
follow-up (years)	PY										
Age in years	Min										
	P25										
	P50										
	Mean										
	P75										
	Max										
Age in categories	0-4										
	5-11										
	12-17										
	18-24										
	25-29										
	30-39										
	40-49										
	50-59										
	60-69										
	70-79										
	80+										
	60+										
Person years across sex	Male										
	Female										
At risk population at January 1-2020	Cardiovascular disease										
	Cancer										
	Chronic lung disease										
	HIV										
	Chronic kidney disease										
	Diabetes										
	Severe obesity										
	Sickle cell disease										
	Use of immunosuppressants										
	Any risk factors										
	Xx other risk factors										

Table 4: Characteristics of populations at first vaccination

Variable	Values	Baseline 1/1/2020 total population	Pfizer 1st dose	Moderna 1st dose	AstraZeneca 1st dose	J&J	Novavax 1st dose	Unknown 1st dose
Study population	N							
follow-up (years)	PY							
Age in years	Min							
	P25							
	P50							
	Mean							
	P75							
	Max							
Age in categories	0-4							
	5-Nov							
	Dec-17							
	18-24							
	25-29							
	30-39							
	40-49							
	50-59							
	60-69							
	70-79							
	80+							
	60+							
Person years across sex	Male							
	Female							
At risk population at January 1-2020	Cardiovascular disease							
	Cancer							
	Chronic lung disease							
	HIV							
	Chronic kidney disease							
	Diabetes							
	Severe obesity							
	Sickle cell disease							
	Use of immunosuppressants							
	Any risk factors							

Table 5: Dosing regimens

Dose	Measure	Dap1		Dap2		Dap3		Dap4		Dap5		Dap x	
Study population		N	%	N	%	N	%	N	%	N	%	N	%
AstraZeneca dose 1, % of total population	Persons												
AstraZeneca dose 2, % of 1st dose	Persons												
Other vaccine dose 2, % of first dose	Persons												
Other vaccine dose 2, % of 1st dose	Persons												
type other dose 2 Pfizer	Persons												
type other dose 2 Moderna													
Amongst persons with AstraZeneca dose 2 distance	Min												
Amongst persons with AstraZeneca dose 2 distance	P25												
Amongst persons with AstraZeneca dose 2 distance	P50												
Amongst persons with AstraZeneca dose 2 distance	P75												
Amongst persons with AstraZeneca dose 2 distance	Max												
Amongst persons with other dose 2 distance	Min												
Amongst persons with other dose 2 distance	P25												
Amongst persons with other dose 2 distance	P50												
Amongst persons with other dose 2 distance	P75												
Amongst persons with other dose 2 distance	Max												
AZ dose 3, % of 1st dose	Persons												
other dose 3, % of 1st dose	Persons												
type other dose 3 Pfizer	Persons												
type other dose 3 Moderna	Persons												
Amongst persons with AstraZeneca dose 3 distance	Min												
Amongst persons with AstraZeneca dose 3 distance	P25												
Amongst persons with AstraZeneca dose 3 distance	P50												
Amongst persons with AstraZeneca dose 3 distance	P75												
Amongst persons with AstraZeneca dose 3 distance	Max												
Amongst persons with other dose 3 distance	Min												
Amongst persons with other dose 3 distance	P25												

Amongst persons with other dose 3 distance	P50												
Amongst persons with other dose 3 distance	P75												
Amongst persons with other dose 3 distance	Max												
Janssen dose 1, % of total population	Persons												
Janssen dose 2, % of 1st dose	Persons												
Other vaccine dose 2, % of 1st dose	Persons												
type other dose 2 Pfizer													
type other dose 2 Moderna													
type other dose 2 AZ													
Amongst persons with dose 2 distance	Min												
Amongst persons with dose 2 distance	P25												
Amongst persons with dose 2 distance	P50												
Amongst persons with dose 2 distance	P75												
Amongst persons with dose 2 distance	Max												
Pfizer dose 1, % of total population	Persons												
Pfizer dose 2, % of 1st dose	Persons												
Other vaccine dose 2, % of first dose	Persons												
Other vaccine dose 2, % of 1st dose	Persons												
type other dose 2 AZ	Persons												
type other dose 2 Moderna	Persons												
XX	Persons												
Amongst persons with Pfizer dose 2 distance	Min												
Amongst persons with Pfizer dose 2 distance	P25												
Amongst persons with Pfizer dose 2 distance	P50												
Amongst persons with Pfizer dose 2 distance	P75												
Amongst persons with Pfizer dose 2 distance	Max												
Amongst persons with other dose 2 distance	Min												
Amongst persons with other dose 2 distance	P25												
Amongst persons with other dose 2 distance	P50												
Amongst persons with other dose 2 distance	P75												
Amongst persons with other dose 2 distance	Max												
Pfizer dose 3, % of 1st dose	Persons												
other dose 3, % of 1st dose	Persons												
type other dose 3 Moderna	Persons												
type other dose 3 AZ	Persons												

Amongst persons with Pfizer dose 3 distance	Min												
Amongst persons with Pfizer dose 3 distance	P25												
Amongst persons with Pfizer dose 3 distance	P50												
Amongst persons with Pfizer dose 3 distance	P75												
Amongst persons with Pfizer dose 3 distance	Max												
Amongst persons with other dose 3 distance	Min												
Amongst persons with other dose 3 distance	P25												
Amongst persons with other dose 3 distance	P50												
Amongst persons with other dose 3 distance	P75												
Amongst persons with other dose 3 distance	Max												
Moderna dose 1, % of total population	Persons												
Moderna dose 2, % of 1st dose	Persons												
Other vaccine dose 2, % of first dose	Persons												
Other vaccine dose 2, % of 1st dose	Persons												
type other dose 2 Pfizer	Persons												
type other dose 2 AZ	Persons												
XX	Persons												
Amongst persons with Moderna dose 2 distance	Min												
Amongst persons with Moderna dose 2 distance	P25												
Amongst persons with Moderna dose 2 distance	P50												
Amongst persons with Moderna dose 2 distance	P75												
Amongst persons with Moderna dose 2 distance	Max												
Amongst persons with Moderna dose 2 distance	Min												
Amongst persons with Moderna dose 2 distance	P25												
Amongst persons with Moderna dose 2 distance	P50												
Amongst persons with Moderna dose 2 distance	P75												
Amongst persons with Moderna dose 2 distance	Max												
Moderna dose 3, % of 1st dose	Persons												
other dose 3, % of 1st dose	Persons												
type other dose 3 Pfizer	Persons												
type other dose 3 AZ	Persons												

Amongst persons with Moderna dose 3 distance	Min												
Amongst persons with Moderna dose 3 distance	P25												
Amongst persons with Moderna dose 3 distance	P50												
Amongst persons with Moderna dose 3 distance	P75												
Amongst persons with Moderna dose 3 distance	Max												
Amongst persons with other dose 3 distance	Min												
Amongst persons with other dose 3 distance	P25												
Amongst persons with other dose 3 distance	P50												
Amongst persons with other dose 3 distance	P75												
Amongst persons with other dose 3 distance	Max												

Table 6: Incidence rates

type	AESI	Group	Characteristic	Dap1	PYs	IR (CI 95%)	Dapn	PYs	IR (CI 95%)
AESI		2019 before COVID-19 infection	Total (all ages/gender)						
AESI			Total age Standardized						
AESI		2020 before COVID-19 infection	Total (all ages/gender)						
AESI			Total age Standardized						
AESI		2020 after COVID-19 infection before vaccination	Total (all ages/gender)						
AESI			Total age Standardized						