Covid Vaccines Effectiveness (CoVE)

Effectiveness of heterologous and booster Covid-19 vaccination in 5 European countries, using a cohort approach in children and adults with a full primary Covid-19 vaccination regimen

Specific Contract No 01 implementing framework contract No EMA/2020/46/TDA/L5.06.

Deliverable D2 - Study protocol

Week 05

V 0.4

Disclaimer & acknowledgement

The research leading to these results was conducted as part of the activities of the EU PE&PV (Pharmacoepidemiology and Pharmacovigilance) Research Network (led by Utrecht University) with collaboration from the Vaccine Monitoring Collaboration for Europe network (VAC4EU).

The project has received support from the European Medicines Agency under the Framework service contract nr EMA/2020/46/TDA/L5.06.

EU PE&PV research network

Title	Effectiveness of heterologous and booster Covid-19 vaccination in 5 European countries, using a cohort approach in children and adults with a full primary Covid-19 vaccination regimen
Protocol version identifier	0.3
Date of last version of protocol	May 31, 2022
EU PAS register number	
Active substance	*NA*
Medicinal product	*NA*
Product reference	*NA*
Procedure number	*NA*
Marketing authorisation holder(s)	*NA*
Research question and objectives	The goal of this study is to assess the effectiveness and waning of immunity of primary Covid-19 vaccinations and the booster in preventing different Covid-19 outcomes.
Country(-ies) of study	The Netherlands, the United Kingdom, Italy, France and Spain
Author	Dr. Elisa Martin, AEMPS Prof. Dr. Lamiae Grimaldi, APHP Dr. Tiago Vaz, UMCU Dr. Satu Johanna Siiskonen, UU Dr. Fabio Riefolo, Teamit Dr. Rosa Gini, ARS
Contributors	ROC12 Consortium

Table of Contents

Document history

Name	Date	Version
Dr. Elisa Martin, AEMPS Prof. Dr. Lamiae Grimaldi, APHP Dr. Tiago Vaz , UMCU Dr. Satu Johanna Siiskonen, UU Dr. Fabio Riefolo, Teamit Dr. Rosa Gini, ARS	17-05-2022	V0.1 1st draft
ROC12 Consortium	27-05-2022	V0.2 Implemented with Consortium review comments
Dr. Elisa Martin, AEMPS Prof. Dr. Lamiae Grimaldi, APHP Dr. Tiago Vaz , UMCU Dr. Satu Johanna Siiskonen, UU Dr. Fabio Riefolo, Teamit	31-05-2022	V0.3 Final version submitted to EMA
Dr. Elisa Martin, AEMPS Dr. Fabio Riefolo, Teamit	15-06-2022	V0.4 Implemented with EMA comments

1 Responsible parties

Organisation	Key person	Role(s)
Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)	Elisa Martin, PI	Core Scientific team, WP1 lead
l'Assistance Publique-Hôpitaux de Paris (APHP)	Lamiae Grimaldi	Core Scientific team WP1 and Data Access Provider SNDS
University Medical Center Utrecht (UMCU), The Netherlands	Tiago Vaz	Core Scientific team, WP2 lead
Utrecht University, Pharmacoepidemiology & Clinical Pharmacology	1.Olaf Klungel 2.Satu Siiskonen	1.lead EU PE&PV research network 2.Consortium Management and liaison EMA
Teamit Institute, S.L.	Fabio Riefolo	Core Scientific team, Project Management

1.1 Key collaborators and roles

Organisation	Key person	Role(s)
Utrecht University, Pharmacoepidemiology & Clinical Pharmacology	Patrick Souverein Helga Gardarsdottir	Data Access Provider (ETL) & expertise
VAC4EU	1.Daniel Weibel, secretary general 2.Patrick Mahy, treasurer	1.Provide access and use of tools, coordination 2.subcontracting
Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)	Mar Martin Patricia García Belen Castillo Miguel-Angel Macia	Data Access Provider BIFAP Statistician
UMC Utrecht (UMCU).	1. Albert Royo 2. Vjola Hoxhaj 3. Judit Riera 4. Ivonne Martin	1. programmer R 2. programmer & review data quality checks 3. Review data quality checks 4. Statistician
PHARMO Institute for Drug Outcomes Research.	Ronald Herings Karin Swart-Polinder Jetty Overbeek Jesse van den Berg	Data Access Provider PHARMO & expertise
Agenzia Regionale di Sanità (ARS)	Rosa Gini Davide Messina Olga Paoletti	Data Access Provider ARS & Programming expertise
Società Servizi Telematici -Pedianet.	Carlo Giaquinto Elisa Barbieri Luca Stona	Data Access Provider Pedianet, expertise
IDIAP JGol	Felipe Villalobos	Data Access Provider SIDIAP, expertise
Academic Spin-Off "INSPIRE" srl	Gianluca Trifiro Ylenia Ingrasciotta Valentina Ientile Salvatore Crisafulli	Data Access Provider, expertise
Instituto Aragonés de Ciencias de la Salud.	Antonio Gimeno Beatriz Poblador Mercedes Aza Aida Moreno Jonás Carmona	Scientific support protocol, statistics, report
Navarre Health Service.	Juan Erviti Leire Leache Luis Carlos Saiz Marta Gutiérrez Valencia	Scientific support protocol, statistics, report
Drug Safety Research Unit	Liz Lynn Dr. Debabrata Roy	Scientific support protocol, statistics, report

RESEARCH TRIANGLE INSTITUTE	Bradley Layton Xabier Garcia de Albeniz Susana Perez-Gutthann Rachel Weinrib	Scientific support protocol, statistics, report
Institute of Public Health, Riga Stradins University	Anda Kivite-Urtane	Scientific support protocol, statistics, report
Democritus University of Thrace	Christos Kontogiorgis Foteini Dermiki-Gkana	Scientific support protocol, statistics, report
National Public Health Agency (RIVM)	Susan Hahne	Scientific support protocol, statistics, report
Teamit Institute, S.L.	Eva Molero Gianmarco Di Mauro	Project Management

2 Abstract

Title

Covid Vaccines Effectiveness (CoVE): Effectiveness of heterologous and booster Covid-19 vaccination in 5 European countries, using a cohort approach in children and adults with a full primary Covid-19 vaccination regimen

Rationale and background

Real-world effectiveness data demonstrated that Covid-19 vaccines' protection against severe SARS-CoV-2 infection is high in the short term but wanes over time, also depending on the virus variants. This study will deepen the real-world data effectiveness evidence of heterologous, homologous, and booster vaccination different regimes on a large population scale.

Research questions and objectives

The goal of this study is to assess the effectiveness and waning of immunity of primary Covid-19 vaccinations and the booster in preventing different covid-19 outcomes.

Primary objectives

To estimate the effectiveness and waning of effectiveness in:

- adults and adolescents, separately, between heterologous and homologous primary vaccinations.
- children between homologous primary vaccinations and non-vaccination.
- adults with full homologous primary regimen between those with a homologous booster and heterologous booster, separately, and those without any booster.
- adults with full heterologous primary regimen between those with and without any booster.

Secondary objective

To estimate the effectiveness against all-cause mortality in adults aged 60+ with a homologous or heterologous full primary regimen between those with or without any booster.

Study design

The study design is a retrospective multi-database cohort study that aims to complement common test-negative case-control studies. The heterologous primary vaccination cohort (different 1st and 2nd dose vaccine brands) will be compared with matched homologous cohort (same 1st and 2nd dose vaccine brand) among adults, and adolescents or with non-vaccinated children. Booster vaccinees will be split into homologous or heterologous (same or different vaccine brands between primary scheme and booster, respectively) and compared with non-boosted persons.

Population

Persons are eligible to participate in the study if they have 2 years of valid data upon first Covid-19 vaccination.

Variables

The date of cohort entry (t=0) is the date of the 2nd dose for the primary vaccination regimen, the booster vaccination date for the boosted vaccinees, or the same calendar date of non-boosted individuals in the same data source, who are matched on time0, birth year, sex and region. The main exposure of interest is the receipt of a different primary regimen or booster Covid-19 vaccine, the dose, and its brand/manufacturer. This study will consider different Covid-19 outcomes: severe Covid-19, Covid-19-related death, all Covid-19 infections, and all-cause mortality (secondary objective).

Data Sources

The study will include data from 8 health care data sources across 5 European countries (Spain, Italy, Netherlands, France, and United Kingdom) which could link event of research interest with vaccination data . Data sources will capture outcomes from hospitalization and/or general practice.

Study Size

The source population will comprise all patients meeting the eligibility criteria, being approximately more than 98 millions, of which 67 millions will have a complete primary vaccination.

Data Analyses

Distributions of baseline and Covid-19 vaccination characteristics at time0 will be assessed in all-Covid-19 vaccinated population and matched populations.

Incident rates differences (95% confidence intervals) of each Covid-19 outcome for both primary vaccination matched and booster/non-booster matched cohorts estimated by overall, age groups, brands, and time since (booster-)time0 will be estimated. IPW-weight Kaplan-Meier curves will be generated to depict the cumulative incidence of the outcomes by matched cohorts over time after (booster-)time0. Cox proportional hazards regression (95% confidence intervals) to derive the average hazard ratio (HR) of Covid-19 related outcomes will be produced. The adjusted vaccine effectiveness for all the outcomes and all-cause death will be estimated as 1 minus the adjusted HR (and 1-95% confidence intervals) for age groups, overall matched cohorts, and brands.

Random-effects meta-analyses using the main estimates from each data source. Different access to Covid-19 testing (restricting to patients with a negative tests) and healthy vaccinee effect will be investigated in sensitivity analyses.

3 Amendments and updates

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

4 Timelines, deliverables and milestones

Deliverable	Time delivery to EMA
D1 Study Plan	4 May 2022
D2 Study Protocol	25 May 2022
D3 Study report & annexes	20 September 2022
D4 Manuscript	20 September 2022

5 List of abbreviations

AEMPS	Spanish Agency for Medicines and Medical Devices
BIFAP	Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria
CI	confidence interval
COVID-19	illness caused by the SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) virus
eHR	electronic health records
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU PAS Register	European Union Electronic Register of Post-Authorisation Studies
GPP	Good Pharmacoepidemiology Practices
GVP	Good Pharmacovigilance Practices
HCU	health care utilisation
HR	hazard ratio
ICPC-2	*International Classification of Primary Care, 2nd Edition*
ICD-9	*International Classification of Diseases, 9th Revision*
ICD-10	*International Classification of Diseases, 10th Revision*
ICU	intensive care unit
IRR	incidence rate ratio
ISPE	International Society for Pharmacoepidemiology
PASS	Post-authorisation safety study
RCT	Randomised clinical trial
RD	risk difference
RR	risk ratio
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology

6 Rationale and Background

Real-world effectiveness data has demonstrated high levels of short-term protection by Covid-19 vaccines against clinical disease and, more so, against severe outcomes including hospitalization and death (1–9). Unfortunately, evidence shows that protection against symptomatic disease wanes over time and depends on circulating variants (1–9). Andrews et al. recently showed the effectiveness of booster vaccination in the UK until December 2021, just prior to the spread of Omicron (9).

Booster/additional doses have now been implemented in many countries to combat the serious effects of the Delta variant and the highly infectious Omicron variant. As ECDC reports (10): "Results from observational studies show that the vaccines authorised in the EU/EEA are currently highly protective against Covid-19 related severe disease, hospitalisation and death caused by the Delta variant of concern. Although overall high, effectiveness can vary, depending on the population groups (e.g., among elderly populations) and the vaccine.".

A literature study conducted by EMA and based on currently published studies suggests that the heterologous combination of mRNA and viral vector vaccines produces good levels of SARS-CoV-2 antibodies and a higher T-cell response compared to homologous vaccination, whether in a primary or booster regimen (11). However, the use of a viral vector vaccine as the second dose in primary vaccination schemes, or the use of two different mRNA vaccines, is less well studied. Therefore, additional real-world evidence is needed on the effectiveness of heterologous and booster vaccination on a larger population scale.

There are also still insufficient real-life data on the effectiveness of the vaccines authorised in the EU against the Omicron variant. Protection also appears to wane over time. Giving an additional or booster covid-19 vaccine dose following a full primary vaccination course is critical to ensure a higher level of protection, particularly in the face of emerging variants such as Omicron.

In Nordic countries, adenovirus platform-based vaccines are hardly used following safety concerns related to thrombotic events. Booster doses have been authorized by the EMA and MHRA and are being rolled out throughout Europe. Table 1 shows the information on the uptake of first vaccination, full regimens, and additional doses per total population and brand administered. The use of adenovirus platforms for primary regimens has been undertaken with heterogeneity among European countries, and these differences in exposure (see Table 1) will allow us to investigate differences in effectiveness based on primary schemes and booster.

The administrative death (linked to the eHR, for instance, from the information on the expiring reason of the Health Identity Card), including the death date, is available in all participating databases (Table 2), however the cause of death is missing in most of them. In those situations, algorithms (i.e. deaths with covid infection recorded in the previous 56 days) will be utilised instead to identify c ovid-19-related death for the primary analysis.

However, also information on the infection status of some patients may be missing in the electronic health records before death (12), due for instance to the periods of disrupted or overwhelmed health system, results of self-diagnosis test not recorded, or among patients not being attended in healthcare centres. Thus, secondary objective is proposed to complement the results of covid-19-related death in the primary objective, as a measure of the global health impact of boostering in the vulnerable population.

