# BY-COVID. Real-world effectiveness of SARS-CoV-2 primary vaccination against SARS-CoV-2 infection: observational federated study across several EU regions.

Version 0


Funder

European Commission||EC

Grant

Beyond COVID/ No 101046203


 DOI: - 09/02/2023

## Researchers

Nina Van Goethem (orcid:0000-0001-7316-6990), Natalia Martínez-Lizaga (orcid:0000-0002-9586-7955), Javier González Galindo (orcid:0000-0002-8783-5478), ENRIQUE BERNAL-DELGADO (orcid:0000-0002-0961-3298), Marjan Meurisse (orcid:0000-0002-4409-0664), Francisco Estupiñán Romero (orcid:0000-0002-6285-8120), Santiago Royo-Sierra (orcid:0000-0002-0048-4370)

## Organizations

Aragon Health Sciences Institute (IACS), Sciensano (Belgium)

  DOI: - 09/02/2023

# Datasets

Title: COVID-19 public health cohort in Aragon (Spain)

Template: Horizon Europe

The COVID-19 public health cohort in Aragon links selected variables from existing population-level registries for COVID-19 public health surveillance and the COVID-19 vaccination program in Aragon, covering a global population for the region of approximately 1.3 million lives.

This cohort includes all registered COVID-19 cases reported to the Health System in Aragon (Spain) in the public health surveillance registry and complementary information on patients' attributes (i.e. underlying health problems, socio-demographic and economic factors), test results, and healthcare use combined with all citizens receiving at least one dose of any of the available vaccines included in the COVID-19 vaccination programme for Aragon.

Original data is maintained and updated daily by BIGAN (Health Data Space in Aragon).

A common data model specification with the required information to achieve the expected results of the baseline use case research question is provided as

  DOI: - 09/02/2023

supporting documentation at BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Common Data Model Specification.

Data is provided pseudonymised and in compliance with the baseline use case data model specification by BIGAN after the data request is processed and granted by the Aragon Ethical Research Committee.

## Dataset Description

### 1.1 Brief description of the described research output

#### 1.1.1 What kind of research output are you describing?

Research Data

#### 1.1.2 Is it physical or digital?

Digital

#### 1.1.3 Are you generating or re-using it?

Re-used

The analysis requires the secondary use of pseudonymised individual-level data from COVID-19 registries (COVID-19 public health surveillance and monitoring information system of COVID-19 cases and SARS-CoV-2 vaccination population registry), insurance registries or health system users databases (i.e., patient administrative information) or data from Electronic Health Records (EHR, i.e., comorbidities). All data sources contain routinely collected data from healthcare or public health information systems.

#### 1.1.4 What is the type of the described dataset?

Observational

The analysis will be based on the secondary use of routinely collected data from healthcare and public health information systems in Aragon (Spain).

#### 1.1.5 What is its format?

Comma Separated Values

  DOI: - 09/02/2023

The original data is processed by BIGAN under the privacy and security policies of the Health Sciences Institute in Aragon and in compliance with the specifications provided by the common data model within the data access request and provided as a pipe-separated CSV file using UTF-8 format.

### 1.1.6 What is its expected size?

GB

### 1.1.7 Why are you collecting/generating or re-using it?

To make informed decisions

The data analysis is expected to respond to the proposed research question detailed in BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol, regarding the real-world effectiveness of the SARS-CoV-2 vaccination programs in several European countries to inform public health vaccination policy.

### 1.1.8 What is its origin / provenance?

Original data sources (for Aragon):

- COVID-19 registries (COVID-19 public health surveillance and monitoring information system of COVID-19 cases and SARS-CoV-2 vaccination population registry),
- insurance registry or health system users databases (i.e., patient administrative information),
- data from Electronic Health Records (EHR, i.e., comorbidities).

  

All data sources contain routinely collected data from healthcare or public health information systems.

The research question is expected to be solved using a federated approach. The scripts for the analysis will be distributed and locally deployed/run at each participant's site using their data. Each participant country/region (Aragon (Spain), Belgium, Austria, Finland; Norway, Estonia and The Netherlands) will pull the data from their respective sources, conditional on data availability and access.

#### 1.1.9 To whom might it be useful ('data utility')?

Research communities

Population health and public health research communities could benefit from the outputs of the BY-COVID WP5 T5.2 Baseline Use case as it demonstrates the use of linked routinely collected real-world data to assess the effectiveness of a public health measure at a population level in several European countries.

### 2.1 Publications

#### 2.1.1 Does the described output support any scientific publication?

No

#### 2.1.2 Is there a data availability statement provided along with the publication?

No

### 2.3 Software

#### 2.3.1 Does the described output use or support any software?

No

  DOI: - 09/02/2023

### 3.1.1 Making data findable, including provisions for metadata

#### 3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

- Data identifiers
- Researchers identifiers
- Projects identifiers

URL

ORCID

Cordis

The COVID-19 public health cohort in Aragon (Spain) is described as a data source in the Health Information Portal (URL: ).

Further information on the baseline use case common data model is published in Zenodo BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Common Data Model Specification, including information metadata and extended information on the causal model, data schema, and a synthetic data set complying with the common data model specifications.

All researchers authoring and contributing to the Baseline Use Case data model specification and the Study protocol are referenced by their ORCID.

The data model and study protocol are published within the BY-COVID, Beyond COVID EU Project 2021-2024 Zenodo community under the CORDIS project id BY-COVID - Beyond COVID (101046203).