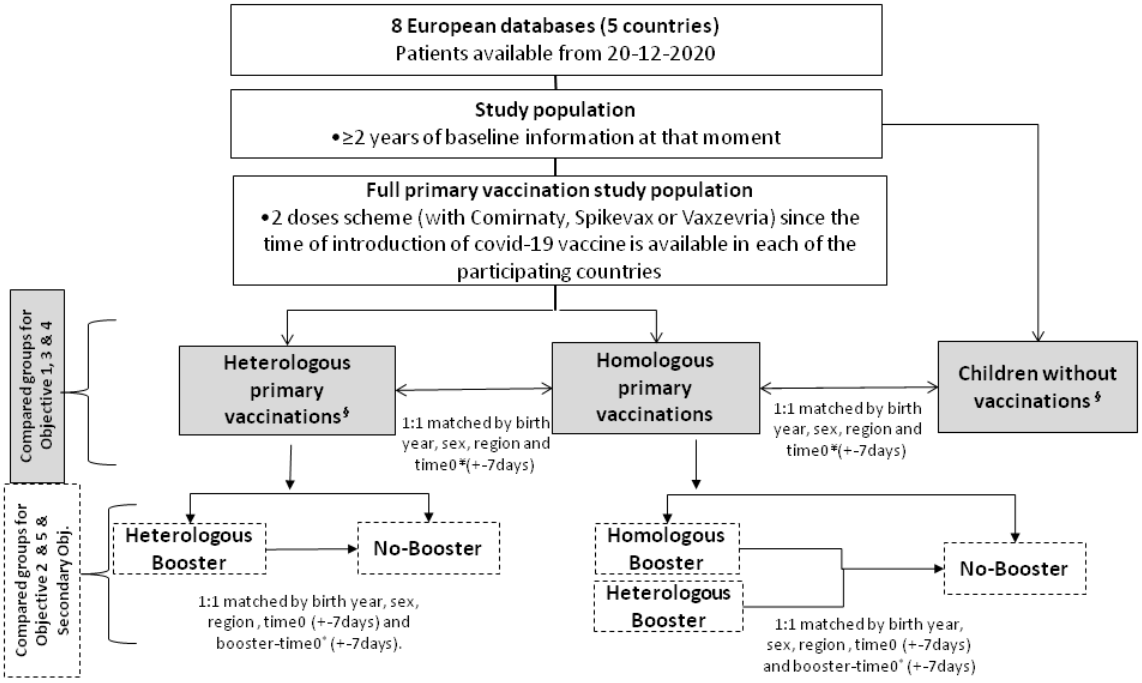
Table 1. *Information on the uptake of covid-19 vaccines in the participating countries and administered brands (as per ECDC Vaccine tracker 19-01-2022)*

Country	First doses	Full primary regimen	Additional doses
Spain	85.4%	75.6%	37.4%
Comirnaty	61,227,693		
Spikevax	18,931,042		
Vaxzevria	9,769,679		
Janssen	1,982,383		
Population size	47.4 million		
Italy	80.5%	74.3%	32.2%
Comirnaty	75,357,550		
Spikevax	21,190,469		
Vaxzevria	11,975,044		
Janssen	1,478,485		
Population size 2022	60.3 million		
Netherlands	76.7%	70.6%	44.5%
Comirnaty	22,646,011		
Spikevax	6,211,691		
Vaxzevria	2,800,808		
Janssen	867,708		
Population size 2022	17.2 million		
France	79.5%	75.6%	47.8%
Comirnaty	96,309,735		
Spikevax	22,098,333		
Vaxzevria	7,834,106		
Janssen	3,598,118		
Population size	67.4 million		
UK	76.2%	70.2%	53.5%
Comirnaty	NA		
Spikevax	NA		
Vaxzevria	NA		
Janssen	NA		
Population size	68.4 million		

7 Research Question and objectives

7.1 Research Question

The aim of this study is to assess the effectiveness and waning of immunity of primary Covid-19 vaccinations and the booster in preventing different covid-19 outcomes. Figure 1 provides an overview of the study design.



*Time0 is the time of 2nd vaccination for 2 dose regimens (homologous or heterologous);
*Booster-time0 is the time of 3rd vaccination for two dose regimens.
§Heterologous primary vaccinations recorded in children will be identified and analysed against no vaccination if numbers allow.

Effectiveness analysis will be stratified by children, adolescent and adults and patients with and without prior covid-19 infection

Figure 1. *Study flow chart.*

7.2 Primary Objectives

7.2.1. Primary objective 1 (adults and adolescents), 3 (children), and 4 (waning of immunity)

- To estimate the effectiveness and waning of effectiveness in adults and adolescents, separately, between heterologous and homologous primary vaccinations.
- To estimate the effectiveness and waning of effectiveness in children between homologous primary vaccinations and non-vaccination.

Up to March 2022, only Comirnaty vaccine was approved for children aged 5-11 years old. On 2 March 2022, the EU commission agreed to extend the marketing authorization of Moderna Covid-19 vaccine to be used in children aged 6 years and older. No heterologous schemes are expected for the time of the study in this population according to the latest updates of heatlh component authorities. Thus, the effectiveness of homologous vaccinations of those two vaccines will be estimated in comparison with non-vaccination among children. Therefore, the methodological approach for evaluating the effectiveness of Covid-19 vaccination in children will differ from the one used in adults and adolescents. If heterologous vaccination were found among children, matched non vaccinated children will be selected for comparisons.

7.2.2. Primary objective 2 (boosting), and exploratory objective 5 (waning of immunity after booster)

- To estimate the effectiveness and waning of effectiveness in adults with full homologous primary regimen between those with a homologous booster and heterologous booster, separately, compared to those without any booster.
- To estimate the effectiveness and waning of effectiveness in adults with full heterologous primary regimen between those with any booster and those without any booster.

On 24/02/2022, EMA recommended authorisation of booster doses of Comirnaty from 12 years of age. Therefore, even if these data are available in the participating databases before the end of the study, numbers are going to be too small to evaluate booster doses in adolescents.

Vaccine effectiveness will be estimated:

- By vaccine brand of the primary homologous scheme, and the combinations in the heterologous scheme
- By age category (children aged 5-11 years; adolescents aged 12-17 years; adults aged 18-29; 30-49; 50-69; 70-79; ≥80; separate groups of adults will be tested for heterogeneity).
- By time since a complete primary vaccination regimen or booster among the compared groups.
- Among clinically meaningful subgroups i.e. associated with a high risk of severe Covid-19 (patients with immunocompromise, cancer, transplants, renal replacement therapy, cystic fibrosis, and Down syndrome) among DAPs allowed to identify them (through diagnosis or medication).
- For patients free of previous Covid-19 infection (all analysis) and for patients with previous Covid-19 infection (analysis for severe Covid-19 and Covid-related death) that will be matched by time since previous covid-19 infection. i.e. <6months≥) .

7.3 Secondary Objective (effectiveness of booster against all-cause mortality)

To estimate the effectiveness against all-cause mortality in adults aged 60+ with a full primary regimen (whether homologous or heterologous) between those with any booster and those without any booster.

Analysis of all-cause mortality will be stratified by the following age groups (60-69; 70-79; ≥80). If a significant number of events will be encountered in age groups below 60 years old (<60), all-cause

mortality analyses will be performed to incorporate additional informative details for these populations in the study results.

8 Research Methods

To meet the proposed goal and objectives we will capitalize on the experience of the EU PE&PV and VAC4EU research network in the creation of readiness, governance, processes, data, people, methods, and tools developed in the ACCESS, Early-Covid-Vaccine-Monitor, and Covid-Vaccine-Monitor studies.

This study protocol describes the approach that will be employed by the tenderer to address the study objectives.

8.1 Study Organization

This study is organized in three work packages organized by specialty of expertise to ensure timely delivery of the data.

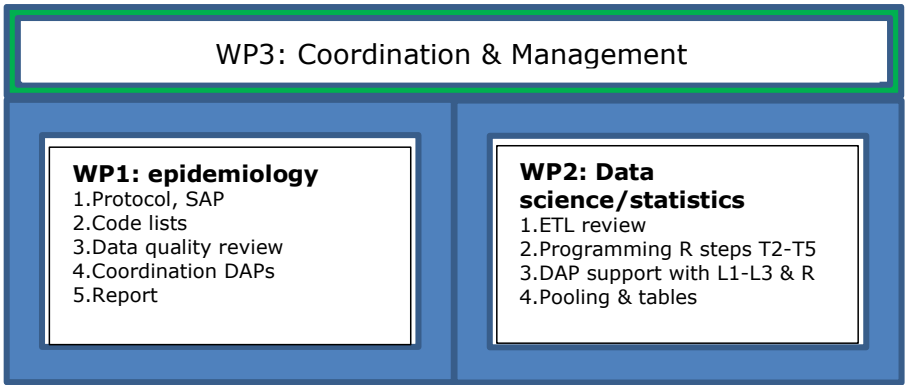


Figure 2. *Study organization scheme.*

WP1 is responsible for the protocol, statistical analysis plan, code lists for events, covariates, medicines, and the coordination of the data access providers in the entire process. WP1 will also draft the study report.

WP2 is responsible for the ETL process towards the Conception CDM, double programming of the T2-T5 steps in the pipeline & code review, and for the technical support of DAPs in running the scripts and debugging. WP2 will also code the creation of input and lay out of tables defined by WP1 as shell tables.

WP3 comprises of experts in coordination and management and will make sure delivery is on time.

EU PE&PV and VAC4EU data access providers (DAPs) have experience in using a distributed approach with a common protocol, common data model (ConcePTION), and common analytics. All DAPs will be prepared in having approvals for template protocols, extraction transformation load designs ready, and quality controlled, such that analysis can be done rapidly.

The governance structure for the project is based on a multi-level organisation that ensures:

- The fulfilment of the work plan.
- The due attention needed on critical activities that aim to ensure the achievement of milestones.

- and that contribute to strategic and scientific objectives.
- The relationships among partners, including conflict resolution.
- The quality and efficiency with which the project activities are carried out.
- The proper follow-up and fulfilment of the contract with EMA, including administrative and financial issues, and of any other legal arrangements with all parties.

To execute the workplan in time, we have distributed work across organizations with experience in leading tasks and accomplish complex projects.

Weekly calls will be organized with the WP leads, and organization of meetings will be created for the WPs. The co-leads and project managers will also have individual calls with each of the task leads to monitor progress and ensure that workstreams are efficient but also interacting.

8.2 Study Design

The study design is a retrospective multi-database cohort study that complements the published and ongoing test negative case control studies by other groups. We will leverage the ability to follow exposure cohorts and the size of the populations, which is important because some of the heterologous primary schemes are low in frequency. Moreover, in many data sources, negative tests are not recorded. A restriction to people with negative tests will be conducted as sensitivity analysis in CPRD and SIDIAP, where negative covid tests are available.

In the current protocol, we will use restriction and matching to ensure comparability and will adjust for additional confounders to ensure the removal of confounding.

Matching on vaccination date of the second dose (i.e. time0, for the primary vaccination analysis) and on the date of the booster (i.e. booster-time0, for the booster analysis) plus region will be crucial for controlling for confounding by calendar time and circulating strains as well as by vaccination prioritization (and consequently by risk of infection and prognosis). **Matching on birth year** will deal with the age-related roll-out of vaccination, risk of severe Covid-19 infection, and the vaccinations (date, brand, homo- or heterologous schemes, and booster date). For both analyses, matching will also be done for Covid-19 infection prior to time zero, according to the time since Covid-19 infection (< 6months; >=6 months).

Individuals are recommended to postpone the vaccination if they do not feel well or had a recent covid-infection, thus a healthy vaccinee effect may occur when comparing boosted patients with non-boosted people on the date the first one received the booster. Boosted and Non-boosted comparators are required to have no contact with the healthcare system in the week prior to time zero (the date the boosted pair received the booster) in main analysis. The strategy will be tailored according to the data sources possibilities (for instance by mean of excluding patients with visits to primary care, of a proxy outpatients' dispensations and other). This healthy vaccinee effect will not be present in objective 1 since patients are compared at the moment they received the 2nd dose (when heterologous versus homologous will be compared). Any other differential access to covid testing will be investigated in a sensitivity analysis restricting to patients with a negative test. Other measured confounders will be controlled for in the analysis through adjustment or stratification.

8.3 Data sources

The study will use data from secondary population-based electronic health care databases. All data sources will have the potential to provide high-quality data on Covid-19 vaccines (product types and dates), Covid-19 outcomes (test results, diagnoses in different settings, procedures, and treatments), and important covariates. Summary information about the data sources is provided in Table 2.

Table 2. *Included VAC4EU and EU PE&PV data sources characteristics*

Country	Data source	DAP & relation	HC Setting	Type of data	CDM	No. Active subjects (children, adolescents and adults)	Lag time	COVID-19 data	COVID-19 vaccine data (approximate N of vaccinations provided by DAPs on the 27/02/2022)	Death information (death date, cause of death)	Reference ENCePP data source Register (no. and link) and experience in COVID-19 studies
IT	CASERTA	Academic Spin-Off "INSPIRE srl"	Local health Unit	Hospital/emergency discharge, Immunization register, COVID-register, population file	CCDM	0.9 million	1-3 months	positive PCR tests from COVID register	COVID-19 vaccine administration date, brand and dose available. Homologous vacc. 1,744 aged 5-11; 22,791 aged 12-17 and 334120 adults. Heterologous: 212307 adults. Booster: Adolescent 12-17: 629; Adults: 27,406. until January 25th, 2022	Administrative death date	45094 Participation in CVM studies on COVID-19 disease and vaccines
IT	PEDIANET	SoSeTe	Pediatric primary care with linkage to the hospitalization database	Primary care medical records Immunization register Hospitalization database	CCDM	0.1 million (80.000 aged <14 years in the Veneto region)	2-4 weeks	positive and negative PCR tests from COVID registy	COVID-19 vaccine administration date and brand. Homologous vac. 88.606 aged 5-11 and 103.584 aged 12-14y	Death information (death date, cause of death) are always available when during hospitalization. If not, most likely to be reported by the FPs in the database.	20131 Participation in ACCESS & VAC4EU studies on COVID-19 disease and vaccines ORCHESTRA – EU funded project VERDI – EU funded project
ES	BIFAP	AEMPS	General practice	Primary care medical records linked with Hospital and ICU admissions and discharge	CCDM	4,237,801 aged <14 years 717,602 aged 14-17 years 12,727,542 aged 18-65 years 3,148,910 aged >65 years	3 months	COVID register, negative (in some regions)/positive tests COVID-related hospitalisations and ICU admissions	COVID-19 vaccine administration date, brand and dose available. 3,919,506 homologous and 8,933 heterologous vaccinated adults up-to October 2021. That needs to be updated for the current study.	Administrative death date	21501 Participation in ACCESS, ECVM, CVM studies on COVID-19 disease and vaccines EUPAS41134: Ongoing COVID-19 vaccines effectiveness study
ES	SIDIAP	IDIAP	General Practice	Primary care medical records Hospital discharge for part of population	CCDM	5.8 million	3 months	COVID tests negative/positive PCR, antigen tests	COVID-19 vaccine administration date, brand and dose available. 72,836 ,aged 5-11 and 570,795, aged 12-19. And 3,565,625 boosters among adults.	Administrative death date	4646 Participation in ACCESS, CVM studies on COVID-19 disease and vaccines

Country	Data source	DAP & relation	HC Setting	Type of data	CDM	No. Active subjects (children, adolescents and adults)	Lag time	COVID-19 data	COVID-19 vaccine data (approximate N of vaccinations provided by DAPs on the 27/02/2022)	Death information (death date, cause of death)	Reference ENCePP data source Register (no. and link) and experience in COVID-19 studies
UK	CPRD-Aurum	UU	Primary care, sample of national	Primary care medical records	CCDM	16 million	2-4 weeks	COVID tests negative/positive PCR, antigen tests	COVID-19 vaccine administration date, brand and dose available. Homologous vac. 49,671 children-adolescent and 4,935,754 adults. Heterologous: 136children/adolescents and 1,163,932 adults. Booster: 3,225,991 adults.	Death date recorded in GP record	https://cprd.com/primary-care Participation in ACCESS, ECVM, CVM studies on COVID-19 disease and vaccines
NL	PHARMO	PHARMO Institute	Primary care, sample of national	Primary care medical records	CCDM	2 million	1-6 months	COVID-19 positive PCR test	COVID-19 vaccine administration date, brand and dose available, with missingness in brand for GP data. Estimated 500,000 adults with booster.	Death date recorded in GP record	45066 Participation in ACCESS, ECVM, CVM studies on COVID-19 disease and vaccines
IT	ARS	ARS Toscana	Regional claims	Hospital/emergency discharge, Immunization register, COVID-register, population file	CCDM	3.7 million	3-4 months	COVID register, positive PCR tests	Administration date, brand and dose. Homologous vac: 102,885 aged 0-17 and 2331741 adults. Heterologous vac. 16,463 adults.	Administrative death date	24417 Participation in ACCESS, ECVM, CVM studies on COVID-19 disease and vaccines
FR	SNDS	APHP	Insurance data	Claims data	CCDM	60 million	3 months	COVID-tests for adults; COVID-19 hospitalization for all ages population	COVID-19 vaccine administration date, brand and dose available. Estimated 39 million people with booster.	Administrative death date	doi: 10.1002/pds.4233.

8.3.1. CPRD, United Kingdom

The CPRD collates the computerised medical records of a network of general practitioners (GPs) in the UK who act as the gatekeepers of health care and maintain patients' life-long electronic health records. The data are sourced from over 2,000 primary care practices and include 62 million patients, of whom 16.5 million are currently registered and active (13). General practitioners act as the first point of contact for any non-emergency health-related issue, which may then be managed within primary care and/or referred to secondary care, as necessary. Secondary care teams also provide feedback information to GPs about their patients, including key diagnoses. The data in the CPRD are updated monthly and include demographic information, prescription details, clinical events, preventive care, specialist referrals, hospital admissions, and major outcomes, including death (14,15). Most of the data are coded using Read or SNOMED codes. Data validation with original records (specialist letters) is available.

Depending on the type of electronic medical software used by the general practice, data are collected into either the CPRD GOLD (General Practitioner Online Database) or the CPRD Aurum database. The dataset is generalisable to the UK population based on age, sex, socioeconomic class, and national geographic coverage when the CPRD GOLD and the CPRD Aurum versions are used. Data include demographics, all GP/health care professional consultations, diagnoses and symptoms, results from laboratory tests, information about treatments (including prescriptions), data on referrals to other care providers, hospital discharge summaries (date and Read/SNOMED codes), hospital clinic summaries, preventive treatment and immunisations, and death (date and cause). Lag time for the CPRD GOLD and CPRD Aurum is 1 month. Information about vaccinations from mass vaccination campaigns during the pandemic is expected to transfer to GPs and into the patient's medical records (via National Health Service [NHS] systems rather than patients informing the GP); however, the lag time for this transfer varies. The present study will include only active CPRD practices (Aurum). The CPRD is listed under the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database, and access will be provided by the University Utrecht (UU).

CPRD approval would be required in order to approve the study protocol and it will require approximately 4-6 weeks.

For the current study, only primary care data in CPRD will be used. This has information on positive tests, as well as diagnoses of Covid-19 (a CPRD provided lists for diagnoses of Covid-19 will be used). Thus, CPRD cannot provide data for severe (hospitalised/ICU) Covid-19 analyses.

8.3.2. PHARMO, the Netherlands

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 the general practices, inpatient and outpatient pharmacies, clinical laboratories, hospitals, the cancer register, the pathology register, and the perinatal register—are linked at the 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 PHARMO Institute is always seeking new opportunities to link with additional databanks and is currently exploring linkage with the Covid-19 immunisation register, which is collected by the Dutch National Institute for Public Health and the Environment (RIVM).

Currently, the PHARMO Database Network covers over 6 million active persons of 17 million inhabitants of the Netherlands. Data collection period, catchment area, and overlap between data sources differ. Therefore, the final cohort size for any study will depend on the data sources included. All electronic patient records in the PHARMO Database Network include information on age, sex, socioeconomic status, and mortality. Other available information depends on the data source. The lag time of all databases is 1 year, except for the General Practitioner Database, which is updated every 3 months or less. For this study we will use the general practitioner database which 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. Primary care data are available for a portion of the population of approximately 3.2 million inhabitants (approximately 20% of the Dutch population). Information on lifestyle variables (e.g., BMI, smoking, alcohol consumption) is available in the General Practitioner Database if recorded by GPs in the electronic medical records.

The PHARMO Institute uses de-identified data from existing databases without any direct enrollment of subjects. Ethical approval or informed consent is not necessary according to the Dutch law regarding human medical scientific research (Wet Medisch-wetenschappelijk Onderzoek met mensen (WMO)), which is enforced by the Central Committee of Research involving Human Subjects (Centrale Commissie Mensgebonden Onderzoek (CCMO)). Studies performed on the PHARMO Database Network are reviewed afterwards by the PHARMO Compliance Committee to assess whether the WMO requirements are met

Covid-19 outcomes and algorithms:

Covid-19 outcomes are captured from data from the primary care level, using an algorithm with identifies: 1) Covid-19 episodes recorded by GP (either ICPC R83.03 or free text) and 2) Covid-19 tests, including test results. Date of infection is determined using multiple sources (date GP journal, mail, or Covid-19 test). Death date recorded in GP records.

8.3.3. ARS Toscana, Italyç

ARS Toscana is a research institute of the Tuscan regional government. Tuscany is an Italian region with approximately 3.6 million inhabitants. The ARS Toscana database comprises all the data that are collected in Tuscany related to health care delivered to those who are official residents of the region. Additionally, ARS Toscana collects data tables from regional initiatives.

The ARS Toscana database routinely collects primary care and secondary care prescriptions of drugs for outpatient use and can link them at the individual level with hospital admissions, admissions to emergency care, records of exemptions from co-payment, dispensing of diagnostic tests and procedures, causes of death, and a pathology registry, which has been available for the last few years and includes complete information only for morphology and topography. Occasionally, ARS Toscana may request retrieval of information from medical records or laboratory results regarding specific subpopulations and link this information to its core data.

Patients in ARS Toscana can be characterised in terms of age, sex, comorbidities (via algorithms), socioeconomic indicators, medication taken regularly on an outpatient basis, date of death, and health care utilisation (including visits to specialists, visits to ambulatory cancer care units, and visits to an emergency department or urgent care centre). Cause of death is available with a lag time of 3 years.

The lag time from a health care encounter to incorporation of data for research purposes is approximately 3 to 4 months. All patients in the ARS Toscana database can be linked to mortality data through deterministic linkage. There is no restriction on reporting small numbers. ARS data were part of the EMA-funded Early Covid-19 Vaccine Monitor study for Covid-19 vaccine safety monitoring.

Covid-19 outcomes and algorithms: the Covid-19 registry is the most reliable source to identify SARS-CoV-2 infection and severity of Covid-19 disease, including hospital admission, ICU admission, and death.

8.3.4. Caserta LHU database, Italy

The Caserta database is a claims database containing patient-level data from the city of Caserta, in the Campania region. The catchment 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.

Covid-19 outcomes and algorithms: the Covid-19 registry is the most reliable source to identify SARS-CoV-2 infection and severity of Covid-19 disease.

8.3.5. Pedianet pediatric data source, Italy

Pedianet, a pediatric general practice research database, was set up in 2000. It contains reason for accessing health care, 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 [International Classification of Diseases, Ninth Revision, Clinical Modification]), 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 approximately 140 family pediatricians distributed throughout Italy.

Pedianet can link to other databases using unique patient identifiers. In the first database, information on routine childhood vaccination is 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 diagnoses) is captured. The family pediatricians' participation in the database is voluntary, and individuals and their parents provide consent for use of their data for research purposes. In Italy, each child is assigned to a family pediatrician, who is the referral for any health visit or any drug prescription; thus, the database contains a 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, Italy, for validation. The Pedianet database can be linked to regional vaccination data, which was successfully tested in several large European projects (e.g., ADVANCE) where it was characterized and deemed fit for purpose to evaluate prescriptions including pediatric routine vaccines.

Timeframe for data availability: data are extracted every trimester.

Covid-19 outcomes and algorithms: Covid-19 positive test from the Covid-19 registry (we have data also for negatives and undetermined). Free text algorithms on Covid-19 signs and symptoms are being developed for specific outcomes. Hospitalization clinical data are available as free text, including ICU admission start and end date.

8.3.6. SNDS France

The SNDS (Système National des Données de Santé) (16) is the French nationwide healthcare database. It currently covers the overall French population (about 67 million persons) from birth (or immigration) to death (or emigration), even if a subject changes occupation or retires. Using a unique pseudonymized identifier, the SNDS merges all reimbursed outpatient claims from all French health care insurance schemes (SNIIRAM database), hospital-discharge summaries from French public and private hospitals (PMSI database), and the national death register. SNDS data are available since 2006 and contains information on:

- General characteristics: gender, year of birth, area of residence, etc.
- Death: month, year, and cause.
- Long-Term Disease registration associated with an ICD-10 diagnostic codes.
- Outpatient reimbursed healthcare expenditures with dates and codes (but not the medical indication nor result): visits, medical procedures, nursing acts, physiotherapy, lab tests, dispensed drugs, and medical devices, etc. For each expenditure, associated costs, prescriber, and caregiver information (specialty, private/public practice) and the corresponding dates are provided.
- Inpatients details: primary, associated ICD-10 diagnostic codes resulting from hospital discharge summaries with the date and duration of the hospital stay, the performed medical procedures, and the related costs. Drugs included in the diagnosis related group cost are not captured.

This study will be conducted through the permanent access of the APHP to the SNDS data. This allows us to avoid the national data protection office (CNIL) processes, and saving time. The SNIIRAM data were not yet characterized in the ADVANCE project but have been used for vaccine studies (http://www.encepp.eu/encepp/viewResource.htm?id=38744).

Covid-19 infection can be identified:

- using tests (PCR, antigens and antibodies). This is available only for adult patients.
- through ICD10 diagnosis codes for hospitalizations for Covid-19.
- 3 types of diagnosis are available: Covid-19 as a main, related or associated hospitalization diagnosis.

8.3.7. SIDIAP, Spain

The Information System for Research in Primary Care (Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària [SIDIAP]) in Catalonia, Spain, is a primary care database set up by the Institute of Research in Primary Care (Institut Universitari D'Investigació en Atenció Primària Jordi Gol [IDIAP Jordi Gol]) and Catalan Institute of Health (Institut Català de la Salut). The database collects information from 278 primary health care centres and includes more than 5.8 million patients covered by the Catalan Institute of Health (approximately 78% of the Catalan population) and is highly representative of the Catalan population (17).

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

Currently, because of the Covid-19 pandemic, having shorter-term updates to monitor the evolution of the pandemic is a possibility. Recent reports have shown SIDIAP data to be useful for epidemiological research. SIDIAP is listed under the ENCePP resources database (http://www.encepp.eu/encepp/resourcesDatabase.jsp). SIDIAP was characterised in the IMI-ADVANCE project and considered fit for purpose for vaccine coverage, benefits, and risk assessment (http://www.encepp.eu/encepp/viewResource.htm?id=4646).