#### 3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

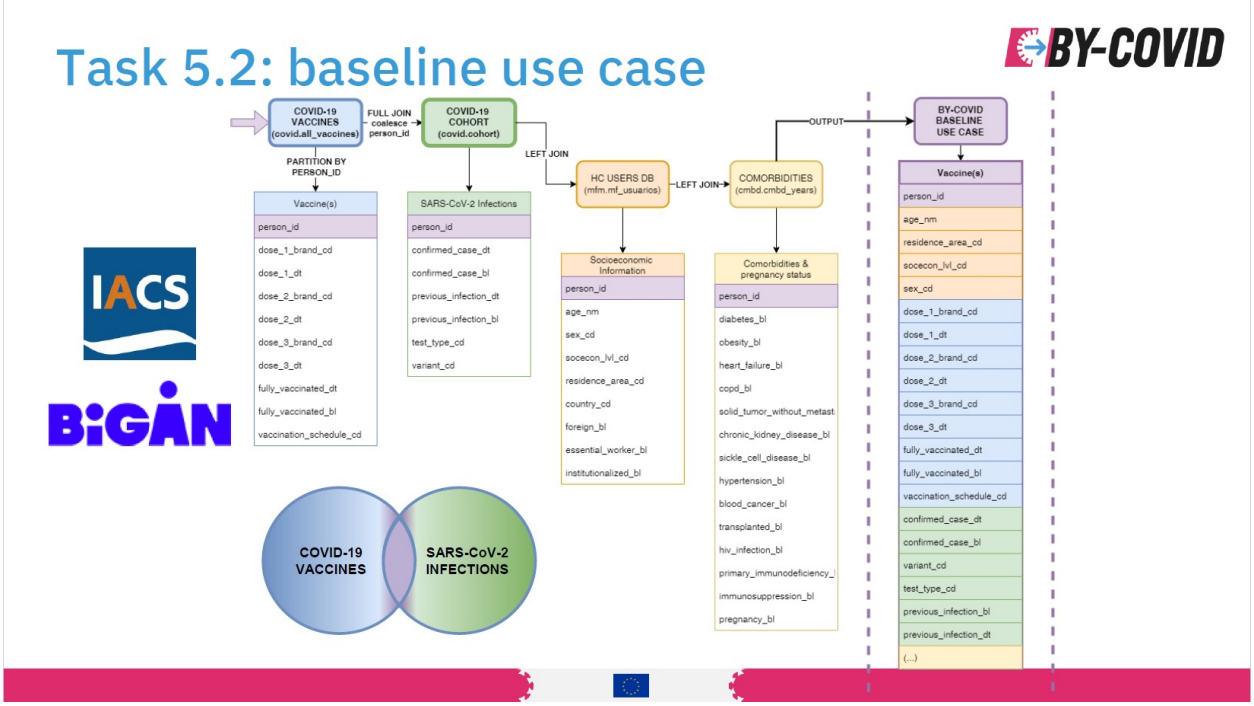


*Image 1*

Complete metadata, both in human-readable and machine-readable formats, are provided for the Baseline Use Case data model at https://doi.org/10.5281/zenodo.6913046:

- COVID-19 vaccine effectiveness data model specification (XLXS) - Human-readable version (Excel)
- COVID-19 vaccine effectiveness data model specification dataspice (HTML) - Human-readable version (interactive report)
- COVID-19 vaccine effectiveness data model specification dataspice (JSON) - Machine-readable version

### 3.1.1.3 What type(s) of metadata?

- Descriptive

• Administrative

• Structural

• Reference

Published Metadata includes:

- ***Administrative information*** on the Project name, Project URL, work package, Use case, Type of document, Version (SEM), Authors, Contributors, General Description of the project, and data model change log.

- *Cohort definition*, including cohort description, inclusion and exclusion criteria, and study period.

- ***Individual-level information requirements***, including

- ***model entities***: entity,
- ***DAG (correspondance with the causal model)***: nodes,

- ***Variable syntactic information***: variable label, variable description (concept), encoding, variable format, variable type, units, requirement level (required/recommended/optional), variable validation rules, transformations at origin, variable property, possible data source(s), and comments

- ***Area-level information requirements*** (similar information),

- and ***entity/variable semantic definition***, including:

- *entity name,*

- *entity description,*

  DOI: - 09/02/2023

- *entity definition*,

- *other specifications,*

- and crosswalks, including:

- *classifications system (international or national standard classifications for diseases, procedures or drugs)*,

- *code*,

- *clean code* (code without special characters),

- *code description*,

- *source of the code,*

**-** *and other observations on the code*

- ***Other administrative information*** published regarding the original data's geographical, population and time coverage of the expected data input is provided in the same publication as additional information.

  DOI: - 09/02/2023

#### 3.1.1.4 Do the metadata use standardised vocabularies?

Yes

Couldn't find it? Insert it manually

#### 3.1.1.5 Please provide URL/Description of used vocabularies

ISO/IEC 5218, NUTS 2021 codes, ISO 3166-1 alpha-3, ISO 8601, WHO-ATC/DDD Index 2023, SNOMED-CT, ICD-10-MC, ICD-9-MC, e90 (belgian pseudopathology codes)

- sex_cd using ISO/IEC 5218,
- residence_area_cd using NUTS 2021 codes,
- country_cd using ISO 3166-1 alpha-3,
- any date using ISO 8601,
- any drug using WHO-ATC/DDD Index 2023,

  

- any diagnoses using SNOMED-CT or ICD-10-MC or ICD-9-MC, or e90 (Belgian pseudopathology codes), following the crosswalks at the entity definition level

#### 3.1.1.6 Are the metadata searchable?

No

#### 3.1.1.8 Are keywords provided in the metadata?

Yes

COVID-19, vaccines, comparative effectiveness, causal inference, international comparison, SARS-CoV-2, common data model, surveillance

Keywords are provided with both in the common data model record in Zenodo and the study protocol in Zenodo.

Keywords do not follow any standard.

#### 3.1.1.9 Are metadata harvestable?

Yes

All metadata of the common data model for the Baseline Use Case is harvestable using the OAI-PMH API from Zenodo https://doi.org/10.5281/zenodo.6913046

### 3.2.1 Repository

#### 3.2.1.1 In which repository will the dataset / output be deposited?

BIGAN - Health Data Space in Aragon

BIGAN is the Health Data Space infrastructure in Aragon (Spain).

More information on BIGAN - Big Data in Healthcare, Aragon https://bigan.iacs.es/es/inicio

BIGAN is a technological platform that integrates all data from the health system for healthcare managers, educators, and researchers.

BIGAN is the Big Data project of the Department of Health of the Government of Aragon, created to improve healthcare using data routinely collected within Aragon's public health system. The project's development has been entrusted to the Aragon Institute of Health Sciences (IACS).

The project aims to integrate all data collected within the health system on a technological platform, where it can be analysed by healthcare professionals, managers, educators, and researchers. The ultimate goal is to improve the healthcare system and the health of residents in Aragon through data observation. To achieve this, collection, analysis, and sharing of information between all involved stakeholders is vital.