After EMA approval, the protocol must be evaluated by the SIDIAP Scientific Committe and by the IDIAPJGol Ethics Committee, the approval can take 4-6 weeks. The timeframe for data availability after the approval by the two local Committees is one month.

Covid-19 outcomes and algorithms

We have two ways to get a COVID+ in SIDIAP data:

- Through diagnosis codes ICD10CM B34.2, B97.21 and U07.1;
- and using tests (PCR, antigens and antibodies).

Covid-19 admitted to hospital or ICU will be able to identify. We also have registered the negatives and the undetermined tests.

8.3.8. BIFAP, Spain

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

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

Information up to the end of 2021 and Covid-19 is also available for several regions from registries linked to the database. The BIFAP database was characterized in the ADVANCE project and considered fit for purpose for vaccine coverage, benefits, and risk assessment (http://www.encepp.eu/encepp/viewResource.htm?id=21501) and participated in the EMA-funded Early Covid-19 Vaccine Monitor study for COVID-19 vaccine safety monitoring.

We estimate that the timeframe for data availability since the EMA approval is less than a month, since the approval from BIFAP Scientific Committee is required, which has a meeting every month, and from an Ethics Committee, which meets twice a month.

Covid-19 outcomes and algorithms

Covid-19 infection can be identified:
- using tests (PCR, antigens and antibodies), which is the most reliable source to identify SARS-CoV-2 infection
- through ICPC-2 and ICD-9 diagnosis codes mapped to SCTSPA (SNOMED). Validations parameters using positive test as gold-standard have been estimated in a previous study, which will allow to select the most predictive algorithm for risk estimation.

The cases of Covid-19 related hospital and ICU admissions will be identified as follows:

- Covid-19 was recorded as the cause of hospital or ICU admissions (information provided by some regions participating in the database);
- or a hospital or ICU admission with Covid-19 diagnosis recorded in the 30 days after a covid-19 positive test (if the previous is not provided by the region).

8.4 Study Population

The source population comprises all persons registered in any of the data sources during the study period (December 2020-latest update). The full**-Covid-19 vaccinated** study population will include all individuals with:

- At least two recorded vaccinations since start of study period in each of the participating countries and
- At least 2 years of baseline information at that moment ensuring that information on covariates is available
    - A person is classified to have a full primary vaccination regimen when a record of a second covid-19 vaccine for persons with a first dose with Comirnaty, Spikevax or Vaxzevria (two doses primary vaccination schemes) is administered >19 days after the first dose (since it will be difficult to distinguish between potential data entry errors, this is, two vaccination records too closed or 3 doses recorded, which may indicate double recording of the same vaccination, or failed/weak effect.)

8.4.1. Matched Populations

Within the all-vaccinated study population, matching will be conducted to answer primary and secondary objectives.

8.4.2. Matched population for the effectiveness of primary vaccination

From the all-Covid-19 vaccinated study population, individuals with a homologous full primary vaccination will be matched 1:1 to persons with A) a full heterologous primary vaccination if they are adolescents or adults, or, B) no vaccination if they are children, based on year of birth (to fall into the ranges 5-11 ; 12-17 ; 18-29; 30-49; 50-69; 70-79; ≥80), sex, region, previous covid-19 infection and time since previous covid-19 infection, calendar date of time zero (+/7 days) and 1st dose and brand of the 1st dose (in heterologous versus homologous analysis). Additional matching variables will be considered if the programming activities are implemented in time by the consortium to include the results in the final study report.

8.4.3. Matched populations for the effectiveness of booster

1. Among individuals with a homologous primary vaccination:

1.1. those with homologous booster will be matched to persons without booster (1:1) based on year of birth (to fall into the ranges 5-11 ; 12-17 ; 18-29; 30-49; 50-69; 70-79; ≥80), sex, region, Covid-19 prior to booster-time0 (yes/no and time since Covid-19, i.e. < 6months; >=6 months), calendar time0 (+/- 7 days) and booster-time0 (+/- 7 days).

1.2. those with heterologous booster will be matched to persons with no booster (1:1) based on year of birth (to fall into the ranges 5-11 ; 12-17 ; 18-29; 30-49; 50-69; 70-79; ≥80), sex, region,

Covid-19 prior to booster-time0 (yes/no and time since Covid-19, i.e. < 6months; >=6 months) and calendar time0 (+/- 7 days), booster-time0 (+/- 7 days) and 1st dose.

1.3. Among individuals with a heterologous primary vaccination, those with booster will be matched to persons without booster (1:1) based on year of birth (to fall into the ranges 5-11 ; 12-17 ; 18-29; 30-49; 50-69; 70-79; ≥80), sex, region, Covid-19 prior to booster-time0 (yes/no and time since covid, i.e. < 6months; >=6 months) and calendar time0 (+/- 7 days) , booster-time0 (+/- 7 days) and 1st dose.

1.4. Boosted and non-boosted with prior Covid-19 infection (any) will be additionally matched by time since previous Covid-19 infection on calendar date of time zero (2nd dose).

Boosted and Non-boosted comparators will be required to have no contact with the healthcare system (proxy for healthy patient in that moment) in the week prior to booster-time0 (as children prior to time0). In sensitivity analysis, stratified by data source setting (i.e. primary care and secondary/hospital-based) will inform about the residual confounding and limitation to control by healthy patients in those last data sources.

8.4.4. Follow-up

Persons without a matching pair will be excluded from the analysis for that objective.

For primary objectives, follow-up will start at time0 (or booster time0) and continue until the earliest of the following dates: Covid-19 disease/infection, death, last date of data extraction, or moving out of the data source. For primary objectives 1 and 4, follow-up will be additionally censored at the date of booster vaccine administration in the non-boosted cohort or any extra dose recorded in the booster cohort. At this censoring date, follow-up for the matched boosted persons will also be censored.

For children analysis, follow-up will be additionally censored at the date of vaccine administration in the non-vaccinated cohort of children. At this censoring date, follow-up for the matched vaccinated children will also be censored.

For secondary objective, follow-up will start at booster time zero and continue until the earliest of the following dates: death, last date of data extraction, or moving out of the data source.

8.5 Variables

Available variables vary by the data sources. Only data sources with variables adequate for addressing a given objective will be included in the analysis of that objective.

8.5.1. Definition of Time Zero (time0)

Aligning the evaluation of eligibility criteria, covariate assessment, exposure assignment, and beginning of follow-up (time zero) avoids selection bias and immortal person-time bias. Time zero is when the vaccination status is assigned; all eligibility criteria must be fulfilled, and Covid-19 infection-related outcomes must start to be followed.

For objectives 1, 3 and 4, time0 is the date when the 2nd dose of Covid-19 vaccine is administered (i.e. recorded). That date will be used to match homologous to heterologous pairs or to non-vaccinated children.

For objectives 2 and 5 and secondary objective, the date when the booster vaccination is recorded (among patients with a booster dose) will be defined as time zero (booster-time0). The booster-time0 will be assigned also as time0 for the non-boosted matched person.

8.5.2. Exposure information

Vaccination information will be based on recorded prescription, dispensing, or administration of the Covid-19 vaccines. Vaccine receipt and date of vaccination will be obtained from all sources that can capture Covid-19 vaccination, such as pharmacy dispensing records, general practice records, immunization registers, vaccination records, medical records, or other secondary data sources. The main exposure of interest is the receipt of a primary regimen or booster Covid-19 vaccine, the dose, and its brand/manufacturer.

8.5.2.1 Exposure assignment

Exposure assignment is based on the brands used in the primary Covid-19 vaccination scheme and the booster vaccines. Table 3 below describes the different options.

Table 3. *Classification of heterologous and homologous primary schemes and heterologous /homologous booster.*

A full primary homologous regimen means:	**Booster classifications**		
	Non-boosted	Homologous	Heterologous
2 doses of Pfizer vaccine as primary regimen	no booster	Pfizer	Non-Pfizer
2 doses of Moderna vaccine as primary regimen	no booster	Moderna	Non-Moderna
2 doses of AstraZeneca vaccine as primary regimen	no booster	AstraZeneca*	Non-AZ
A full primary heterologous regimen means:			
1st dose Pfizer, 2nd dose Moderna or AstraZeneca	no booster		any
1st dose Moderna, 2nd dose Pfizer or AstraZeneca	no booster		any
1st dose AstraZeneca, 2nd dose Pfizer or Moderna	no booster		any

*these may occur infrequently

8.5.2.2 Exposure assessment

8.5.2.2.1 Primary objective 1, 3, and 4: effectiveness of heterologous versus homologous primary vaccinations (or homologous vaccination versus no-vaccination among children)

Individuals will be assigned to the primary heterologous Covid-19 vaccination group if they receive a different Covid-19 vaccine brand for the 1st and 2nd doses of a two-dose primary (initial) course (see table 3 for the different options).

Heterologous primary vaccination is defined as the receipt of a different Covid-19 vaccine for the 1st and 2nd doses of a two-dose primary (initial) course.

Homologous primary vaccination is defined as the receipt of the same Covid-19 vaccine brand for the 1st and 2nd of a two-dose primary course.

Among children, non-vaccination is defined as no receipt of a Covid-19 vaccine at the time0.

A person is classified to be a boosted individual when any Covid-19 dose is administered at least 28 days after the 2nd dose (AstraZeneca, Pfizer, or Moderna). Unboostered individuals will be selected among those with at least 28 days of follow-up and with no third dose in that period. Until the time a third dose is received, the person is considered non-boosted.

8.5.2.2.2 Primary objective 2 and exploratory objective 5: effectiveness of booster versus non-booster vaccinations

Persons receiving a dose >=28 days after the 2nd dose (AstraZeneca, Pfizer, or Moderna) will be assigned as heterologous or homologous boosting, until the time a third dose is received, a person is considered non-boosted.

8.5.2.2.3 Induction time and effectiveness over time as a proxy of the waning of immunity

Effectiveness of heterologous versus homologous vaccinations or booster versus no-booster will be assessed over time in the first 0-6, 7-13, 14-30 days followed by monthly post (booster-)time0 intervals.

Among children, effectiveness will be assessed for the main period of interest, i.e., after 2nd dose and over time.

The date, dose, and type of vaccine administrated will be collected as reported in the previous section for each data source.

8.5.3. Covid-19 outcomes

This study will consider different Covid-19 outcomes: severe Covid-19, Covid-19-related death and all Covid-19 infections.

8.5.3.1 Severe Covid-19 disease

A person will be considered to have severe covid-19 disease when:

- Covid-19 was recorded as the cause of hospital or ICU admissions (if available in the data sources; see Data Sources Section)
- or, the hospital or ICU admission occurred within 30 days of a covid-19 disease or positive test (If reason for admission is not available in a particular database; see Data Sources Section).

8.5.3.2 Covid-19-related death

A person will be considered to have Covid-19-related death when:

- Covid-19 was recorded as the cause of death (if available in the data sources; see Data sources section).
- Or, the death occurred within 8 weeks of a Covid-19 disease or positive test (if the cause of death is not available in the database; see Data Sources section).

8.5.3.3 All Covid-19 infections

Covid-19 infection is defined as a positive test (PCR or antigens) or a Covid-19 diagnosis (depending on the database algorithm) regardless the prognosis.

8.5.3.4 Covid-19 Information by data sources

Information on availability of tests and/or diagnosis for Covid-19 is listed in Table 2.

8.5.4. All cause mortality

Deaths of any cause will be included that is available in all data sources.

8.5.5. Covariates

To control for measurable confounders in the analysis, the following factors will be considered: lifestyle characteristics (BMI, smoking alcohol abuse), comorbidities, comedications, and health care utilization prior to or at time zero or booster-time zero. The listed covariates are available in each database.

The following variables will all be assessed as last status before time zero where available:

- Smoking status
- BMI
- alcohol abuse (no for SNDS).

The following comorbidities (that may be shown to be associated with Covid-19 prognosis) will be assessed ever before time zero and booster time-zero:

- Diabetes mellitus (types 1 and 2)
- Hypertension
- Coronary artery disease
- Cerebrovascular disease
- Chronic respiratory disease

- Chronic kidney disease
- Chronic liver disease
- Cancer
- Immunodeficiencies (including Human immunodeficiency virus, Sickle cell, and other immunosuppressing conditions)
- Autoimmune disorders
- Cystic fibrosis
- Down Syndrome
- Parkinson disease
- Dementia, sepsis
- Heart failure
- Bladder incontinence
- Arthritis
- Coagulation deficiencies.

Comedication use will be assessed as proxies for comorbidities. The following comedications will be assessed during the year before time zero and booster-time zero:

- Antibiotics
- Antiviral
- Corticosteroids
- Non-steroidal anti-inflammatory drugs
- Other analgesic
- Psychotropics
- Statins
- Immunosuppressant
- Influenza vaccine.
- >5 drugs (as a proxy of high level of morbidity)

Antibiotic and antiviral prescriptions will also be assessed in the month prior to time0 and booster-time0 as a marker of acute illness.

Health care utilization in the year before time0 and booster time0 will be evaluated as a proxy measure of health care–seeking behavior, overall health status, unmeasured confounders and access to health care. Additionally, short-term health care utilization in the week before 1st and 2nd dose (for children) and booster-time0, will be extracted separately, as short-term markers of current health status that may influence individuals' vaccination decisions (to minimize healthy vaccine effect). Considered variables will include the following:

- Health care utilization (number of visits to PC or claims)
- Influenza vaccination (number in the previous 5 years)
- Other non-childhood vaccinations (number in the previous 5 years)
- Covid-19 tests (total number, including positive and negative test if available in the data source)

Information on institutionalization or residency in a care home and on being a high-risk professional will be explored in the data sources and collected for confusion controlling if available.

8.5.5.1 Genetic variant of SARS-Cov2 virus

The presence of dominant genetic variants of Sars-Cov-2 at time zero and time zero-booster will be defined based on the periods of dominant circulation obtained from surveillance data in the respective countries (according to SARS-CoV-2 variants dashboard in the European Centre for Disease Prevention and Control). For instance, in Spain, beta and alpha variants were dominant from 1st January-30th June 2021, delta (1st July-30th November 2021), and omicron (from 1st December 2021).

8.6 Study size

All patients meeting the eligibility criteria in the sources of data will be included in the source population, which will comprise more than 98 million patients (Table 2). Out of the data sources able to provide data at proposal deadline (Table 4) a minimum of 7 million patients aged ≤19 years will be included in the study. However, we expect many greater number of children and adolescent contributing due the contribution of SNDS France which includes data on around 7 million people and 29 millions of person-years (of which 24% will be patients aged <19 years) with an approximate distribution of age as follows . The distribution of population by age is estimated in Table 4.

Table 4. *Aggregated age distribution of databases population, excluding SNDS data, calculated based on previous C-19 projects*

Age category	Overall	% overall population
0-19 y	7054129	21.0
20-29 y	3716304	11.1
30-39 y	4569059	13.6
40-49 y	5052251	15.0
50-59 y	4599200	13.7
60-69 y	3623643	10.8
70-79 y	2885769	8.6
80+ y	2094119	6.2
Total	**33594474**	

In Table 5, we provide the estimation of the probability that the upper limit of the 95% CI of the risk ratio (RR) is below 1.00 (a correlate of the lower bound of the vaccine effectiveness estimate being above 0.00 demonstrating a protective effect of vaccination) according to:

- the population in the data sources participating in each analysis: around 67 millions of fully vaccinated patients for Covid-19 infections analysis and 59 millions for severe covid-19 analysis (excluding CPRD and Pharmo without link to hospital data) according to the ROC20 interim reports (https://www.encepp.eu/encepp/viewResource.htm?id=42637) and SNDS data.

- the incidences in two different periods:

1. at the beginning of booster recommendations, i.e. November 2021 with low incidence of infection (21/100,000 hab.) but high hospitalization proportion (4.3%; incidence: 0.9 hospitalised Covid-19 per 100,000 habitants), and

2. at the beginning of 2022 with higher incidences of infections (716/100,000 hab.) but lower hospitalisation proportion (1.4%; 10 hospitalised Covid-19 per 100,000 habitants)

Under different VE scenarios (I.e. 25%, 51.3%, 75% or 95%), selecting 1 vaccine non-boosted to each boosted patient (ratio of 1:1) and a low incidence of 21 infections per 100,000 habitants in non-boosted patients, the probability would be always 1 as estimated and displayed in the following Table 5 (overall and by brand).

Table 5. *Study Size Precision Estimates by boosted versus non-boosted patients over all data sources (including SNDS)* ***for Covid-19 infection*** *in adults.*

Type of vaccine	Expected Vaccine Effectiveness [a, b]	Expected RR (boosted versus non-boosted)	Expected Ratio of Unexposed (non-boosted) / Exposed (boosted)	Expected Sample Size (boosted + non-boosted=pob complete vaccinated)	Expected Risk of the Outcome in the non-boosted (cases per one person)[b,c]	Probability of the Upper Limit of the 95% CI to Be Below 1.00
Overall	51.3% [b]	0.487	1:1	67,271,319	0.000212 [c]	1
Overall	25%	0.75	1:1	67,271,319	0.000212 [c]	1
Overall	75%	0.25	1:1	67,271,319	0.000212 [c]	1
Overall	95%	0.05	1:1	67,271,319	0.000212 [c]	1
Overall	51.3% [b]	0.487	1:1	67,271,319	0.00716 [b]	1
Overall	25%	0.75	1:1	67,271,319	0.00716 [b]	1
Overall	75%	0.25	1:1	67,271,319	0.00716 [b]	1
Overall	95%	0.05	1:1	67,271,319	0.00716 [b]	1
By type of the primary vaccination:						
Pfizer/Biontech (BNT162b2) (Comirnaty)	49.7%	0.503	1:1	50,173,453	0.00716[b]	1
Moderna (mRNA-1273) (Spikevax)	55.3%	0.447	1:1	7,387,961	0.00788[b]	1
Oxford/Astrazeneca (ChAdOx1 nCoV-19) (Vaxzevria)	58.6%	0.414	1:1	9,163,478	0.00578[b]	1

CI = confidence interval; RR = relative risk. Source: Rothman, K. Episheet: spreadsheets for the analysis of epidemiologic data. 2015. Available at: http://www.krothman.org/episheet.xls. Accessed January 2022. [a] Vaccine effectiveness measured as 1 minus the RR, where the RR compared the risk of the outcome in vaccine-exposed versus that in unexposed individuals (i.e., an RR below 1 indicates a protective effect of the vaccine, corresponding to positive vaccine effectiveness). [b] VE and Incidence among non booster during 3-January and 6-february 2022 according to Monge, Susana and Rojas-Benedicto, Ayelén and Olmedo, Carmen and Mazagatos, Clara and Sierra, María José and Limia, Aurora and Martín-Merino, Elisa and Larrauri, Amparo and Hernán, Miguel A., The Effectiveness of mRNA Vaccine Boosters for Laboratory-Confirmed COVID-19 During a Period of Predominance of the Omicron Variant of SARS-CoV-2. Available at SSRN: https://ssrn.com/abstract=4035396 . [c] Based on cumulative incidence of 21.2 cases of covid-19 infection per 100,000 habitants in Spain in the 43rd week of year 2021 (at the beginning of the booster recommendation. Ref. Informe nº 103 Situación de COVID-19 en España a 3 de noviembre de 2021.pdf (isciii.es)

Based on European/national policies and previous studies of the same data sources (ROC19 EMA-2018-28-PE_ROC19_ECVM_Cohort Monitoring of Adverse Events), the proportion of patients receiving a heterologous second dose among those fully vaccinated (initial course) ranged by the source of data from 0.0-3.8% for AZ, 0.0-0.47% for Pfizer, and 0.0-0.16% for Moderna at the beginning of the campaign.

The numbers of heterologous primary vaccinations are expected to be much lower than that of the homologous ones; however, we expect enough power based on the large population covered by data sources, and the high percentage of full primary regimens (see Table 1) in the participating countries. Therefore, we will be able to match by several factors.

Excluding the SNDS population, the expected sample would be around 19 million fully vaccinated individuals. If we hypothesize a very low theoretical 25% effectiveness value of booster vs non-booster against Covid-19 infection and considering a 21.2/100,000 habitants risk value: only in the case of the study population would suffer a strong reduction equal to 1/10 of the total expected (reduced till 1 million in each compared matched cohort as displayed in the Table 6), the probability will be lower than 0.97 (i.e. 0.77). However, even in this unlikely condition, the sample size will always be reached.

Table 6. *Study Size Precision Estimates assuming an effectiveness of 25% of the booster against covid-19 infection in comparison with non-boosted patients, under an incidence of 21.2 covid-19 infections per 100,000 non-boosted patients, allocation ratio of 1, and maximum sample size of 19,287,049 adults over all data sources (excluding SNDS).*

Size	Number in boosted	Number in non-boosted	SE (ln(RR))	Span	Probability of the Upper Limit of the 95% CI to Be Below 1.00
1928704.9	964352.5	964352.45	0.106823	-0.20937	**0.768257**
3857409.8	1928705	1928704.9	0.075535	-0.14805	**0.967744**
5786114.7	2893057	2893057.35	0.061674	-0.12088	**0.996581**
7714819.6	3857410	3857409.8	0.053411	-0.10468	**0.999694**
9643524.5	4821762	4821762.25	0.047773	-0.09363	**0.999976**
11572229.4	5786115	5786114.7	0.04361	-0.08547	**0.999998**
13500934.3	6750467	6750467.15	0.040375	-0.07913	**1.000000**
15429639.2	7714820	7714819.6	0.037767	-0.07402	**1.000000**
17358344.1	8679172	8679172.05	0.035608	-0.06979	**1.000000**
19287049	9643525	9643524.5	0.03378	-0.06621	**1.000000**

Using simulated calculations under difficult and uncommon conditions, the severe Covid-19 analysis could undergo a reduction in the probability to 0.87 in case the VE was 25% under an incidence of hospitalised Covid-19 of 0.9/100,000, as showed in Table 7. Only in the case that we consider the worst possible scenario, which consists of the exclusion of SNDS data, the sample size will reduce to around 11 million in the same VE condition of fully vaccinated patients, thus, the probability to 0.28.

Table 7. *Study Size Precision Estimates by boosted versus non-boosted patients against severe Covid-19 in adults over data sources with hospital data (including CASERTA, BIFAP, SIDIAP, ARS y SNDS and excluding CPRD and PHARMO)* ***for severe Covid-19****.*

Type of vaccine	Expected Vaccine Effectiveness (1-RR)	Expected RR (boosted versus non-boostered)	Expected Ratio of Unexposed (non-boosted) / Exposed (boosted)	Expected Sample Size (boosted + non-boosted=pob complete vaccinated)	Expected Risk of the Outcome in the non-boosted (cases per one person)[b,c]	Probability of the Upper Limit of the 95% CI to Be Below 1.00
Overall	25%	0.75	1:1	59,001,387	0.000009116[c]	0.870448
Overall	50%	0.50	1:1	59,001,387	0.000009116[c]	0.999998
Overall	75%	0.25	1:1	59,001,387	0.000009116[c]	1
Overall	95%	0.05	1:1	59,001,387	0.000009116[c]	1
Overall	25%	0.75	1:1	59,001,387	0.00010024[b]	1
Overall	50%	0.50	1:1	59,001,387	0.00010024[b]	1
Overall	75%	0.25	1:1	59,001,387	0.00010024[b]	1
Overall	95%	0.05	1:1	59,001,387	0.00010024[b]	1

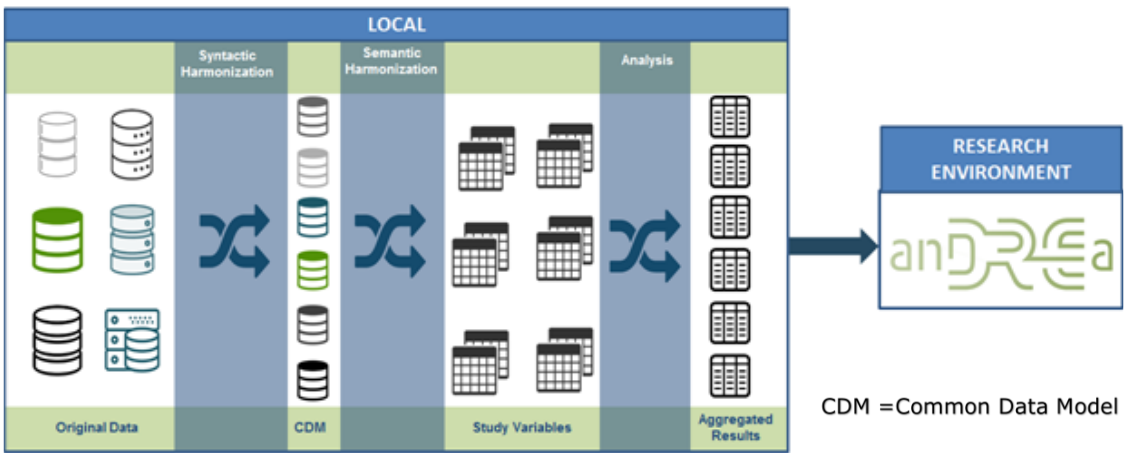
CI = confidence interval; RR = relative risk. Source: Rothman, K. Episheet: spreadsheets for the analysis of epidemiologic data. 2015. Available at: http://www.krothman.org/episheet.xls. Accessed January 2022. [b] Based on a cumulative incidence of 716 infections per 100,000 habitants among non-booster during 3-January 6-February 2022 reported by Monge S et al. The Effectiveness of mRNA Vaccine Boosters for Laboratory-Confirmed Covid-19 During a Period of Predominance of the Omicron Variant of SARS-CoV-2. Available at SSRN: https://ssrn.com/abstract=4035396; and a 1.4% of hospitalisation (Ref. Informe nº 117 Situación de COVID-19 en España a 09 de febrero de 2022.pdf (isciii.es)). [c]Based on cumulative incidence of 21.2 cases of covid-19 infection per 100,000 habitants in Spain in the 43rd week of year 2021 (at the beginning of the booster recommendation in Spain in this period the 4.3% of infection were hospitalized). Ref. Informe nº 103 Situación de COVID-19 en España a 3 de noviembre de 2021.pdf (isciii.es)

8.7 Data processing

This study will be conducted in a distributed manner using a common protocol, common data model (CDM), and common analytics programs based on existing health data. The following steps will be implemented:

- Extraction, transformation, and loading (ETL) of data to a CDM. To harmonise the structure of the data sets stored and maintained by each data partner, a shared syntactic foundation will be used. The CDM is the Conception CDM v2.2 for EHR. In this CDM, data are represented in a common structure, but the content of the data remain in their original format.
- The ETL design is shared in a searchable FAIR catalogue. The VAC4EU FAIR Molgenis data catalogue is a meta-data management tool designed to contain searchable meta-data describing organisations that can provide access to specific data sources.
- To reconcile differences across terminologies, a shared semantic foundation is built for the definition of events under study by collecting relevant concepts in a structured fashion using a standardised event definition template. The Codemapper tool will be used to create diagnosis code lists based on completed event definition templates for each outcome and comorbid risk condition . Based on the relevant diagnostic medical codes and keywords, as well as other relevant concepts (e.g., medications), one or more algorithms are constructed (typically one sensitive, or broad, algorithm and one specific, or narrow, algorithm) to operationalise the identification and measurement of each event. These algorithms may differ by database, as the components involved in the study variables may differ. Specifications for both ETL and semantic harmonisation will be shared in the catalogue.
- Third, following conversion to harmonised study variable sets, R programs for the calculation of incidence and risk will be distributed to data access providers for local deployment.
- The aggregated results produced by these scripts will then be uploaded to the Digital Research Environment (DRE) for pooled analysis and visualisation (see Figure 3). The DRE is made available through UMCU (University Medical Center Utrecht) (https://www.andrea-consortium.org/). The DRE is a cloud-based, globally available research environment where data are stored and organised securely and where researchers can collaborate (https://www.andrea-consortium.org/azure-dre/).