BIGAN is a technological infrastructure owned by the Government of Aragon, managed by the Aragon Institute of Health Sciences, and financed by the Department of Health and the Aragon Health Service, in which information from health information systems and other useful sources is captured, anonymized, safeguarded, and analysed, to fulfil BIGAN’s objectives.

BIGAN guarantees the protection and anonymization of data at all times. No researchers who use the data will ever have access to any personal information that would enable identification of either the patient or their environment.

BIGAN and the project's development is regulated by ORDER SAN/1355/2018, of August 1, published in the Official Gazette of Aragon on August 22 and in which the terms of its creation are established.

#### 3.2.1.2 Is the selected repository a trusted source?

Yes

- Follows repository standards
- Details terms of use
- Supports retention
- Supports withdrawal
- Supports back up
- Assigns PIDs
- Follows metadata standards
- Uses non-proprietary formats
- Supports mid- and long-term preservation

  DOI: - 09/02/2023

• Follows curation processes

• Supports authentication and authorization of users

• Has data security mechanisms in place

#### 3.2.1.4 Add appropriate arrangements made with the repository(ies) where the described dataset will be deposited

BIGAN Research is the service provided by the BIGAN platform that provides access to the data contained therein to researchers who wish to carry out biomedical research projects involving the secondary use of health data.

The data access service is managed by the Aragon Institute of Health Sciences and requires prior approval of the research protocol by the Clinical Research Ethics Committee of Aragon (CEICA).

#### 3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

#### 3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

The repository resolves a persistent identifier based on the project supporting each data request.

#### 3.2.1.7 Does the repository support versioning?

Yes

### 3.2.2 Data

#### 3.2.2.1 What is the described dataset / output title?

COVID-19 public health cohort in Aragon (Spain)

#### 3.2.2.2 How is the dataset / output shared?

Closed

**Data request**

  DOI: - 09/02/2023

To request access to BIGAN data, the following steps must be followed:

Preparing and presenting the research protocol and other necessary documentation for approval by the CEICA.

Once the CEICA has approved, a data request form must be provided

The data extraction and preparation process for research projects are subject to fees. The fee for the data extraction service, which must be taken into account when estimating the budget of the projects presented, covers the cost of the technological infrastructure and the services provided by the IACS in making the data available and in no case should be considered the “sale” of data.

### 3.2.2.3 What is the reason of limiting access to the dataset / output?

The type of data required to fulfil the study is individual-level pseudonymised health data routinely collected from healthcare and public health information systems, considered highly sensitive.

That access is only granted within the system-level security policies under specific circumstances and after undergoing a data access process following the procedure detailed below.

**Data access**

Any research group or consortium that wishes to carry out a biomedical research project with health data in Aragon can request access to health data managed by BIGAN, provided that it meets the following requirements:

In the case of a consortium of research groups, the applicant group or one of its members must be part of the Aragon research system.

The objective of the research must be non-profit and have a clear social interest.

The research project protocol must be approved by the Aragon Clinical Research Ethics Committee (CEICA) or another recognized Clinical Research Ethics Committee.

The project must have a corresponding feasibility report from the IACS Biocomputing Unit, verifying the requested data's availability, suitability for the project at hand, and compliance with all security, privacy, and data minimization requirements required by current regulations. To this end, the applicant must submit a data management plan (DMP) and the research protocol.

Projects being carried out or completed using data extracted from BIGAN will be publicized via this website. The key findings will demonstrate to citizens the use and benefits associated with analysing their health data information.

#### 3.2.2.5 Are there any methods or tools required to access the dataset / output?

  DOI: - 09/02/2023

Yes

Couldn't find it? Insert it manually

**BIGAN - Secure Processing Environments (BIGAN SPE)**

BIGAN provides researchers with access to the analytical power of its technological infrastructure via Secure Processing Environments (SPE). This allows for the execution of the necessary analyses and algorithms within BIGAN’s secure private cloud, without the need for removal of data from the platform at any time, and also allows the user to leverage the parallel processing infrastructure that has been developed specifically for this purpose.

The researcher can choose from various analytical tools and languages for this purpose.

### 3.2.2.6 Please provide information about the method(s) needed to access the dataset / output.

After data access request approval, researchers can access the expected data through BIGAN SPE Authentication, Authorization and Identification services or alternatively researchers can ask to be served the data as an encrypted file via the government of Aragon internal network.

https://bigan.iacs.es/en/services/data-access

### 3.2.2.8 Is the described dataset / output supported by a data access committee?

Yes

The research project protocol must be approved by the Aragon Clinical Research Ethics Committee (CEICA) or another recognized Clinical Research

Ethics Committee.

In the case of a consortium of research groups, the applicant group or one of its members must be part of the Aragon research system.

The objective of the research must be non-profit and have a clear social interest.

The project must have a corresponding feasibility report from the IACS Biocomputing Unit, verifying the requested data's availability, suitability for the project at hand, and compliance with all security, privacy, and data minimization requirements required by current regulations. To this end, the applicant must submit a data management plan (DMP) and the research protocol.

### 3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Data can be accessed during the project to produce the expected results of analyses planned within the study protocol following the research project's purposes, scope and objectives.

Data will not be accessible after the project ends, as data access is only granted for the research project's scope, purpose and duration.

Researchers commit within the data access request to delete the data upon the project's completion.

Aggregated data outputs and other outputs of the analysis can (or must) be published under open licenses as research outputs/outcomes at any time during or after the project's completion.

  

#### 3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

Aggregated data outputs and other outputs of the analysis can (or must) be published under open licenses as research outputs/outcomes at any time during or after the project's completion.

### 3.2.3 Metadata

#### 3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

See the information provided in section 3.1. Making data findable (...).

#### 3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

#### 3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

Information on the data access process, data request procedure and data access are provided as part of this Data Management Plan (DMP).

#### 3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

The common data model is published under an open licence (CC BY 4.0 International) in Zenodo.

This DMP will be published under an open licence (CC BY 4.0 International) in Zenodo and will be updated during the project.

## 3.3 Making data and other outputs interoperable

### 3.3.1 Does your (meta)data use a controlled vocabulary?

No

### 3.3.3 Have you applied a standard schema for your (meta)data?

Yes

Couldn't find it? Insert it manually

schema.org using the 'dataspice' R package

RO-crate using GitHub as the version control system (see https://by-covid.github.io/baseline-use-case-synthetic-crate/)

### 3.3.5 What is the methodology followed?

We use the 'schema.org' metadata standard using the 'dataspice' R package to produce the machine-readable version of the data model specification as both HTML and JSON files.