Figure 3. *Data management flow.*



8.7.1. Quality management and control

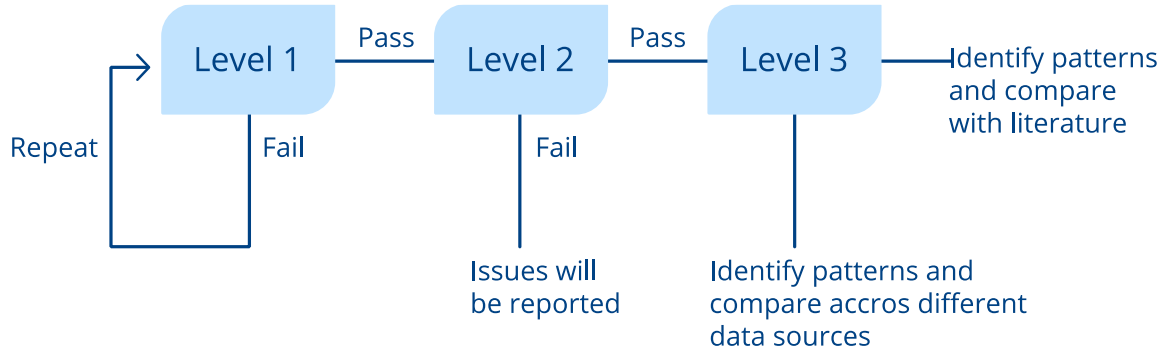
Data transformation into the CDM will be conducted by each subcontracted research partner in its associated database, using the processes described in the following sections (see below), each of these steps is fully transparent and will be signed of/reviewed.

Standard operating procedures or internal process guidance at each research centre will be used to guide the conduct of the study. These procedures include rules for secure and confidential data storage, backup, and recovery; methods to maintain and archive project documents; Quality control procedures for programming; standards for writing analysis plans; and requirements for scientific review by senior staff.

8.7.2. Quality checks of DAPs data

Data partners will be asked to provide data quality checks (Figure 4) for level 1 (Completeness) and 2 (consistency) (or any EMA quality framework checks when available); level 3 is checking for study variables and assess whether data are fit for purpose.

Figure 4. *Data quality check cycle as developed in ConcePTION Quality Framework.*



Generic open-source data quality check scripts are available from the IMI-ConcePTION quality framework that are publicly available on GitHub.

8.7.2.1 Level 1 - Data completeness

The purpose of the level 1 check is to verify the completeness of the ETL process and the data in the variables. Examples of tests are:

- Presence of variables in each of the CDM tables in D2
- Checks for misspelling and letter case in variable names in each of the CDM tables
- Verification of vocabularies
- Check date formats

- Check conventions of values
- Missing data analysis
- Frequency tables for categorical variables

https://github.com/IMI-ConcePTION/Level-1-checks

8.7.2.2 Level 2 - Data logic/consistency

Real data are not random but follow certain logical constraints that reflect rules governing real-world situations. Examples of indicators generated by level 2 checks are:

- Event dates before date of birth
- Event dates after date of death
- Event dates out of observation periods
- Subjects having an observation but not present in the PERSONS table
- Observations associated with a visit id and occurred before/after the visit start/end date
- Subjects younger than 12 years old reported as parents
- Age at the observation period older than 115 y old

https://github.com/IMI-ConcePTION/Level-2-checks

8.7.2.3 Level 3 – Fit for purpose

Level 3 checks review patterns of study variables over time, age within and between datasources. There are 8 modules, which may be used depending on the study variables:

- Source and study population.
- Medicines
- Vaccines
- Diagnoses
- Pregnancy
- Populations of interest.
- Health-seeking behaviour and lifestyle factors.
- EUROCAT indicators.

https://github.com/IMI-ConcePTION/Level-3-checks

8.7.3. Quality checks of R-coding

Data Management and Statistical Analysis will follow standard operating procedures for UMCU. All Statistical Analysis programs will be double coded or reviewed by UMCU and ARS.

UMCU will create clear documentation (graphical and in Excel spreadsheet) of the data management steps, beginning with describing the required variables from the CDM and each of the subsequent transformation steps and intermittent data tables. ARS will double code or conduct code review of

the datasets built in R by UMCU using R and from instructions provided by UMCU. Discrepancies will be resolved.

8.7.3.1 Coding conventions (process quality)

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

The main coordinator of the Github is the UMCU who creates a repository for the study and provides the 'main' functions to be used in each study.

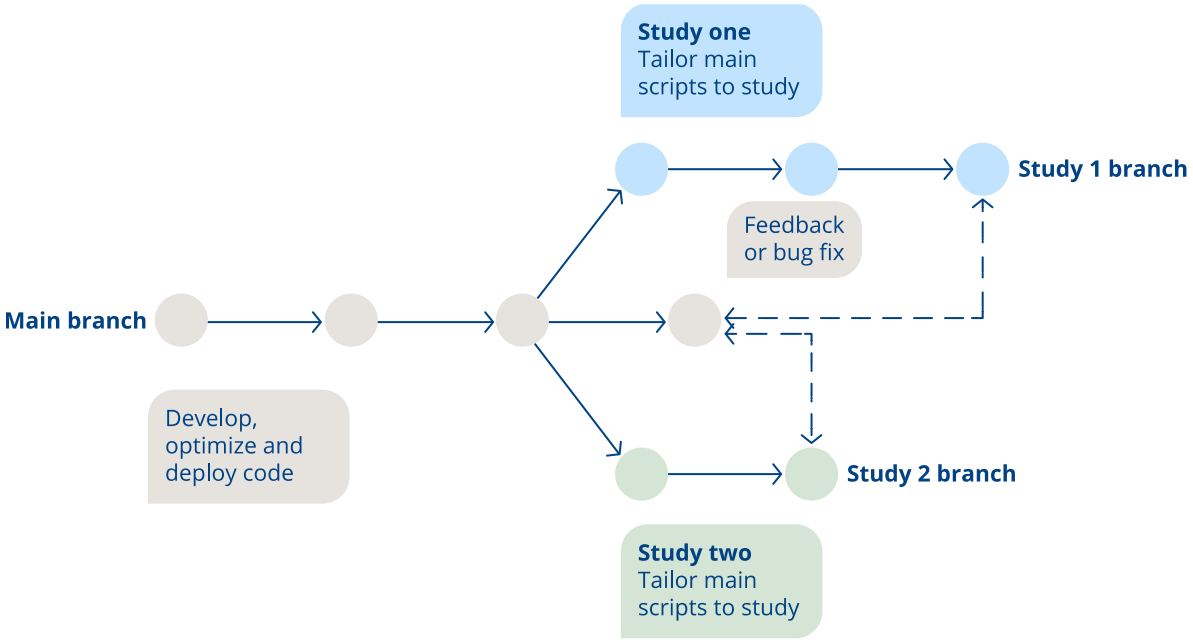
- A readme file is initialized with relevant information about the scripts.
- For each study, a 'branch' is created in which scripts are tailored to the respective study.
- After each version update, the coordinator requests from all teams to incorporate the changes using 'merge'. One responsible from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

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

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

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

Figure 5. *A workflow for collaborating on code with several studies in GitHub.*



8.7.3.2 Function creation and release

Functions are modules of code that accomplish a specific limited task. The development of functions consists of a series of 5 steps:

- Specifications of the function by the initiator
- Approval of the function specification by the owner.
- Programming of the function by a qualified programmer
- Testing of the function by another qualified programmer (the tester), who will complete a test specification form. For unit testing, all the test scenarios are added to a test script that needs to be performed after every update of the function.
- The function owner checks if all steps are complete, and deployment is approved.

8.7.3.3 Standard/bespoke analyses script creation, testing and release

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

1. Defining a map of the script, which includes specification of the folder structure, data model, graphical representation of the steps, use of functions, allocation of responsibilities and timelines, plus review schedules.

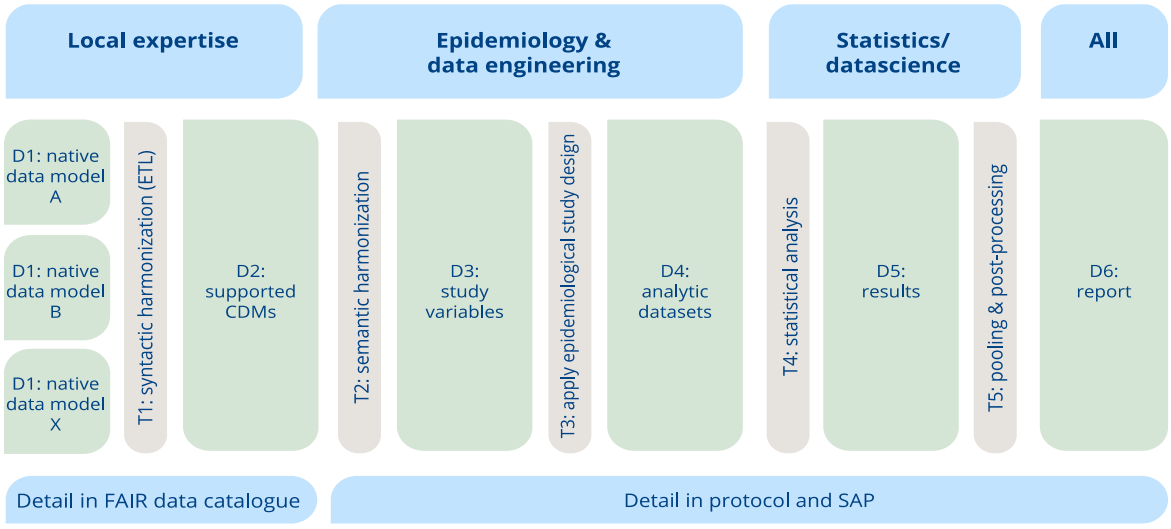
2. Programming of the code by a qualified programmer plus statistician. Test with code profiler to monitor bottlenecks in the code.
3. Testing the scripts by second programmer on:
    - Small simulated dataset in the specific CDM
    - Test on big data: run the script on the 1 million test dataset and test performance (see QAC1 for creation)
    - Test script on one real data partner before making it available for all the DAPS.
4. Take the script into deployment

For each update, steps 1-4 are repeated.

8.7.4. Data transformation

Based on the diagram in (Figure 6) the transformation steps (T) listed below in this section will be implemented by the data partners.

Figure 6. *Data Management Process.*



CDMs, common data models; Dn, data type; FAIR, findable, accessible, interoperable, and reusable; Tn, transformational step, for explanation, see below.

8.7.4.1 T1: Syntactic harmonisation (ETL)

Syntactic harmonisation through an extraction, transformation, and loading (ETL) process of native data into the ConcePTION CDM. In this CDM, data are represented using a common structure, but the content of the data remain in their original format. The CDM version that is used is v2.2, which is available as an open-source CDM. The CDM was developed as part of the IMI-ConcePTION project (project number IMI-821520). The ETL process has various structured steps as described in Thurin et al. 2021 (18,19):

1. DAPs are asked to share the data dictionaries of their data banks (tables and variable names/structure) with the principal Investigator.

2. Based on the data dictionaries of the data banks, an interview is conducted by the principal Investigator that explores what action(s) prompts the creation of a record, what is missing, and the context of each of the data banks.

3. Metadata (descriptive data about the data sources and databanks), data dictionaries, and interview answer sheets are uploaded in the VAC4EU FAIR (findable, accessible, interoperable, and reusable) data catalogue, according to the metadata structure for electronic health data that was defined in IMI-ConcePTION and the European Medicines Agency (EMA)–funded MINERVA project.

4. An overview is created of all the required study variables and definitions in order to create the code lists to identify the outcomes and covariates (see T2).

5. Instructions for the ETL design are provided by the WP2, these instructions comprise:

    - The required CDM tables

    - Mandatory variables

    - The calendar period over which data needs to be extracted

    - Code lists for the data to be extracted

    - The ETL design for each study can be conducted on paper or directly in the FAIR catalogue. The VAC4EU FAIR data catalogue (software and server provided through MOLGENIS: https://www.molgenis.org) is a metadata management tool designed to contain searchable metadata describing organisations that can provide access to specific data sources (VAC4EU). The ETL design section will detail how the native data are mapped to the different CDM tables. Review of the ETL design is conducted by the study team.

    - Once the ETL design is approved, it can be executed by the DAP using its programming language. The output are CSV (comma-separated values) files.

    - Once the ETL has been conducted, Level 1 and 2 data quality checks will be conducted to measure the integrity of the ETL, as well as internal consistency within the context of the CDM. University Medical Center Utrecht (UMCU) is responsible for the Level 1 and 2 scripts.