We additionally have set up a BY-COVID GitHub repository linked to the Zenodo publication of the data model to document a synthetic dataset simulated following the specifications of the data model as an RO-crate.

### 3.3.6 What community-endorsed interoperability best practices are followed?

We follow the general best practices and approach provided by the European Interoperability Framework (EIF)

### 3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

The study protocol has a qualified reference ('requires') with the common data model specification both published in Zenodo.

## 3.4 Increasing data and other outputs reuse

### 3.4.1 What internationally recognised licence will you use for your dataset / output?

Creative Commons Attribution-NonCommercial 4.0

All publications are under a CC BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/deed.en)

### 3.4.2 What reusability and / or reproducibility methods are followed?

• Readme files

• Codebooks

• Data cleaning

• Analyses

• Variable definitions

• Units of measurement

• Other

Other: We also share the data quality analysis (EDA), missing data imputation and exposed/controls matching algorithms.

### 3.4.3 Will you provide the described dataset / output in the public domain?

No

### 3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

We intend to ensure the use of the project outputs (including aggregated data) by third parties after the end of the project.

We will not share or publish any original individual-level data as this is subject to all restrictions within the data access request and national and international data protection regulations.

### 3.4.5 Is provenance well documented?

Yes

The provenance of the original data is well-documented in the study protocol, although not fully documented in the published metadata.

We only include as provenance information some information on the authors of the study and data managers of the original dataset regarding their names, contact information, affiliation (i.e. institution) and some information on the source of data provided by BIGAN (*see other links above*).

#### 3.4.6 What documented procedures for quality assurance do you have in place?

• Set up of scientific and technical committee

• Use of tools for automatic checks

• Data conform to format specification

• Consistency verified with data models and standards

BIGAN has a data quality assurance system for data processing to enable the secondary use of health data in Aragon (Spain). In addition, we implement our own data quality assessment and assurance procedures within the BY-COVID baseline use case, aiming to provide insight into the impact of data quality in interpreting the research outcomes and producing high-quality research. Those procedures include the assessment of the information requirements and the data model specification by a scientific and technical committee, the implementation of an automatic data quality check (i.e. exploratory data analysis - EDA) on the original dataset of each participant on-site, the implementation of data conformance and consistency checking following the common data model specification, and a final missing values assessment on core variables to inform decisions on imputation requirements.

### 4.1 Allocation of resources

#### 4.1.1 What will be the cost of making the described output FAIR?

a. 15,68

Euro

• Re-use

• Security

• Other

Data mining and processing (Data request processing)

Direct cost

Direct costs are 15,68€/hour (more information on fees is available in 'BIGAN>Services>Fees' https://bigan.iacs.es/en/services/fees) and include data request processing, a) data mining, ETL processing, and c) security and quality assurance.

The service includes defining queries and extracting information on available information sources, extracting the requested data and completing the ETL processes for complying with the requested common data model specification.

The global cost of a typical data request for a population-level research question is usually under 2000 euros and is served in approximately under 3 weeks time.

b. Euro

• Re-use

• Other

Data management, Data analysis

Indirect cost

Indirect costs are associated with the data management and data analysis to respond to the research question after the data request is served and considers only the person/time of the data manager and data scientists assigned to the task.

### 4.1.2 How will this cost be covered?

• Use of institution infrastructure

• Other

The Health Sciences Institute in Aragon usually covers the direct costs associated with the data request processing and extraction through BIGAN (regional infrastructure for Aragon). Indirect costs associated with the development and completion of the research use case are covered within the

  

research project budget dependent on Regional/National/European funding and the participant institution funding rate for the call supporting the project.

#### 4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

a. Natalia Martínez-Lizaga (orcid:0000-0002-9586-7955)

Data Manager (Research Project)

b. Javier González Galindo (orcid:0000-0002-8783-5478)

Couldn't find it? Insert it manually

Data Engineer

c. Santiago Royo-Sierra (orcid:0000-0002-0048-4370)

Couldn't find it? Insert it manually

Data analyst

d. Francisco Estupiñán Romero (orcid:0000-0002-6285-8120)

Couldn't find it? Insert it manually

Data Scientist

e. ENRIQUE BERNAL-DELGADO (orcid:0000-0002-0961-3298)

Couldn't find it? Insert it manually

Data Scientist. Baseline use case coordinator. WP5 co-coordinator.

### 5.1 Data Security

#### 5.1.1 What security measures are followed?

• Encryption

• Passwords

• Hash functions

Original data is de-identified and pseudonymised using a Hash function before being extracted from BIGAN (Data holder and secure processing environment)

and then served for research purposes encrypted and compressed, requiring a secure password to decompress the files.

The data request process includes a feasibility and security assessment of the information requirements specified in the data model considering data minimization to achieve the research objectives.

#### 5.1.2 What conditions do the security measures meet?

• Data access

• Data storage

• Data transmission

• Data recovery

• Data sharing

Secure data access is provided at servicing the data both within a secure processing environment (SPE) or through the internal network of the government of Aragon.

Data storage, transmission, recovery and sharing are assured by BIGAN as the regional Health Data Space.

Researchers are responsible for complying with data access conditions and data protection and privacy regulation.

#### 5.1.3 How will you preserve the described dataset / output in the long term?

BIGAN has a regional mandate from the government of Aragon to curate and maintain original data requests (and related ETL queries) in the long term and report on the re-use of health data for research purposes.

### 6.1 Ethical aspects

#### 6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

yes

Original individual-level data can not be shared as it is considered highly sensitive even after de-identification, pseudonymization and minimization.

  DOI: - 09/02/2023

Access to original data is granted only after approval by the Ethics Committee in Aragon of the research project and only within the scope of the research project.

https://www.iacs.es/investigacion/comite-de-etica-de-la-investigacion-de-aragon-ceica/ceica-evaluaciones-y-otras-presentaciones/ceica-proyectos-de-investigacion/

#### 6.1.2 Does the described dataset / output contain sensitive information?

Yes

#### 6.1.3 Does the described dataset / output contain personal data?

No

#### 6.1.4 What are the methods used for processing and accessing sensitive/personal information?

• Anonymising data where necessary

• Privacy constraints and applicable ethical norms

• Privacy policies

• National laws

• Other

See the information above regarding the data de-identification, pseudonymization and minimization upon extraction and processing.

More information on the BIGAN system-level security policies is available https://bigan.iacs.es/en/privacy

### 7.1 Other

#### 7.1.1 Do you make use of other procedures for data management?

No

Title: LINK-VACC: Linking of Registries for COVID-19 Vaccine Surveillance

Template: Horizon Europe

The LINK-VACC project links **selected variables from existing national registries** for COVID-19 vaccine surveillance to ensure the monitoring of COVID-19 vaccines in the phase following their marketing authorization (post-authorization surveillance). This includes the measurement of uptake and coverage of the vaccination, the estimation of vaccine effectiveness, and continuous monitoring of the vaccine's safety. For these purposes, existing pseudonymized data on COVID-19 laboratory test results, hospitalized COVID-19 patients, COVID-19 vaccinations, underlying health problems, socio-demographic and -economic factors, and healthcare worker status are linked.

By linking existing databases, the LINK-VACC project enables to create a **prospective cohort** of all individuals that are recorded in the vaccine registry as having received at least one dose of a registered COVID-19 vaccine and/or recorded in the COVID-19 Healthdata Database as having a positive test for SARS-CoV-2 (PCR or rapid antigen). This cohort covers more than 90% of the total population in Belgium. The COVID-19 vaccine related data from the vaccine registry (**VACCINNET+**) and the COVID-19 test result database (**COVID-19 Healthdata Database**) are continuously updated and linked (on a daily basis) to the database of hospitalized patients with a confirmed COVID-19 diagnosis (**COVID-19 Clinical Hospital Survey**) and to the common database for the different public institutions responsible for the recognition of healthcare actors in Belgium (**COBRHA**). In addition, there are non-continuous linkages (monthly, bi-annual or *ad hoc*) to the database on reimbursed care and medicines of citizens insured in Belgium (**Intermutualistic Agency - IMA**) and to the socioeconomic information (civil status, employment status, income decile, ...) from the **STATBEL** database.

  DOI: - 09/02/2023

Data from the vaccine registry (VACCINNET +), the TestResult database (COVID-19 Healthdata Database) and the COVID-19 clinical database (COVID-19 Clinical Hospital Survey), are hosted in the HealthData COVID-19 data Center (secured environment), where personal patient data are available. As described above, links are organized with the databases external to healthdata.be (COBRHA, IMA and STATBEL) using the national registry number. The Trusted Third Party (TTP) service of the eHealth platform pseudonymizes the patient's identifier and the data from the six databases are stored in **Healthdata's pseudonymised environment** (separation of secured environment and research environment).

In the context of the BY-COVID project, Sciensano will be one of the nodes participating in the baseline use case. These analyses will take place within the architecture of the LINK-VACC project (pseudonymised research environment) which allows to meet the assumptions of the common data model specifications. The required information to achieve the expected results of the baseline use case research question is provided as supporting documentation at BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Common Data Model Specification.

## Dataset Description

### 1.1 Brief description of the described research output

#### 1.1.1 What kind of research output are you describing?

Research Data

#### 1.1.2 Is it physical or digital?

Digital

#### 1.1.3 Are you generating or re-using it?

  DOI: - 09/02/2023

Re-used

The analysis specified in the BY-COVID baseline use case requires the secondary use of pseudonymised individual-level data from the vaccine registry containing information on administered COVID-19 vaccines (VACCINNET+), the COVID-19 test result database containing positive and negative PCR or rapid antigen test results for SARS-CoV-2 (COVID-19 Healthdata Database), the Belgian statistical office (STATBEL) containing socioeconomic information, the reimbursement data from the Intermutualistic Agency (IMA), and information on healthcare actors (COBRHA). All data sources contain routinely collected data from healthcare or public information systems. Individual-level linkage of these data sources based on the national registry number within a pseudonymised environment is organized by the IT architecture of the LINK-VACC project, hosted by the Healthdata.be service of Sciensano. No new data collection will be set up. The objectives of the BY-COVID baseline use case (studying vaccine effectiveness) are covered by the existing deliberation of the Information Security Committee of the LINK-VACC project.

### 1.1.4 What is the type of the described dataset?

Observational

The analysis will be based on the secondary use of routinely collected data from healthcare information systems and administrative data sources in Belgium that are linked based on the national reference number in the context of the LINK-VACC project.

### 1.1.5 What is its format?

Comma Separated Values

The original data in the LINK-VACC project are contained within a DB2 database, and will subsequently be processed in compliance with the specifications provided by the common data model and provided as a pipe-separated CSV file using UTF-8 format.

### 1.1.6 What is its expected size?

GB

### 1.1.7 Why are you collecting/generating or re-using it?

To make informed decisions

The data analysis is expected to respond to the proposed research question detailed in BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol, regarding the real-world effectiveness of the SARS-CoV-2 vaccination programs in several European countries to inform public health vaccination policy.

### 1.1.8 What is its origin / provenance?

Original data sources linked in the context of the LINK-VACC project and re-used to fulfill the requirements of the common data model specification of the baseline use case:

- The vaccine registry (VACCINNET+)
- The COVID-19 test result database (COVID-19 Healthdata Database)
- Socioeconomic data from the Belgian Statistical Office (STATBEL)
- Reimbursement data from the Intermutualistic Agency (IMA)
- The common database for the different public institutions responsible for the recognition of healthcare actors in Belgium (COBRHA)

All data sources contain routinely collected data from healthcare or public health information systems.

The research question is expected to be solved using a federated approach. The scripts for the analysis will be distributed and locally deployed/run at each participant's site using their data. Each participant country/region (Aragon (Spain), Belgium, Austria, Finland, Norway, Estonia and The Netherlands) will pull the data from their respective sources, conditional on data availability and access.

#### 1.1.9 To whom might it be useful ('data utility')?

Research communities

Population health and public health research communities could benefit from the outputs of the BY-COVID WP5 T5.2 Baseline Use case as it demonstrates the use of linked routinely collected real-world data to assess the effectiveness of a public health measure at a population level in several European countries.