    - Level 1 data checks are performed to assess the completeness and content for each variable in all CDM tables to ensure that the mandatory variables contain data and conform to the formats specified by the CDM specifications, eg, data types, variable lengths, formats, acceptable values. Each DAP will be responsible for running the Level 1 R-scripts to complete the Level 1 checks (see https://github.com/IMI-ConcePTION/Level-1-checks/tree/master). A standard R Markdown report describing results of the checks for each table of the CDM will be produced by the script. Level 1 checks will be reviewed by the UMCU team with each of the DAPs, to assess completeness and consistency. After addressing any issues identified in Level 1 checks, DAPs may rerun the script and inspect the results. This control and correct process can be repeated until the ETL is judged to be sufficiently complete and correct by the DAP.

6. Level 2 data checks are performed to assess the logical relationship and integrity of data values within a variable, or between 2 or more variables within and between tables. Records occurring outside of recorded person-time, ie, before birth, after death, or outside of recorded observation

periods, will be assessed. Each DAP is responsible for running the Level 2 R-scripts to complete the 2 checks https://github.com/IMI-ConcePTION/Level-2-checks). An R Markdown report describing results of the Level 2 checks for each CDM table will be produced. After addressing any issues identified in Level 2 checks, DAPs may rerun the script and inspect the results together with the UMCU team. This control and correct process can be repeated until the ETL is judged to be sufficiently complete and correct by the DAP.

7. Each of the corrections, changes, and edits between sequential Level 1 and 2 checks and adaptations is documented in specific database instance reports and signed off by the DAP using specific forms.

8.7.4.2 T2: Semantic harmonisation

To reconcile differences between terminologies and native data availability, a shared semantic foundation needs to be built for the creation of relevant study variables. This is a multistep process:

1. **Definition of study variables**, which is done using an event definition form that systematically captures the following items and is a living document that will be closed upon study ending.

    - Purpose of the event: covariate or outcome
    - Version
    - Document history
    - Clinical definition
    - Synonyms/lay terms (for text mining purposes)
    - Laboratory tests specific for diagnosing event
    - Diagnostic tests specific for diagnosing event
    - Drugs that are used to treat event
    - Procedures used to treat event
    - Setting where condition is diagnosed (hospital, outpatient, GP)
    - Diagnosis codes or algorithms used in other papers (health outcomes of interest)
    - Codes used for study
    - Algorithm proposal
    - References
    - Examples of event definition forms can be found in the VAC4EU Zenodo repository (https://zenodo.org/communities/vac4eu/?page=1&size=20)

2. **Initial code lists** are created using the VAC4EU CodeMapper tool (https://vac4eu.org/codemapper/) to assist in the creation of code sets from case definitions for several coding systems simultaneously while keeping a record of the complete mapping process. Study variables are named in a standard hierarchical fashion based on body system.

3. **Review of the codes by DAPs:** The output of the CodeMapper is a Microsoft Excel list, which will be inspected by the DAPs and commented on at the VAC4EU SharePoint.

4. **Consolidation:** Comments from DAPs are consolidated by the study team. The code lists are read automatically through R/SAS code and will:

    - Check for ranges in output
    - Check for strange codes

- Insert the codes in the creation of concept sets, which are used to extract the data from the various CDM tables

5. Based on the relevant diagnostic medical codes and keywords, as well as other relevant components (eg, medications), 1 or more algorithms may be constructed to operationalise the measurement of each study variable. These algorithms may differ by database, as the components relevant for the study variables may differ.

6. During the T2 step, transformations occur for a series of steps related to completion of missing features in the data, eg, dose of vaccines, sorting on record level, combination of concepts for algorithm, and rule-based creation of study variables on a personal level for the study population, specific if needed per DAP.

7. Once the study variables are created, Level 3 checks will be deployed, which will be targeted to assess the patterns of study variables between data sources and against external benchmarks. A public example is available from the IMI-ConcePTION GitHub (https://github.com/IMI-ConcePTION/Level-3-checks).

8. The Level 3 checks are currently divided in 8 major modules, which can be tailored to the specific study variables:

- Source and study population
- Medicines
- Vaccines
- Diagnoses
- Pregnancy
- Populations of interest
- Health-seeking behaviour and lifestyle factors
- EUROCAT indicators

An R Markdown report describing results of the Level 3 checks will be produced by the script. Level 3 checks will be reviewed by the study team with each of the DAPs, to assess whether study variables are fit for purpose. After addressing any issues identified in Level 3 checks, DAPs may rerun the script and inspect the results. This control and correct process can be repeated until the Level 3 checks are judged to be sufficiently complete and correct by the DAP.

8.7.4.3 T3: Application of epidemiological study design

Based on the creation of the study variables on a person level or a medicines level, epidemiological designs will be applied such as sampling, matching (on specific variables and/or propensity scores), and selection based on inclusion and exclusion criteria.

For some data access providers (DAPs), before extensive medical data can be extracted, preliminary matching of vaccinated and comparators on key demographic (eg age, sex, calendar time, and region) is needed since not all the medical data can be extracted for the D1 data instance. These designs are defined in the statistical analysis plan (SAP) and may differ per study objective. Subsequent matching on medical conditions and other variables will then be conducted as part of the T3 step, once the required medical information is extracted.

The designs will be implemented for the various study objectives using R-scripts, and these may use the existing functions (R-cran) or macros.

8.7.4.4 T4: Statistical analysis

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

8.7.4.4.1 Scripting and deployment

The analytical R scripts that produce the T2-T4 steps are produced on VAC4EU GitHub for version control, links to the latest script will be distributed to DAPs for local deployment. Any issues can be notified on the GitHub, and the data engineers who are responsible for the R code will work with the local DAP to resolve issues if they occur. Scripts will be developed independently, based on a data engineering program and codebook in R (UMCU/ARS), and the output will be compared against each other for validation.

8.7.4.5 T5: Results and pooling post-processing

The aggregated results produced through T4 will be uploaded to the Digital Research Environment (DRE) for pooled analyses and visualisation. The DRE is made available through VAC4EU and UMCU (The anDREa consortium 2021). The DRE is a Microsoft Azure cloud-based, research environment with double authentication where researchers can collaborate using data that are stored and organised securely. UMC Utrecht is responsible for data processing and data security.

All researchers who need access to the DRE will be granted access to study-specific secure workspaces by VAC4EU/UMCU. Access to the workspaces will be possible only after double authentication using an identification code and password together with the user's mobile phone for authentication.

Uploading of files will be possible for all researchers with access to the workspace within the DRE. Downloading of files will be possible only after requesting and receiving permission from a workspace member with an "owner" role, who will be a UMCU team member.

8.8 Data analyses

Analyses will be conducted using R version R4.0.3 or higher (Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org). Additional tools as RevMan 5.4 or Stata® version 16 or higher will also be used when needed.

8.8.1. Descriptive study

Distributions of baseline and Covid-19 vaccination characteristics at time zero will be assessed and reported in the all-Covid-19 vaccinated population (Table 1 and 2 in Annex 1_Data Specification) and in the matched populations to describe and illustrate differences between the compared groups by scheme (homologous versus heterologous; or non-vaccinated in children; booster and non-booster) (Tables 5, 6, and 7).

The number and proportion of different brands used in homologous and heterologous primary vaccinations and boosters, over the total patients vaccinated by month will be reported. Table 3-4.

For continuous variables, means, standard deviations, medians, and quartiles will be estimated and reported by compared groups. For categorical variables, counts and proportions will be reported by compared groups. The missingness of variables will also be described.

Furthermore:

- Incident rate (IR) of each Covid-19 outcome (i.e. severe covid, death with covid and all covid) and 95% confidence intervals (CIs) by **primary vaccination** matched cohorts will be estimated overall, by age groups (adults, adolescents and children), by brand of the 1st dose and time after full primary vaccination (time0; only the pairs in which both individuals are still at risk at the beginging of each studied period will be included).

- IR (and 95% CI) of each Covid-19 outcome (i.e. severe covid, death with covid and all covid) CIs will be estimated in the **booster and non-booster** matched cohorts overall adults and by type of primary vaccination scheme (heterologous or homologous), type of booster (heterologous or homologous) and time since booster-time0 (only the pairs in which both individuals are still at risk at the beginging of each studied period will be included).

- Generate IPW-weighted Kaplan-Meier curves to depict the cumulative incidence of the outcomes by matched cohorts over time after time0 (for adults, adolescents and children separately, and overall and by brand i.e. PF, MD, AZ) and booster-time0 (for adults by booster and no-booster matched cohorts overall, by primary scheme (heterologous and homologous separately).

8.8.2. Comparative effectiveness study (primary objectives 1,2,3, 4 and secondary objective)

For each matched compared cohort, the following will be performed:

- Study the risk of Covid-19 related outcomes, applying Cox proportional hazards regression with robust variance estimators to derive average hazard ratio (HR) and 95% CIs. The assumption of proportionality of the survival curves in the compared groups will be evaluated. If not filled, flexible parametrical models will be used to provide final effectiveness estimations. Matching is used to deal with the key parameters, and additional confounding, if any, will be adjusted in the hazard models. Only the pairs in which both individuals are still at risk at the beginging of each studied period will be included in the analysis of that period.

- The adjusted vaccine effectiveness (VE) will be estimated as 1 minus the adjusted HR (and 1-95% CIs). VE for Covid-19 outcomes will be presented for adults, adolescents and children separately, overall matched cohorts and by brand of the 1st dose (i.e. PF, MD, AZ)

Severe covid-19 and covid-19 related death analysis will be stratified by previous Covid-19 infection (yes/no) and by periods of predominant circulating Covid-19 genetic variants in every country covered in the study at time zero and time zero-booster as a proxy of variant-specific vaccine effectiveness (i.e. pre-alpha; alpha; delta; omicron) (20).

VE will be provided by data source.

8.8.2.1 Meta-analysis

For analysis restricted to subgroups of people (reaching reduced sample size) or rare outcomes, meta-analysis will be performed.

Using the main estimates from each data source, appropriate random-effects meta-analytic methods (inverse variance method or others, as needed) will obtain a combined effect estimate.

The analysis of outcomes will be based on individuals, not on number of events. We will use Stata and RevMan Version 5.4 for analyses. Both adjusted hazard ratios and incidence rate differences will be used to summarize findings and a forest plot will be produced with the data sources' estimates and the pooled estimate.

We will use the Chi^2 and I^2 statistics to test for heterogeneity of treatment effect between trials. We will consider a Chi^2 value $p < 0.05$ or I^2 value > 50% as indicative of heterogeneity. If data exhibit substantial heterogeneity ($I^2 > 60\%$), we will investigate possible causes.

8.8.3. Sensitivity analyses

Sensitivity analyses will be conducted restricting to pairs with a negative Covid-19 test recorded prior to time0 or booster-time0 in data sources with negative Covid-19 tests (SIDIAP and CPRD). A comparison of the VEs resulted between that analysis and main analysis in those data sources will provide an estimation of the effect of the differential access (if any) to Covid-19 testing among compared groups in the effectiveness.

As explained above, a sensitivity analysis for VE in children and for booster effectiveness (in which compared groups will not receive a vaccination at time0 and booster-time0) will be performed stratifying by whether the data source includes information of primary care consultations or not. This analysis will provide an estimation of the confusion of healthy vaccinee effect present in data sources not including primary care consultations/data.

9 Limitations

Regarding the access to the data, the majority of the data sources present in this study do not expect to encounter any particular limitations or threats due to their previous participation in VAC4EU and EU PE&PV studies (see table 2). The unique exception to these circumstances lies in the French SNDS data access. We have designed two strategies to overcome this limitation and access French SNDS data through: (a) the so-called APHP "permanent" access (obtained in February 2022), which frees our researchers from the regulatory approval circuit, only requiring that the APHP data scientist has the authorization (special training) to access and program in the SNDS bubble (which is the case for our data scientist); (b) if strategy (a) fails, then: signing an agreement between APHP and EPI-PHARE, a scientific interest group formed by the French Agency for Medicines and the National Health Insurance Fund that can permanently access SNDS data for more than 10 years. APHP will contract an EPI-PHARE data scientist to work on this study, as it would be of interest to the EPI-PHARE group.

According to our estimations, the exclusion of data sources without link to hospital data (CPRD and Pharmo), would not affect the found high probability to demonstrate a protective effect of the booster against severe Covid-19.

Although this protocol addresses many design considerations to avoid common biases of vaccine effectiveness research, studies of COVID19 vaccines may be subject to limitations common to non-randomized studies based in existing health care data.

The patient's test-seeking behavior may be associated with both severity of infection symptoms and personal health-seeking behavior, which may introduce selection bias or confounding. To avoid that effect, adjustment by number of previous visits among compared groups (controlling by healthy vaccinee effect), as well as a sensitivity analysis restricting to patients with negative test results (controlling by differential access to covid testing).

Confounding of the relationship between booster receipt and Covid-19 outcomes may be likely. The use of eligibility criteria to define a comparable exposure group (i.e. matched on the date of the booster, date of full vaccination, type of primary vaccination, region, sex, age and having no visits in the week before) and further adjustment by covariates taken into account in the statistical models will allow investigators to adequately minimize confounding.

Even though efforts will be made for proper control of bias and multivariate adjustment minimizing confusion, no information about the job and type of residence may not be available in the data sources (i.e. Prioritised groups to be boosted or certain scheme receipt, such as people living in nursing homes or health care workers). Consequently, confusion may still be present due to the higher probability of infection among them versus other social groups. That aspect would direct towards a reduction in the effectiveness estimations. Matching by date of the 1st and 2nd dose, will prevent the differential effect of prioritized groups in all databases, because it determines the prioritized group (preference for full vaccination and boostering). In other words, people with homologous scheme will be compared to people with heterologous vaccination call to vaccinate with 1st and 2nd dose in the same moment.