## 2.1 Publications

#### 2.1.1 Does the described output support any scientific publication?

No

#### 2.1.2 Is there a data availability statement provided along with the publication?

No

## 2.3 Software

#### 2.3.1 Does the described output use or support any software?

No

#### 3.1.1 Making data findable, including provisions for metadata

##### 3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

- Data identifiers
- Researchers identifiers
- Projects identifiers

URL

  DOI: - 09/02/2023

ORCID

Cordis

A description of the LINK-VACC project is published on Sciensano's website (URL: https://www.sciensano.be/en/projects/linking-registers-covid-19-vaccine-surveillance) and the LINK-VACC cohort is described as a data source in the Health Information Portal (URL: https://www.healthinformationportal.eu/health-information-sources/linking-registers-covid-19-vaccine-surveillance). Further, the data is made discoverable on Sciensano’s FAIR portal (https://fair.healthdata.be/dataset/d43a158e-7d13-4660-bbc3-9d3f8d5501e5).

Further information on the baseline use case common data model is published in Zenodo BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Common Data Model Specification, including information metadata and extended information on the causal model, data schema, and a synthetic data set complying with the common data model specifications. All researchers authoring and contributing to the Baseline Use Case data model specification and the Study protocol are referenced by their ORCID. The data model and study protocol are published within the BY-COVID, Beyond COVID EU Project 2021-2024 Zenodo community under the CORDIS project id BY-COVID - Beyond COVID (101046203).

#### 3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

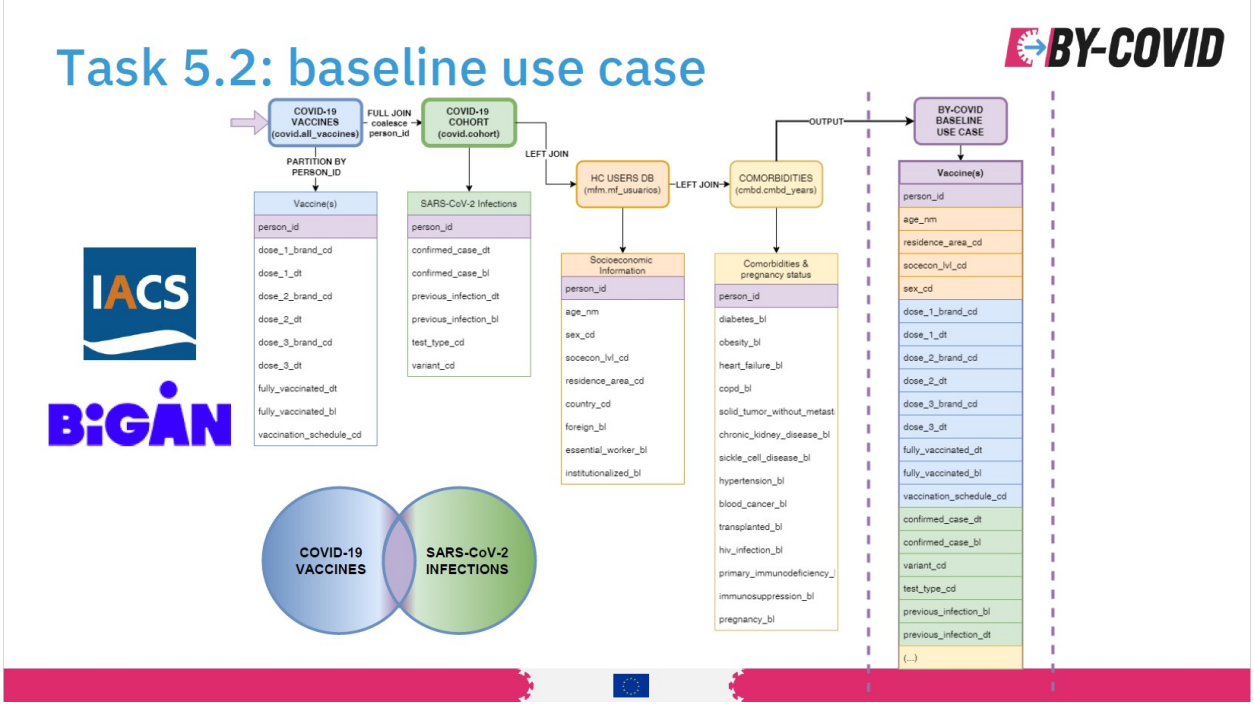


*Image 2*

Complete metadata, both in human-readable and machine-readable formats, are provided for the Baseline Use Case data model at https://doi.org/10.5281/zenodo.6913046:

- COVID-19 vaccine effectiveness data model specification (XLXS) - Human-readable version (Excel)
- COVID-19 vaccine effectiveness data model specification dataspice (HTML) - Human-readable version (interactive report)
- COVID-19 vaccine effectiveness data model specification dataspice (JSON) - Machine-readable version

### 3.1.1.3 What type(s) of metadata?

- Descriptive

• Administrative

• Structural

• Reference

Published Metadata includes:

- ***Administrative information*** on the Project name, Project URL, work package, Use case, Type of document, Version (SEM), Authors, Contributors, General Description of the project, and data model change log.

- *Cohort definition*, including cohort description, inclusion and exclusion criteria, and study period.

- ***Individual-level information requirements***, including

- ***model entities***: entity,
- ***DAG (correspondence with the causal model)***: nodes,

- ***Variable syntactic information***: variable label, variable description (concept), encoding, variable format, variable type, units, requirement level (required/recommended/optional), variable validation rules, transformations at origin, variable property, possible data source(s), and comments

- ***Area-level information requirements*** (similar information),

- and ***entity/variable semantic definition***, including:

- *entity name,*

- *entity description,*

  DOI: - 09/02/2023

- *entity definition*,

- *other specifications,*

- and crosswalks, including:

- *classifications system (international or national standard classifications for diseases, procedures or drugs)*,

- *code*,

- *clean code* (code without special characters),

- *code description*,

- *source of the code,*

**-** *and other observations on the code*

- ***Other administrative information*** published regarding the original data's geographical, population and time coverage of the expected data input is provided in the same publication as additional information.

  DOI: - 09/02/2023

#### 3.1.1.4 Do the metadata use standardised vocabularies?

Yes

Couldn't find it? Insert it manually

#### 3.1.1.5 Please provide URL/Description of used vocabularies

ISO/IEC 5218, NUTS 2021 codes, ISO 3166-1 alpha-3, ISO 8601, WHO-ATC/DDD Index 2023, SNOMED-CT, ICD-10-MC, ICD-9-MC, e90 (belgian pseudopathology codes)

- sex_cd using ISO/IEC 5218,
- residence_area_cd using NUTS 2021 codes,
- country_cd using ISO 3166-1 alpha-3,
- any date using ISO 8601,
- any drug using WHO-ATC/DDD Index 2023,

  

- any diagnoses using SNOMED-CT or ICD-10-MC or ICD-9-MC, or e90 (Belgian pseudopathology codes), following the crosswalks at the entity definition level

#### 3.1.1.6 Are the metadata searchable?

No

#### 3.1.1.8 Are keywords provided in the metadata?

Yes

COVID-19, vaccines, comparative effectiveness, causal inference, international comparison, SARS-CoV-2, common data model, surveillance

Keywords are provided both in the common data model record in Zenodo and the study protocol in Zenodo.

Keywords do not follow any standard.

#### 3.1.1.9 Are metadata harvestable?

Yes

All metadata of the common data model for the Baseline Use Case is harvestable using the OAI-PMH API from Zenodo https://doi.org/10.5281/zenodo.6913046

### 3.2.1 Repository

#### 3.2.1.1 In which repository will the dataset / output be deposited?

Healthdata.be

Data from the vaccine registry (VACCINNET +), the database I & II of the Cooperation Agreement of August 25, 2020 (COVID-19 Healthdata Database) and the COVID-19 clinical database (Clinical Hospital Survey), are hosted in the **Healthdata.be COVID-19 Data Center**.

The Healthdata.be platform (a service of Sciensano) has been established following the Belgian eHealth plan and offers secured environments to link individual-level data from different sources based on the national registry number, with the necessary guarantees in terms of information security and the protection of privacy. The mission of Healthdata.be is to facilitate the data exchange between healthcare professionals, patients and researchers according to the only once principle and the re-use of data, in order to increase public health knowledge and to adjust health care policy, with respect for the privacy of the patient, the healthcare professional and the medical confidentiality. More information on healthdata.be.

#### 3.2.1.2 Is the selected repository a trusted source?

Yes

- Follows repository standards
- Details terms of use
- Supports retention
- Supports withdrawal
- Supports back up
- Assigns PIDs
- Follows metadata standards
- Uses non-proprietary formats
- Supports mid- and long-term preservation

• Follows curation processes

• Supports authentication and authorization of users

• Has data security mechanisms in place

#### 3.2.1.4 Add appropriate arrangements made with the repository(ies) where the described dataset will be deposited

At Healthdata.be, the information collected may only be passed onto authorized researchers and supervisory doctors for the purpose of improving the quality and management of the health and healthcare sector. The data can therefore only be passed on for the purposes of scientific research, health monitoring and the promotion of knowledge. The data are never communicated to third parties for a purpose other than research or the protection of public health.

#### 3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

No

#### 3.2.1.7 Does the repository support versioning?

Unknown

### 3.2.2 Data

#### 3.2.2.1 What is the described dataset / output title?

Belgium COVID-19 public health cohort

#### 3.2.2.2 How is the dataset / output shared?

Closed

The objectives of the BY-COVID baseline use case (studying vaccine effectiveness) are covered by the existing deliberation of the Information Security Committee of the LINK-VACC project. Only researchers at Sciensano involved in the LINK-VACC project have access to the linked pseudonymized data. Only fully aggregated results (obtained after deploying the analysis pipeline) will be exported from the secure processing environment. As such, the BY-COVID baseline use case, with only anonymous data exports, does not require a deliberation from the Information Security Committee Social Security and Health.

  DOI: - 09/02/2023

Regarding the governance of Healthdata.be the following steps will be taken:

- Submission of a new project request via https://sciensano.service-now.com/. As Healthdata.be need to assist with the exports, this project should be notified and integrated within the healthdata.be planning.
- Healthdata.be will notify, as part of transparency policy the Healthdata.be Steering Committee about the BY-COVID baseline use case.

#### 3.2.2.3 What is the reason of limiting access to the dataset / output?

The individual-level datasets generated or analyzed during the current study do not fulfill the requirements for open data access. The data is too dense and comprehensive to preserve patient privacy. The data of the individual data sources within the LINK-VACC project are kept in the pseudonymized environment of Healthdata.be, and a link between the individual data in each of them takes place thanks to the use of a pseudonymized national reference number managed by Healthdata.be under a ‘project mandate’. A ‘project mandate’ consists of a group of individuals, a group of variables, and a time period. Access rights to the pseudonymized data in the Healthdata.be data warehouse are granted ad nominatum for the scientists involved in the surveillance activities at Sciensano.

#### 3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

### 3.2.2.6 Please provide information about the method(s) needed to access the dataset / output.

A cascade of users and passwords were assigned to the Healthdata.be platform, of only which a limited number of Healthdata.be employees, have access to the most disclosed information. Only the Sciensano researchers implied in the LINK-VACC project have access to the pseudonymised research environment. The access for the authorised users from the outside to the inside will be done by means of a secured Server Based Computing, which is also provided by the Healthdata.be platform and which offers a protection of the data and of the traffic.

Access to the secure research environment via Citrix.

Data from the DB2 database are loaded in a local R environment within the healthdata.be secure data environment, using the R package RJDBC which is loading a JDBC driver.

### 3.2.2.8 Is the described dataset / output supported by a data access committee?

Yes

The protocol of the LINK-VACC project was approved by the medical ethics committee University Hospital Brussels – Vrije Universiteit Brussel (VUB) on 03/02/2021 (reference number 2020/523) and obtained authorization from the Information Security Committee (ISC) Social Security and Health (reference number IVC/KSZG/21/034).

  DOI: - 09/02/2023

The BY-COVID baseline use case aims to use the data for the same objective (vaccine effectiveness) as defined in the context of the LINK-VACC project, and as no sensitive data will be shared outside the secured environment, the current use case in BY-COVID is covered by the existing deliberation of the ISC.

The architecture of the Healthdata.be-platform has an authorization of the ISC-SSH: https://www.ehealth.fgov.be/ehealthplatform/file/view/AXGCtcLOmTlaOSp4NmlQ?filename=15-009-f102-HD4Patient-modifi%C3%A9e%20le%203%20mars%202020.pdf

### 3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Data within the secured environment can only be accessed by authorized users (see 3.2.2.6) to produce the expected results of analyses planned within the study protocol following the research project's purposes, scope and objectives.

Data available in the LINKVACC secured data environment is used for other purposes than the investigation of the research question specified in this project, e.g., the measurement of uptake and coverage of the vaccination, the estimation of vaccine effectiveness, and continuous monitoring of the vaccine’s safety. Hence data will remain available for authorized researchers at Sciensano involved in the LINK-VACC project after the project ends.

Aggregated data outputs of the analysis will be published under open licenses as research outputs/outcomes at any time during or after the project's completion.

See 3.2.2.6

#### 3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

Aggregated data outputs and other outputs of the analysis can (or must) be published under open licenses as research outputs/outcomes at any time during or after the project's completion.

### 3.2.3 Metadata

#### 3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

See the information provided in section 3.1. Making data findable (...).

#### 3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

#### 3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

Information on the data access process, data request procedure and data access are provided as part of this Data Management Plan (DMP).

#### 3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

The common data model is published under an open licence (CC BY 4.0 International) in Zenodo.

This DMP will be published under an open licence (CC BY 4.0 International) in Zenodo and will be updated during the project.

## 3.3 Making data and other outputs interoperable

### 3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

Couldn't find it? Insert it manually

### 3.3.3 Have you applied a standard schema for your (meta)data?

Yes

Couldn't find it? Insert it manually

schema.org using the 'dataspice' R package

RO-crate using GitHub as the version control system (see https://by-covid.github.io/baseline-use-case-synthetic-crate/)

### 3.3.5 What is the methodology followed?

We use the 'schema.org' metadata standard using the 'dataspice' R package to produce the machine-readable version of the data model specification as both HTML and JSON files.

We additionally have set up a BY-COVID GitHub repository linked to the Zenodo publication of the data model to document a synthetic dataset simulated following the specifications of the data model as an RO-crate.

### 3.3.6 What community-endorsed interoperability best practices are followed?

We follow the general best practices and approach provided by the European Interoperability Framework (EIF)

#### 3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

The study protocol has a qualified reference ('requires') with the common data model specification both published in Zenodo.

### 3.4 Increasing data and other outputs reuse

#### 3.4.1 What internationally recognised licence will you use for your dataset / output?

Creative Commons Attribution-NonCommercial 4.0

All publications are under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/deed.en)

#### 3.4.2 What reusability and / or reproducibility methods are followed?

• Readme files

• Codebooks

• Data cleaning

• Analyses

• Variable definitions

• Units of measurement

• Other

Other: We also share the data quality analysis (EDA), missing data imputation and exposed/controls matching algorithms.

#### 3.4.3 Will you provide the described dataset / output in the public domain?

No

#### 3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

We intend to ensure the use of the project outputs (including aggregated data) by third parties after the end of the project.

We will not share or publish any original individual-level data as this is subject to all restrictions within the data access request and national and international data protection regulations.

### 3.4.5 Is provenance well documented?

Yes

The provenance of the original data is well-documented in the study protocol, although not fully documented in the published metadata.

We only include as provenance information some information on the authors of the study and data managers of the original dataset regarding their names, contact information, affiliation (i.e. institution) and some information on the source of data provided within the context of the LINK-VACC project (*see other links above*).

### 3.4.6 What documented procedures for quality assurance do you have in place?

• Set up of scientific and technical committee

• Use of tools for automatic checks

• Data conform to format specification

• Consistency verified with data models and standards

We implement our own data quality assessment and assurance procedures within the BY-COVID baseline use case, aiming to provide insight into the impact of data quality in interpreting the research outcomes and producing high-quality research. Those procedures include the assessment of the information requirements and the data model specification by a scientific and technical committee, the implementation of an automatic data quality check (i.e. exploratory data analysis - EDA) on the original dataset of each participant on-site, the implementation of data conformance and consistency checking following the common data model specification, and a final missing values assessment on core variables to inform decisions on imputation requirements.

## 4.1 Allocation of resources

### 4.1.1 What will be the cost of making the described output FAIR?

a. ¿¿??

Euro

• Re-use

  DOI: - 09/02/2023

• Security

• Other

Data mining and processing (Data request processing)

Direct cost

¿¿???

b. ¿¿??

• Re-use

• Other

¿¿?¿?

Indirect cost

Indirect costs are associated with ....

#### 4.1.2 How will this cost be covered?

• Use of institution infrastructure

• Other

¿¿??? we will change the above if necessary

#### 4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

a. Marjan Meurisse (orcid:0000-0002-4409-0664)

Data analyst

b. Nina Van Goethem (orcid:0000-0001-7316-6990)

Couldn't find it? Insert it manually

BY-COVID WP5 co-coordinator

## 5.1 Data Security

### 5.1.1 What security measures are followed?

• Encryption

• Passwords

Data from the vaccine registry (VACCINNET +), the TestResult database (COVID-19 Healthdata Database) and the COVID-19 clinical database (COVID-19 Clinical Hospital Survey), are hosted in the HealthData COVID-19 data Center (secured environment), where personal patient data are available. As described above, links
are organized with the databases external to healthdata.be (COBRHA, IMA and STATBEL) using the national registry number. The Trusted Third Party (TTP) service of the eHealth platform pseudonymizes the patient's identifier and the data from the six databases are stored in Healthdata's pseudonymised environment (separation of secured environment and research environment).

#### 5.1.2 What conditions do the security measures meet?

• Data access

• Data storage

• Data transmission

• Data recovery

• Data sharing

¿¿??? we will change the above if necessary

#### 5.1.3 How will you preserve the described dataset / output in the long term?

End date of LINK-VACC project?

### 6.1 Ethical aspects

#### 6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

yes

Original individual-level data can not be shared as it is considered highly sensitive even after de-identification, pseudonymization and minimization.

#### 6.1.2 Does the described dataset / output contain sensitive information?

  DOI: - 09/02/2023

Yes

#### 6.1.3 Does the described dataset / output contain personal data?

No

#### 6.1.4 What are the methods used for processing and accessing sensitive/personal information?

- Anonymising data where necessary
- Privacy constraints and applicable ethical norms
- Privacy policies
- National laws
- Other

¿¿??? we will change the above if necessary

### 7.1 Other

#### 7.1.1 Do you make use of other procedures for data management?

No