If data about determinants, such as personal health-seeking behavior, were not recorded in the database, the selection of patients included in the compared groups and/or their dates to start the comparison could be biased, i.e.

- Patients prone to infections or to develop a severe/hospitalized infection or those with active respiratory infection/disease (prevalent active pneumonia, COPD, etc.) decided to attend to

receive the vaccination earlier/more than those less prone to it, we could observe patients more affected by covid among the vaccinated group. That aspect would direct towards a biased lower effectiveness estimation due to selection bias).

- The opposite could also be true, i.e., patients more adhered to general prevention measures (personal health-seeking behavior such as applying social distance, wearing masks, volunteer confinement, etc.) were more vaccinated, thus we could be observing more Covid-19 infections in the unvaccinated group. That aspect would direct towards a biased higher effectiveness estimation due to selection bias.

Selective recruitment into the study of subjects for compared groups recorded in the database with quality criteria (up-to-standard information) that are not representative of the general respective groups' subjects respectively in the source population could produce selection bias. For instance:

- If we were losing people having died from coronavirus disease even though they were vaccinated (i.e., non-effective vaccinations), selection bias would be present. Similarly, if more vaccinated participants were survivors of previous Covid-19 infections than vaccinated non-participants (i.e., they died by Covid-19) because those who attend the PCP (having information and minimum anamnesis in the database required to participate), while vaccinated non-participants do not.

- Also, if vaccinated participants were recorded in the database because they are more closely surveyed by the PCP/nurse due to their predisposition to complicated Covid-19 infection (i.e., baseline health conditions), while vaccinated non-participants were not recorded in the database because they do not seek healthcare, we would be including in the study patients with more probability to severe infection than the real overall vaccinated individuals.

The selection bias could direct towards any direction the effectiveness estimates.

If an individual's personal health beliefs and behaviors increase the likelihood of both Covid-19 vaccines booster receipt and seeking health care for the milder disease, then the analysis of this outcome may be particularly subject to confounding by health care–seeking behavior.

Misclassification of vaccine exposure, outcome status, or covariates is possible in existing health care data not collected for research purposes. Information about Covid-19 infection (through test results or codes) may not be captured reliably in databases. Additionally, covid-19 testing was not systematic all through the study period, and laboratory confirmation of case status may not be available always. Bias may be minimised since Covid-19 testing becomes near universal and repeated with individuals, with similar testing capacity and practice across regions, though. However, the marketing of patients' auto-diagnosis became popular over the months which could imply missing positive cases recorded in tests performed in clinical settings.

Since we will not have any better gold-standard for case status than the information about confirmed Covid-19 infection through lab results (used for case definition in most databases in the current study), its precision will not be evaluated but with the expected external incidences published by the health authorities. For that, before effectiveness estimation, a comparison (benchmarking) between the incidences of Covid-19 events in the participating databases and those reported by the corresponding country will be performed. Benchmarking will allow to confirm the events are the expected in the DAPs.Also, we consider that if COVID misclassification occurred it would be similar in both, heterologous and homologous primary vaccinations and booster and no-booting, however, surveillance or detection bias (different likelihood of screening or testing for COVID between the groups) cannot be discarded. If boosted individuals had less likelihood of screening or test than non-

boosted, we would artificially observe more cases among non-boosted, directing toward a biased higher booster effectiveness estimation. This may be less frequent when comparing heterologous and homologous primary vaccinations.

Data from multiple regions and sources will be included in the current study, thus, there may be variation in the capture and recording of various clinical elements. Additionally, distinct types of data sources will be used (e.g., records from general practice, hospital, and lab results from other registers) as well as different coding systems. Thus, the variables defined in different data sources may not exactly represent the same concept across data sources. The heterogeneity of effectiveness estimates across data sources may be due to the underlying heterogeneity of confounding control, misclassification, or other data source factors rather than true differences in the effectiveness.

Patterns of routine health care delivery and utilization may be disrupted during the Covid-19 pandemic as patients and providers forgo or delay routine preventive, elective, or non-emergency care. These disruptions in health care may result in under-ascertainment of important patient comorbidities in existing health care databases during periods of disruption. However, we will control this effect (in al databases by design) by matching by date that will prevent the differential effect of disrupted periods between compared groups (i.e. moment of vaccination with homologous versus heterologous second vaccination and boostering date).

With the increase in coverage, there is the potential for rapidly changing herd immunity in the population. This study is not designed to assess the overall and indirect effects of vaccination with Covid-19 vaccines.

Comparative analyses may not be possible in every setting and brand combination if, for instance, confounding is deemed to be insurmountable. However, descriptive information about vaccine recipients and crude incidence rates may still be informative and meaningful, even without calculations of vaccine effectiveness measures.

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

The use of secondary data could lead to a misclassification of both the exposure and the outcomes in terms of the result and the date. Nonetheless, both outcomes and exposures have been validated in previous studies. Asymptomatic Covid-19 infection could be underestimated as it is difficult that people without symptoms to do the test. Underestimation of positive At-Home COVID-19 Test will be higher over time, especially during the last months of 2021 and during omicron period. At-Home COVID-19 Test may not follow surveillance or not be compulsory declared in all countries. Some data sources do not contain the cause of deaths, but a proxy of death by covid has been established (as death occurring in Covid-19 people in a plausible temporary window).

The specific outcomes missed in a datasource, the latter will be excluded only where a specific analysis requires that outcome (i.e., CPRD will be excluded from analyses where hospitalization/ICU are required; SNDS will be excluded in the analysis of non-severe covid in children; Pedianets will be included only in analysis of pediatric population).

Perfect planned matching may not be reached for all heterologous primary vaccination or boosted patients. If perfect planned matching would reduce the sample to insufficient size, matching criteria will be relaxed (opening by each birth year into the aforementioned ranges; or days-by-day of date of 1st dose, time0 or booster-time0) till reaching 1 comparable individual per each one.

This study aims to compare different brand options, available in participant countries, for the 2nd dose and subsequent booster. Thus, one-dose Janssen Covid-19 vaccine or one-dose among patients with prior Covid-19 infection (I.e. homogeneous by defaults) were not considered in the current study. The comparability between the time after one-dose scheme and two-doses schemes may be compromised in observational studies, by confusion immediately after the start of the vaccination. Thus, people vaccinated with Janssen's vaccine (around 7 million in the covered countries; Table 1) will not be represented in the current study.

10 Protection of human subjects

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

10.1 Patient information

This study involves data that exist in anonymized structured format and contain no patient personal information.

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

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

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

10.2 Patient consent

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

10.3 Ethical conduct of the study

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

The study is a post-authorisation study of vaccine effectiveness and will comply with the definition of the non-interventional (observational) study referred to in the International Conference on Harmonisation tripartite guideline Pharmacovigilance Planning E2E and provided in the EMA Guideline on Good Pharmacovigilance Practices (GVP) Module VIII: Post Authorisation Safety Studies and with the 2012 EU pharmacovigilance legislation, adopted June 19, 2012.

The study will be registered in the EU PAS Register before data collection commences.

The research team and study sponsor should adhere to the general principles of transparency and independence in the ENCePP Code of Conduct and the ADVANCE Code of Conduct. There is no sponsor in the current study. The research team will apply for the ENCePP Study Seal.

The study will be conducted in accordance with legal and regulatory requirements, as well as with scientific purpose, value, and rigour, and will follow accepted research practices described in the Guidelines for Good Pharmacoepidemiology Practices (GPP) issued by the International Society for Pharmacoepidemiology (ISPE), and Good Epidemiological Practice guidelines issued by the International Epidemiological Association.

10.4 Institutional review board (IRB)/Independent ethics ccommittee (IEC)

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

11 Plans for Disseminating and Communicating Study Results

In its *Guidelines for Good Pharmacoepidemiology Practices (GPP)*, ISPE contends that "there is an ethical obligation to disseminate findings of potential scientific or public health importance".

Study results will be published following guidelines, including those for authorship, established by the International Committee of Medical Journal Editors. When reporting results of this study, the appropriate Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist will be followed.

Communication via appropriate scientific venues will be considered.

The study reports will be circulated among the participants of the collaborating public institutions for communication and review.

12 References

1. Andrews N, Tessier E, Stowe J, Gower C, Kirsebom F, Simmons R, et al. Duration of Protection against Mild and Severe Disease by Covid-19 Vaccines. New England Journal of Medicine [Internet]. 2022 Jan 27 [cited 2022 May 30];386(4):340–50. Available from: https://www.nejm.org/doi/10.1056/NEJMoa2115481

2. Pouwels KB, Pritchard E, Matthews PC, Stoesser N, Eyre DW, Vihta KD, et al. Impact of Delta on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. medRxiv [Internet]. 2021 Aug 24 [cited 2022 May 30];2021.08.18.21262237. Available from: https://www.medrxiv.org/content/10.1101/2021.08.18.21262237v1

3. Bernal JL, Andrews N, Gower C, Robertson C, Stowe J, Tessier E, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. BMJ [Internet]. 2021 May 13 [cited 2022 May 30];373. Available from: https://www.bmj.com/content/373/bmj.n1088

4. Hyams C, Marlow R, Maseko Z, King J, Ward L, Fox K, et al. Effectiveness of BNT162b2 and ChAdOx1 nCoV-19 COVID-19 vaccination at preventing hospitalisations in people aged at least 80 years: a test-negative, case-control study. The Lancet Infectious Diseases [Internet]. 2021 Nov 1 [cited 2022 May 30];21(11):1539–48. Available from: http://www.thelancet.com/article/S1473309921003303/fulltext

5. Pritchard E, Matthews PC, Stoesser N, Eyre DW, Gethings O, Vihta KD, et al. Impact of vaccination on SARS-CoV-2 cases in the community: a population-based study using the UK's COVID-19 Infection Survey. medRxiv [Internet]. 2021 Apr 23 [cited 2022 May 30];2021.04.22.21255913. Available from: https://www.medrxiv.org/content/10.1101/2021.04.22.21255913v1

6. Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, et al. Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. The Lancet [Internet]. 2021 May 1 [cited 2022 May 30];397(10285):1646–57. Available from: http://www.thelancet.com/article/S0140673621006772/fulltext

7. Ismail SA, Garcia Vilaplana T, Elgohari S, Stowe J, Tessier E, Andrews N, et al. Effectiveness of BNT162b2 mRNA and ChAdOx1 adenovirus vector COVID-19 vaccines on risk of hospitalisation among older adults in England: an observational study using surveillance data. 2021;

8. Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. New England Journal of Medicine [Internet]. 2021 Aug 12 [cited 2022 May 30];385(7):585–94. Available from: https://www.nejm.org/doi/10.1056/NEJMoa2108891

9. Andrews N, Stowe J, Kirsebom F, Toffa S, Sachdeva R, Gower C, et al. Effectiveness of BNT162b2 COVID-19 booster vaccine against covid-19 related symptoms and hospitalization in England. Nature Medicine. 2022 Jan 14;

10. Questions and answers on COVID-19: Vaccines [Internet]. [cited 2022 May 30]. Available from: https://www.ecdc.europa.eu/en/covid-19/questions-answers/questions-and-answers-vaccines

11. EMA and ECDC recommendations on heterologous vaccination courses against COVID-19: 'mix-and-match' approach can be used for both initial courses and boosters | European Medicines Agency [Internet]. [cited 2022 May 30]. Available from: https://www.ema.europa.eu/en/news/ema-ecdc-recommendations-heterologous-vaccination-courses-against-covid-19-mix-match-approach-can-be

12. Pottegård A, Kurz X, Moore N, Christiansen CF, Klungel O. Considerations for pharmacoepidemiological analyses in the SARS-CoV-2 pandemic. Pharmacoepidemiology and Drug Safety [Internet]. 2020 Aug 1 [cited 2022 May 30];29(8):825–31. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/pds.5029

13. Data | CPRD [Internet]. [cited 2022 May 30]. Available from: https://cprd.com/data

14. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology [Internet]. 2015 Jun 1 [cited 2022 May 30];44(3):827–36. Available from: https://academic.oup.com/ije/article/44/3/827/632531

15. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. International Journal of Epidemiology [Internet]. 2019 Dec 1 [cited 2022 May 30];48(6):1740–1740g. Available from: https://academic.oup.com/ije/article/48/6/1740/5374844

16. Qu'est-ce que le SNDS ? | SNDS [Internet]. [cited 2022 May 30]. Available from: https://www.snds.gouv.fr/SNDS/Qu-est-ce-que-le-SNDS

17. Willame C, Dodd C, van der Aa L, Picelli G, Emborg HD, Kahlert J, et al. Incidence Rates of Autoimmune Diseases in European Healthcare Databases: A Contribution of the ADVANCE Project. Drug Safety [Internet]. 2021 Mar 1 [cited 2022 May 30];44(3):383–95. Available from: https://link.springer.com/article/10.1007/s40264-020-01031-1

18. Thurin NH, Bosco-Levy P, Blin P, Rouyer M, Jové J, Lamarque S, et al. Intra-database validation of case-identifying algorithms using reconstituted electronic health records from healthcare claims data. BMC Medical Research Methodology [Internet]. 2021 Dec 1 [cited 2022 May 30];21(1):1–8. Available from: https://link.springer.com/articles/10.1186/s12874-021-01285-y

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

20. Data on SARS-CoV-2 variants in the EU/EEA [Internet]. [cited 2022 May 31]. Available from: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea