BY-COVID Spring 2024 Baseline Use Case Workshop:

Integration of individual-level socioeconomic data for infectious diseases research and prevention in Europe

Co-organised by KNAW-DANS/CESSDA, Sciensano and IACS

Funded by
the European Union

Programme

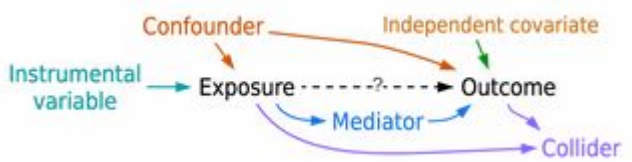
09:30 - 09:40	Brief tour de table, Introduction, Recap previous workshop & goals of the workshop
09:40 - 09:55	Overview of the BY-COVID Baseline Use Case (Enrique Bernal-Delgado)
09:55 - 10:20	Experience from Baseline Use Case partners: mapping data availability, procedures, and challenges (Francisco Estupiñán-Romero and Marjan Meurisse)
10:20 - 10:30	Break
10:30 - 11:15	Breakout sessions
	1. Exploring socioeconomic data sources in Europe: availability, limitations, and mobilisation
	2. Socioeconomic and health data linkage for EU-level research: challenges and solutions
	3. Advancing socioeconomic data integration: generalisation, sustainability, and policy relevance
11:15 - 11:30	Reporting back from breakout sessions & Wrap-up

Welcome, goals of the workshop, recap previous workshop & brief tour de table

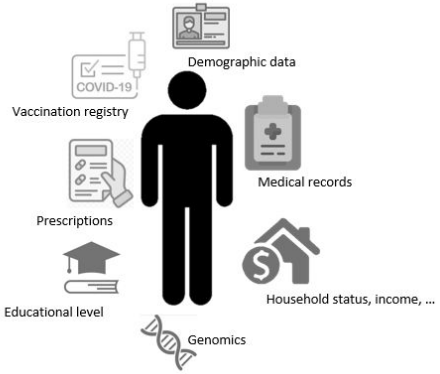
09:30 - 09:40

BY-COVID Baseline Use Case - Intro

- Answering policy-relevant research questions through **causal research**

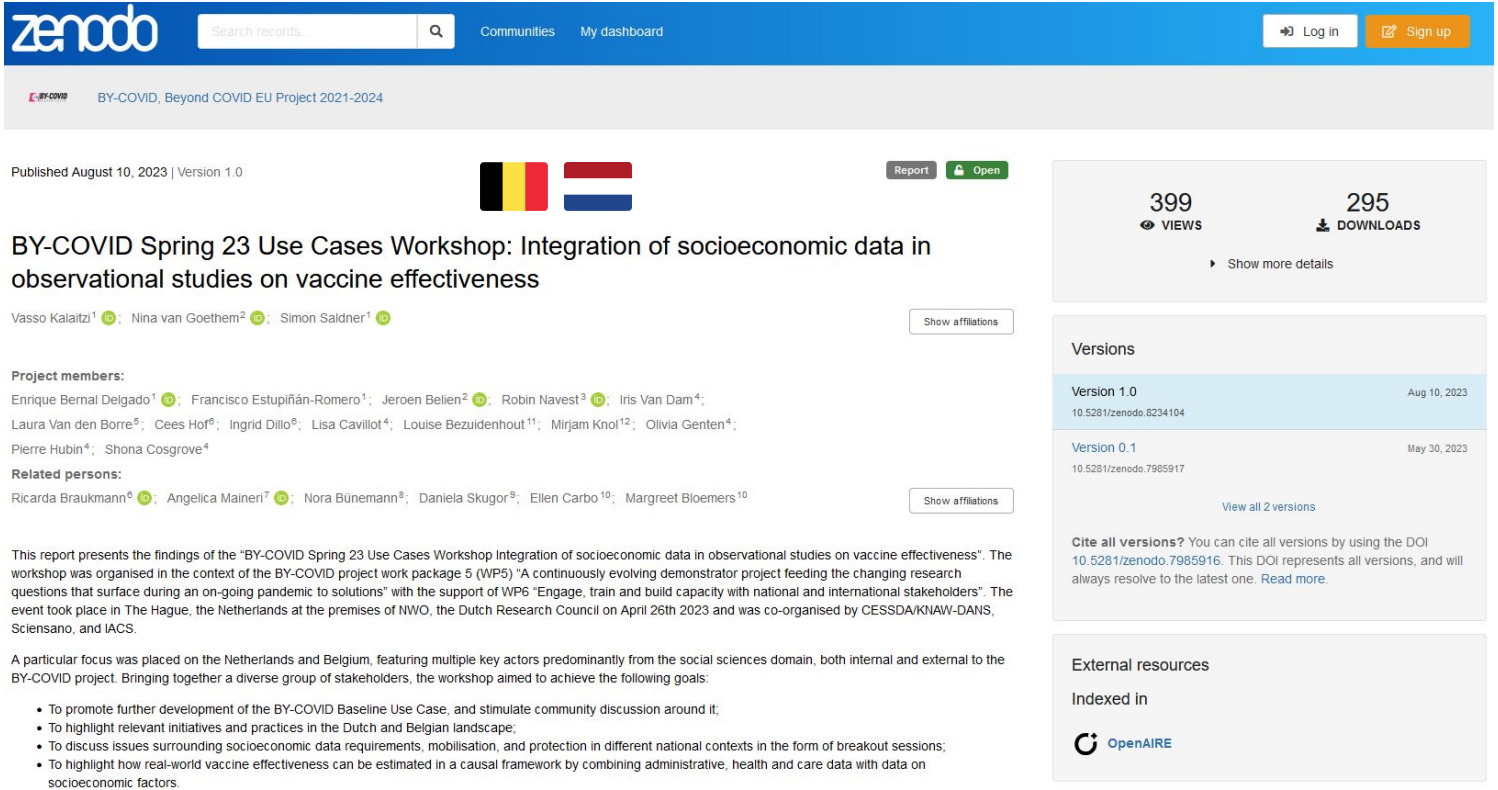


- Combining **individual-level data** from several heterogeneous (real-world) **data sources**, *including socioeconomic data*



- **Across borders** (EU-level)

Previous workshop outcomes



https://zenodo.org/records/8234104

Workshop goals

(1) Identify **barriers and facilitators** related to the identification, linkage, and analysis of individual-level socioeconomic data

(2) Identify **solutions** for integrating individual-level socioeconomic data in population health research

(3) **Generalise** such solutions in various disciplinary and geographical contexts (EU-level focus) and **translate** the workshop findings into an innovative workflow standard to federated population health research

Integrate insights from this Workshop to the report

Overview of the BY-COVID Baseline Use Case

Enrique Bernal-Delgado

09:40 - 09:55

Prototyping a workflow that is standard to population health research

- **Policy relevant** research question: *real-life vaccine effectiveness*
- Seeking to assess **interventions** (e.g., emulate RCT with obs data)
- Mobilising **sensitive routine patient-level data**
  - Sensitive data stays under the jurisdiction and governance of data holders - data visiting principle
  - Data minimisation and purpose limitation relies on the minimum common data model (CDM) required
  - Data application is made by local researchers (DMP); data access is granted according to the local prescriptions
- Linking **data from multiple sources** (1 to 1; 1 to N)
- Using a **federated approach** (strong reliance on LOST interoperability)
  - Organisation follows a common step-wise workflow supported by a computational master/worker topology
  - All the code is implemented in software container acting as secure process environment (SPE), deployed locally
- **FAIR**ification of the workflow
- Compatible with **HealthData@EU** developments

Methodological framework

Meurisse et al.
BMC Medical Research Methodology (2023) 23:248
https://doi.org/10.1186/s12874-023-02068-3

BMC Medical Research Methodology

RESEARCH Open Access

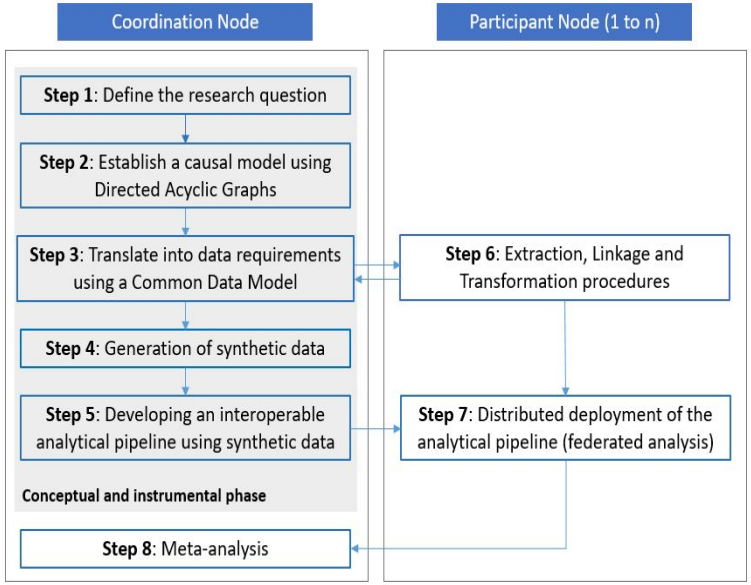
Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment

Marjan Meurisse[1,2*†], Francisco Estupiñán-Romero[3†], Javier González-Galindo[3], Natalia Martínez-Lizaga[3], Santiago Royo-Sierra[3], Simon Saldner[4], Lorenz Dolanski-Aghamanoukjan[5], Alexander Degelsegger-Marquez[5], Stian Soiland-Reyes[6,7], Nina Van Goethem[1†], Enrique Bernal-Delgado[3†] and for the BeYond-COVID project

Abstract

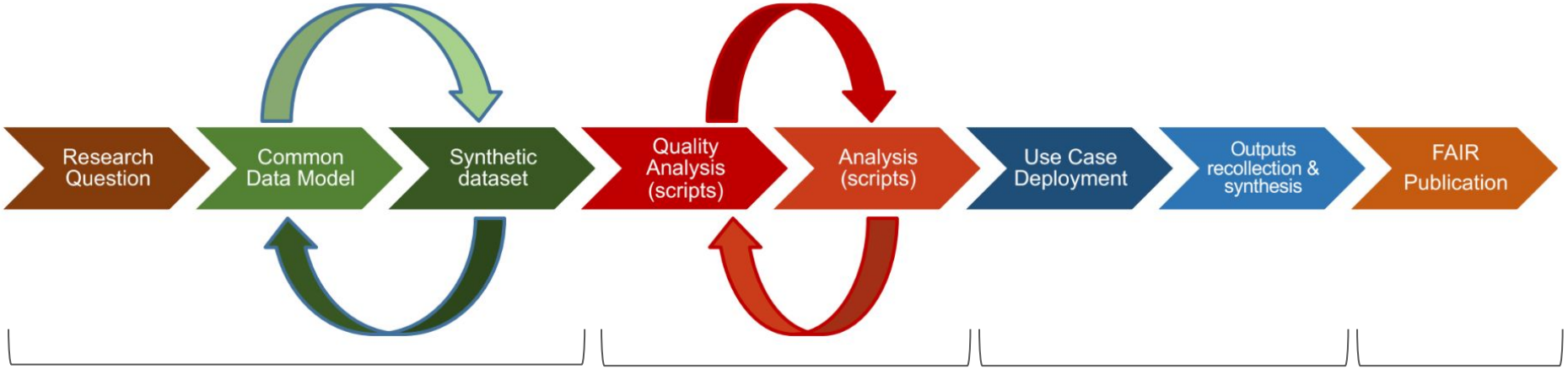
Introduction Causal inference helps researchers and policy-makers to evaluate public health interventions. When comparing interventions or public health programs by leveraging observational sensitive individual-level data from populations crossing jurisdictional borders, a federated approach (as opposed to a pooling data approach) can be used. Approaching causal inference by re-using routinely collected observational data across different regions in a federated manner, is challenging and guidance is currently lacking. With the aim of filling this gap and allowing a rapid response in the case of a next pandemic, a methodological framework to develop studies attempting causal inference using federated cross-national sensitive observational data, is described and showcased within the European BeYond-COVID project.

Methods A framework for approaching federated causal inference by re-using routinely collected observational data across different regions, based on principles of legal, organizational, semantic and technical interoperability, is proposed. The framework includes step-by-step guidance, from defining a research question, to establishing a causal model, identifying and specifying data requirements in a common data model, generating synthetic data, and developing an interoperable and reproducible analytical pipeline for distributed deployment. The conceptual and instrumental phase of the framework was demonstrated and an analytical pipeline implementing federated causal inference was prototyped using open-source software in preparation for the assessment of real-world effectiveness of SARS-CoV-2 primary vaccination in preventing infection in populations spanning different countries, integrating a data quality assessment, imputation of missing values, matching of exposed to unexposed individuals based on confounders identified in the causal model and a survival analysis within the matched population.



https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-02068-3

Prototyping the workflow



CONCEPTUAL PHASE

- Common Data Model specification (v1.1.0) https://doi.org/10.5281/zenodo.7572373
- Study Protocol (v1.0.3) https://doi.org/10.5281/zenodo.7551181

ANALYTICAL PIPELINE

- Reproducible scripts (*continuous update*) https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case

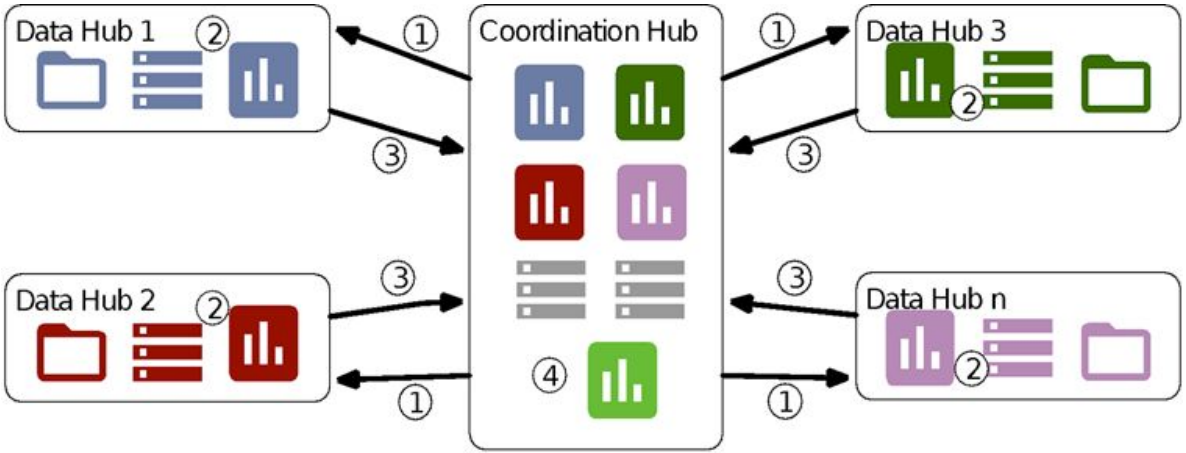
DEPLOYMENT

- Data Management Plan (v1.0.0) https://doi.org/10.5281/zenodo.7625783
- Bilateral meetings with nodes

SHARING

- Zenodo
- Github
- RO-Crate
- WorkflowHub

Code moves; sensitive data stays



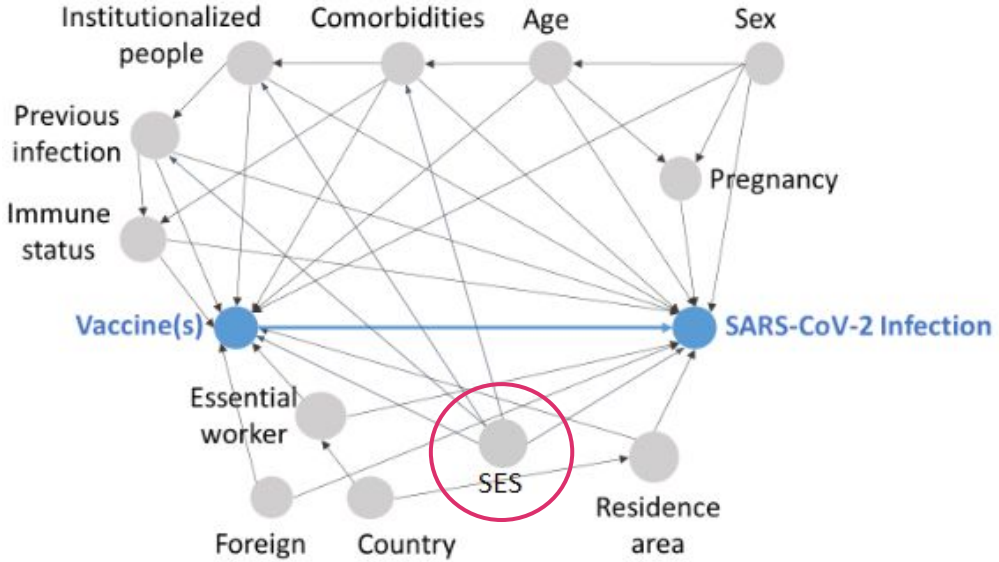
1) Coordination hub orchestrates; also prepares and sends code

2) Data hubs run code, after ETL, and get local outputs

3) Data hubs send back aggregated results

4) Coordination hub compiles and does meta-analysis

Socioeconomic data integration

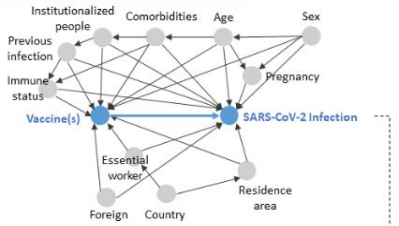


Causal model

Common Data Model (CDM) specification

Synthetic data

Causal model



Common data model

model_metadata

PROJECT
PROJECT URL
DOCUMENT
VERSION (SEM)
AUTHORS
CONTRIBUTORS
DESCRIPTION
CONVENTIONS
VERSION CHANGE LOG

cohort_definition

cohort_name
cohort_description
inclusion_criteria
exclusion_criteria
study_period

model_description

entity	DAG node	variable label	variable description	encoding	variable format	variable type	units	requirement level	variable validation rules	variable property	possible data source(s)
person	Age	age_nm	Age of the person ...	-	integer	Numerical	Years	Required	age_nm >= 5 | age_nm <= 115	Observed	Insurance registry or Health system users database
person	Comorbidities	diabetes_bl	Does the person suffer from type II diabetes ...	crosswalks	logical	Binary	-	Required	Options: TRUE, FALSE	Observed	Electronic Health Records
person	Comorbidities	hypertension_bl	Does the person suffer from hypertension ...	crosswalks	logical	Binary	-	Required	Options: TRUE, FALSE	Observed	Electronic Health Records
Person	Residence area	residence_area_cd	Area of residence of the person (NUTS 3) ...	NUTS 2021 codes	string / character	Categorical	-	Required		Observed	Insurance registry or Health system users database
...											
area	Not applicable	residence_area_cd	Administrative areas (NUTS 3)	NUTS 2021 codes	Character	Categorical	-	Required		Observed	Insurance registry or Health system users database
area	Not applicable	socecon_lvl_area_nm	Ratio of population in the lowest quintile ...		Integer	Numerical	-	Recommended		Observed	Eurostat
...											

detailed_variable_description (def)

entity_name	diabetes_bl		
entity_description	Comorbidities		
entity_definition	Does the person included in the cohort suffer from type II diabetes at the time of entering the cohort?		
classification_system	icd-10-mc	icd-09-mc	snomed-ct
code	E08.00, E08.01, E08.10, ...	249.30, 249.40, 249.50, ...	44054006

Synthetic data

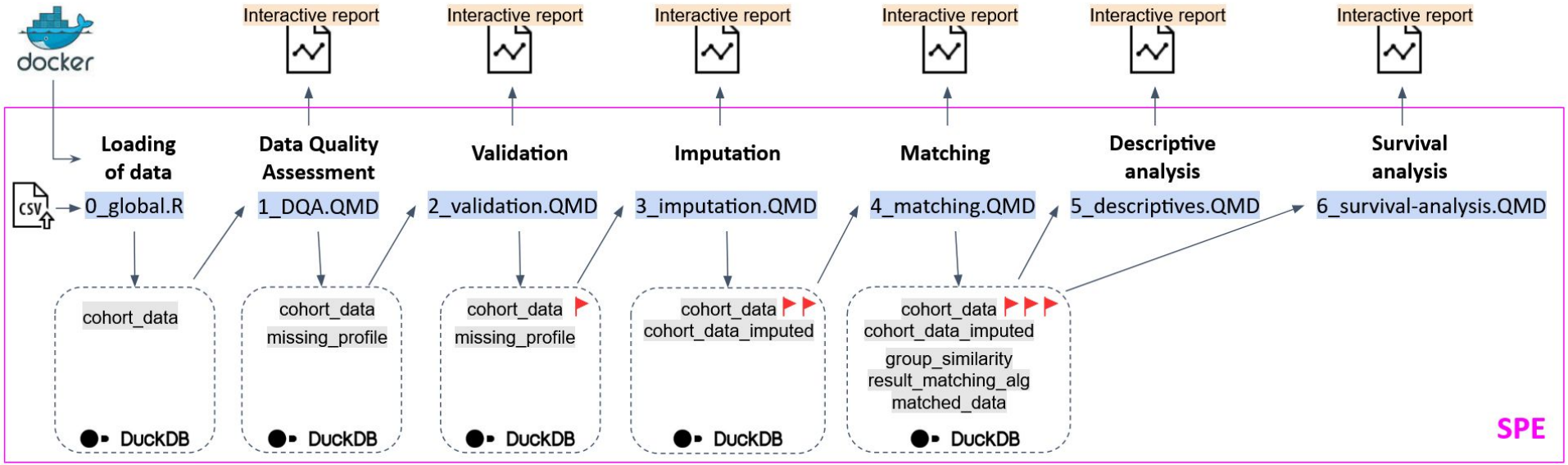
Individual-level data

person_id	age_nm	diabetes_bl	hypertension_bl	...	residence_area_cd
LKxsSbZqAj	47	FALSE	TRUE		BE231
jrpCJeYWJd	89	FALSE	FALSE		BE341
MXJzQXAXsw	56	TRUE	FALSE		BE211
...					

Area-level data

residence_area_cd	socio_econ_lvl_area_nm	...
BE231	0.77	
BE341	0.30	
BE211	1.38	

Analytical pipeline (implementation phase)



Updated cohort_data table including the flag_violation_val

Updated cohort_data table including the flag_violation_val and flag_listwise_del

Updated cohort_data table including the flag_violation_val and flag_listwise_del and flag_inclusion_record

Results analytical pipeline

DATA QUALITY ASSESSMENT

Dataset statistics | Variables | Missing data profile | Alerts | Duplicates

Dataset statistics

Number of variables	42
Number of rows	10 000
Total observation	420 000
Total missing cells	59 599
Missing cells (%)	14.2%
Memory usage	2.9 Mb

VALIDATION RULES

Validation table | Validation plot

Validation rule	Name rule	Items	Passes	Fails	Percentage of fails	Number of NAs	Per c
is.na(age_nm) | age_nm - 5 >= -1e-08 & age_nm - 115 <= 1e-08	V01	10000	10000	0	0%	0	
is.na(sex_cd) | sex_cd %vin% c(0, 1, 2, 9)	V02	10000	10000	0	0%	0	
is.na(dose_1_brand_cd) | dose_1_brand_cd %vin% c("BP", "MD", "JJ", "AZ", "NV")	V03	10000	10000	0	0%	0	

IMPUTATION

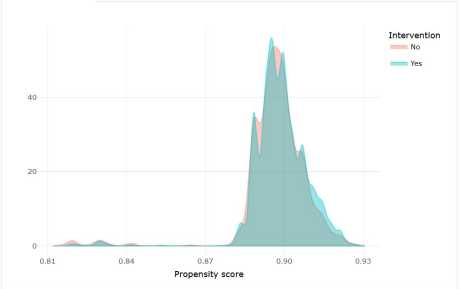
Imputation process | Distribution imputed data

Variable_name	Imputation method	Number of imputed values	Missing_values	Required	Cc
age_nm	No missing values	0	FALSE	TRUE	
blood_cancer_bl	No missing values	0	FALSE	TRUE	
chronic_kidney_disease_bl	No missing values	0	FALSE	TRUE	

MATCHING

Propensity score distribution

Before matching | After matching



DESCRIPTIVE ANALYSIS

Description of the study population: table 1

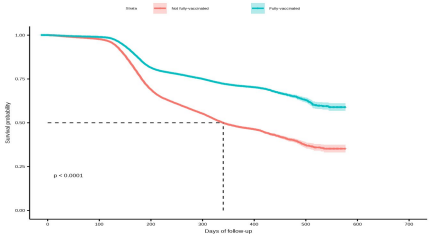
Population eligible for matching | Matched population

Table 1

	Intervention group (Completed a primary vaccination schedule) (N=8378)	Control group (Not completed a primary vaccination schedule) (N=949)	P-value
sex_cd			
Mean (SD)	1.47 (0.537)	1.48 (0.541)	0.356
Median [Min, Max]	1.00 [0, 2.00]	2.00 [0, 2.00]	
age_cd[a]			
Mean (SD)	11.6 (2.96)	11.6 (3.00)	0.914
Median [Min, Max]	12.0 [2.00, 18.0]	12.0 [2.00, 18.0]	
residence_area_cd			
731	1170 (14.0%)	139 (14.6%)	0.574
732	1182 (14.1%)	123 (13.0%)	
733	6026 (71.9%)	687 (72.4%)	

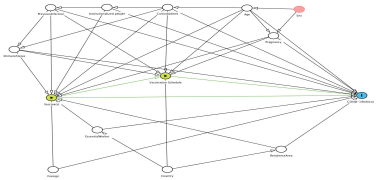
SURVIVAL ANALYSIS

Survival plot



	estimate	coefficient	SE	P-value	HR)
fully_vaccinated_blTRUE	-0.196	0.074	0.095	0.039	0.822 (0.636, 1.008)

Digital objects persisting in Zenodo (OPEN AIRE)



CAUSAL MODEL

Model entity	DAG	Variables	
entity	nodes	variable label	variable description (concept)
person	Vaccine(s) - SARS-CoV-2 Infe	person_id	Pseudoid of the person included in th
person	Age	age_nm	Age of the person included in the coh
person	Sex	sex_cd	Sex of the person included in the coh
person	Socioeconomic level	socecon_M_cd	Socioeconomic level of the person in
person	ResidenceArea	residence_area_cd	Area of residence of the person inclu
person	Country	country_cd	Country of residence of the person in
person	Foreign	foreign_bl	Is the country of residence different f
person	Death	exitus_dt	Date of death of the person (if the per
person	Death	exitus_bl	Date of death of the person (if the per
person	EssentialWorker	essential_worker_bl	Is the person included in the cohort a
person	Institutionalized people	institutionalized_bl	Is the person included in the cohort in
person	Vaccine(s) type	dose_1_brand_cd	Brand of first dose of the vaccine
person	Vaccine(s)	dose_1_dt	Date of the first dose of the vaccine
person	Vaccine(s) type	dose_2_brand_cd	Brand of second dose of the vaccine
person	Vaccine(s)	dose_2_dt	Date of the second dose of the vaccir
person	Vaccine(s) type	dose_3_brand_cd	Brand of third dose of the vaccine

COMMON DATA MODEL SPECIFICATION

vaccines_effectiveness_synthetic_dataset
- by-covid_wp5_baseline_generate_synthetic_data_v.1.0.1.ipynb
- vaccine_effectiveness_synthetic_dataset_eda_v.1.0.1.html
- vaccines_effectiveness_synthetic_dataset_eda_v.1.0.1.json
- vaccines_effectiveness_synthetic_dataset_pop_650k_v.1.0.1.csv

SYNTHETIC DATASET

DATA QUALITY ASSESSMENT

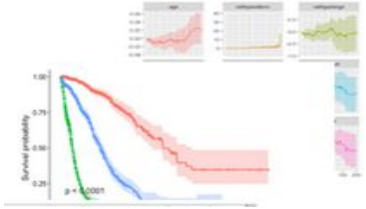
Scripts Including:

Syntactic Conformance
Compliance with data model
Conformance with ETL Rules
Null, missing, outliers values
Density distributions and frequencies
Collinearity

ANALYTICAL PIPELINE

Scripts including:

Propensity score matching for time-variant exposure in large populations
Survival analysis



RESULTS

Including

IPW or B coefficients for full distribution
Aggregated data tables
Html reports

DATA SOURCES

Including

Meta data of the cohorts
Logic data model of the data sources composing the cohort

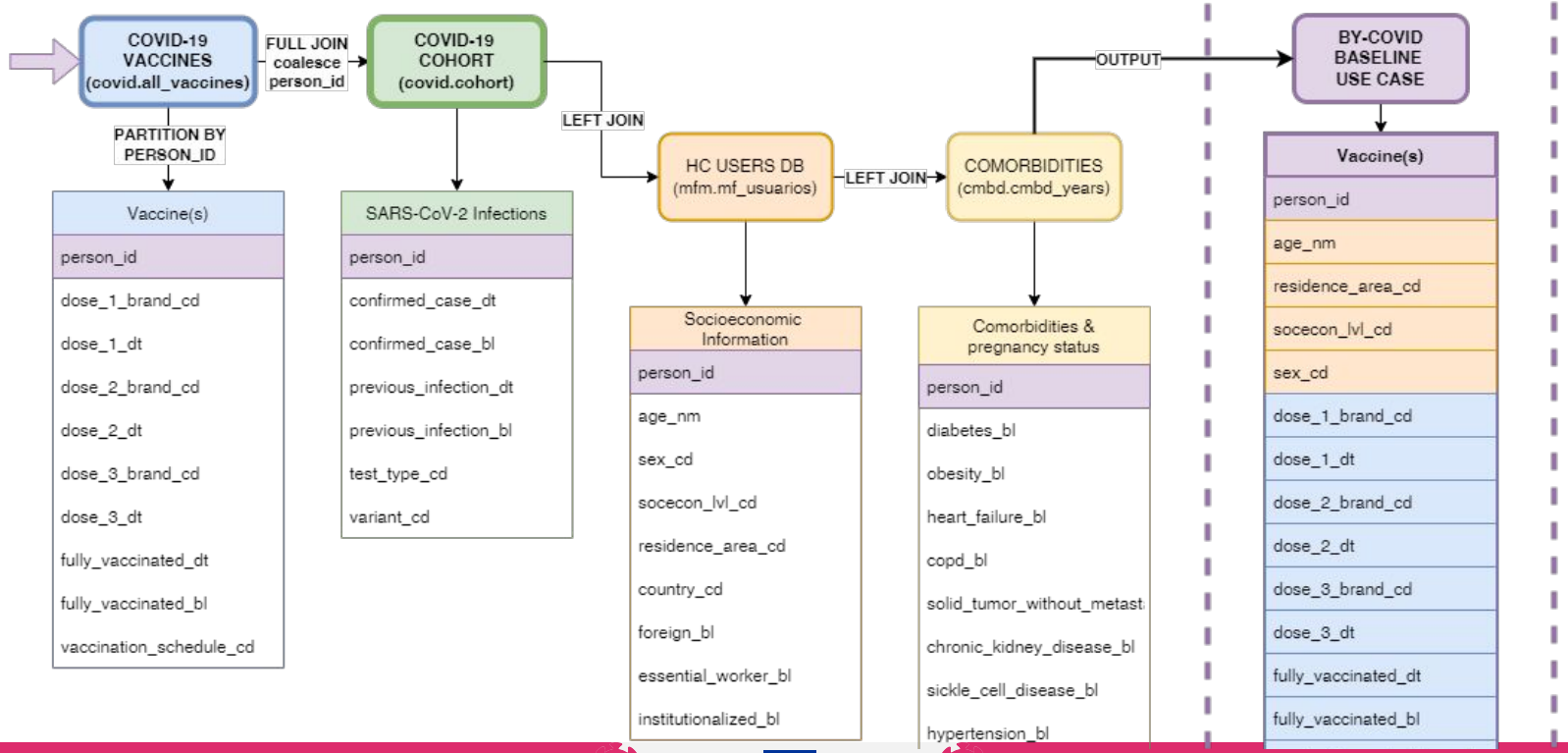
Experience from Baseline Use Case partners: mapping data availability, procedures, and challenges

09:55 - 10:20

Site: Aragon (Spain)

Instituto Aragonés de Ciencias de la Salud

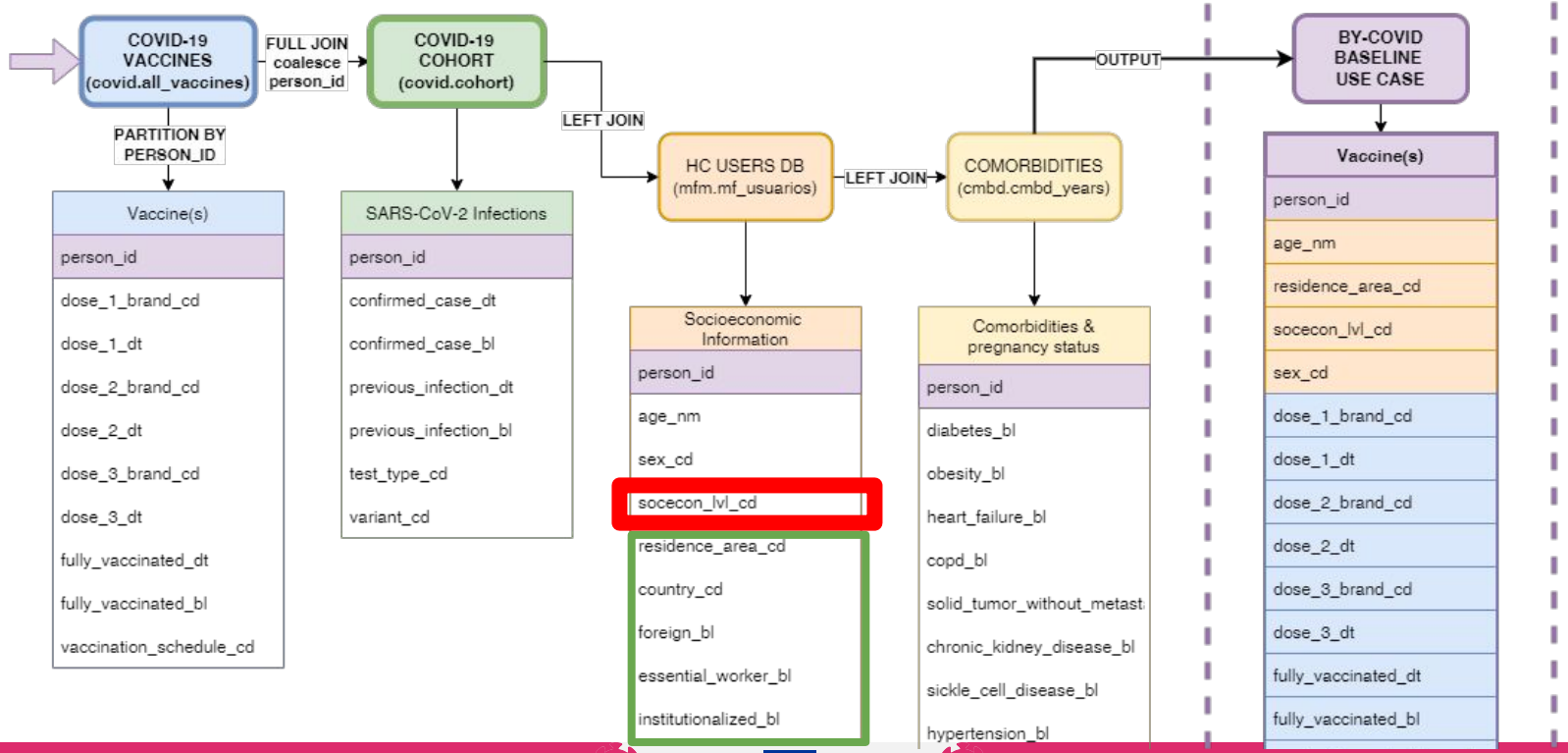
No individual-level (socioeconomic) data available



Site: Aragon (Spain)

Instituto Aragonés de Ciencias de la Salud

No individual-level (socioeconomic) data available

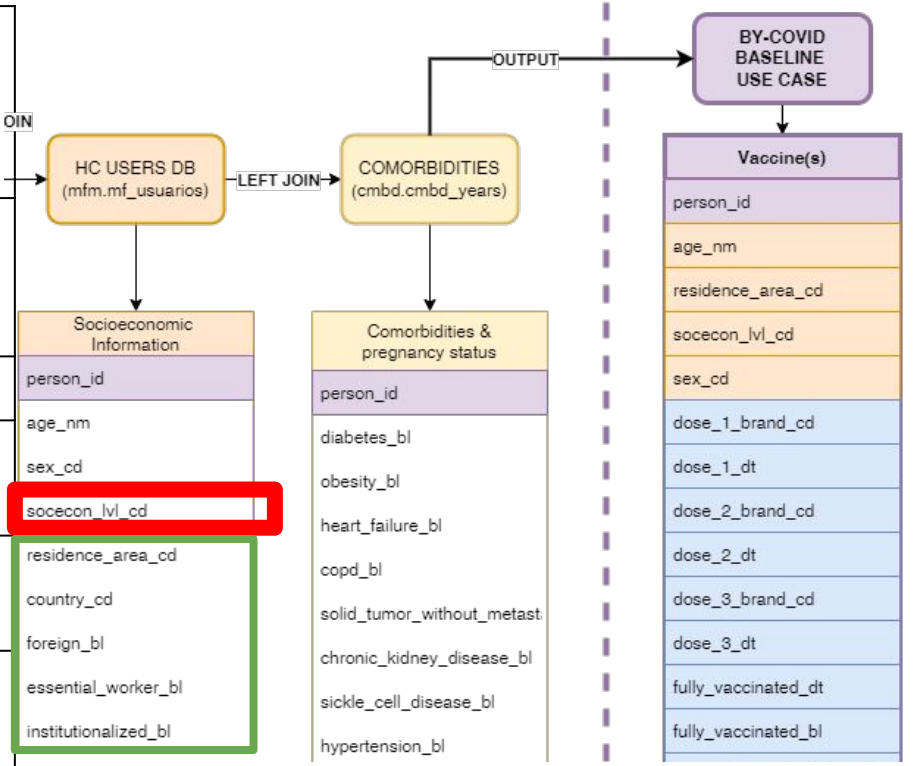


Site: Aragon (Spain)

Instituto Aragonés de Ciencias de la Salud

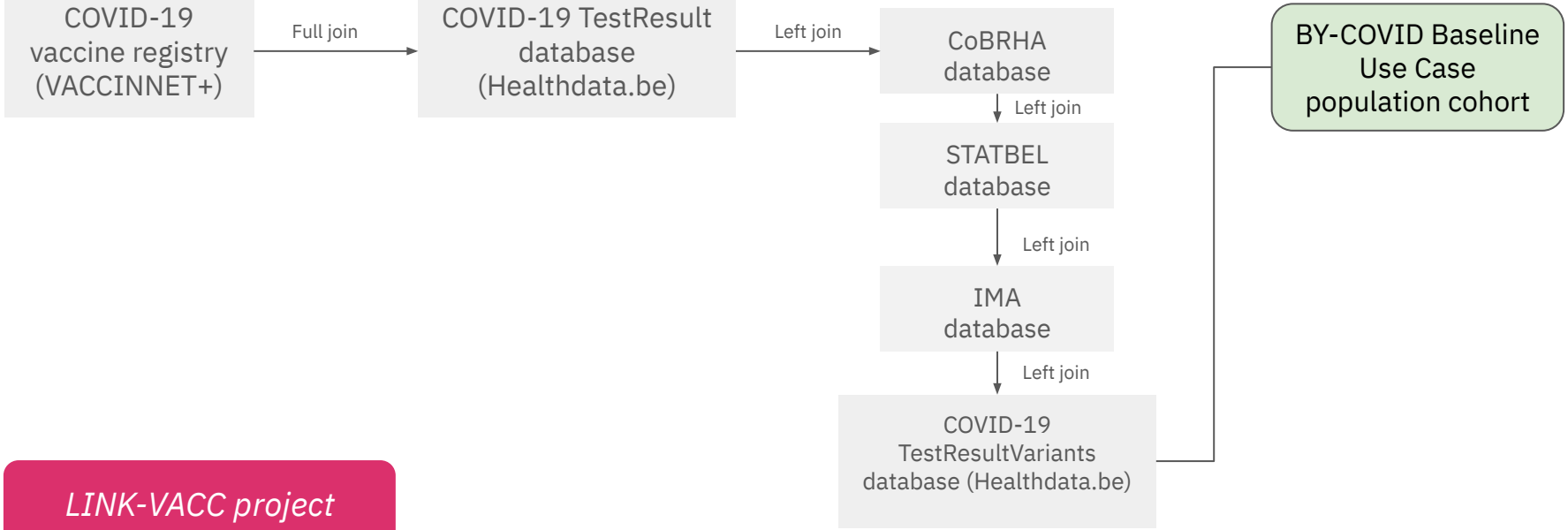
No individual-level (socioeconomic) data available

Data sources	**No individual-level socioeconomic data is available in healthcare** apart from residence area, country of origin, institutionalization or essential worker status. There is a proxy variable based on drug copayment levels depending on individual-level income and labour status. Still, the categories were deemed too broad to be used in the baseline use case.
	There is individual-level socioeconomic administrative and surveys data at the National (or regional) Statistics Institute (INE): Census, Labour force survey, Household budget survey, Survey on income and living conditions.
Data provider(s)	The National (or Regional) Statistics Institute (INE).
Data discoverability	Metadata of the INE dataset for the Survey on Living conditions is included in the **Health Information Portal**. Metadata on all INE datasets is available at INE Open Data Catalogue
Data access specifications	Formal application using a **standardised form**. For each variable requested, motivation must be given as to why the data is needed for the research.
Data discoverability and access barriers	Data access requires formal request from a public entity with proper justification of the need for research and public interest and INE to perform pseudonymization and linkage of the relevant information with the source



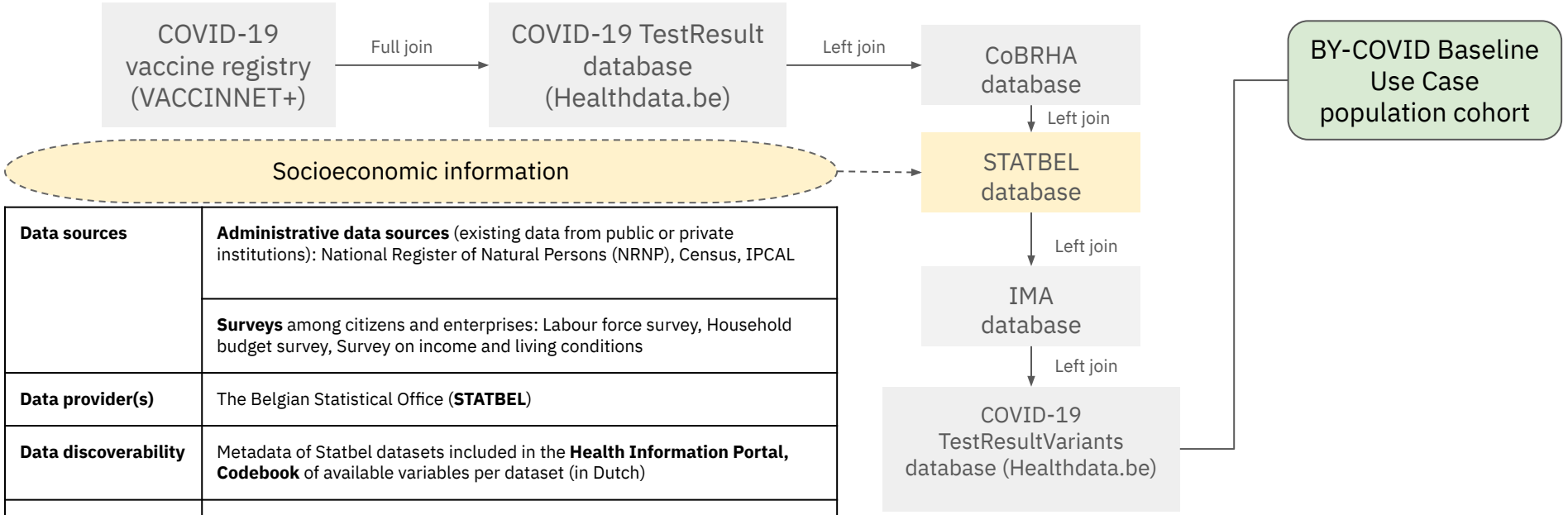
Site: Belgium

Individual-level (socioeconomic) data availability and integration



Site: Belgium

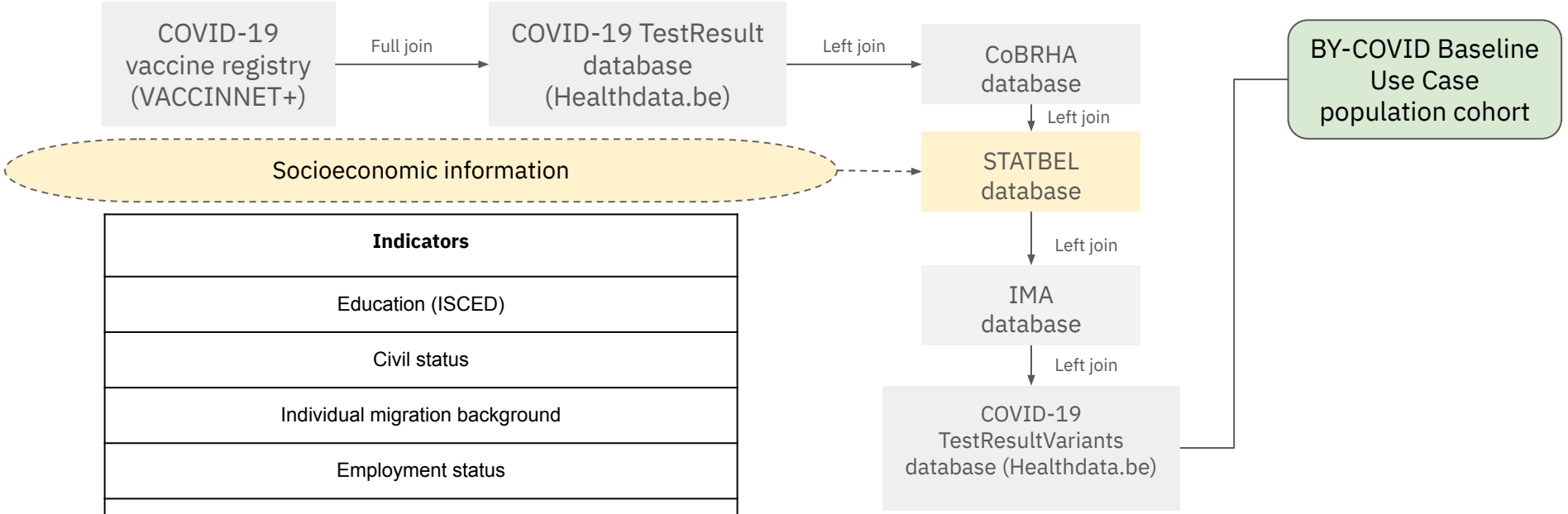
Individual-level (socioeconomic) data availability and integration



Data sources	**Administrative data sources** (existing data from public or private institutions): National Register of Natural Persons (NRNP), Census, IPCAL
	Surveys among citizens and enterprises: Labour force survey, Household budget survey, Survey on income and living conditions
Data provider(s)	The Belgian Statistical Office (**STATBEL**)
Data discoverability	Metadata of Statbel datasets included in the **Health Information Portal, Codebook** of available variables per dataset (in Dutch)
Data access specifications	Formal application using a **standardised form**. For each variable requested, motivation must be given as to why the data is needed for the research.
Data discoverability and access barriers	No clear overview of available datasets and variables

Site: Belgium

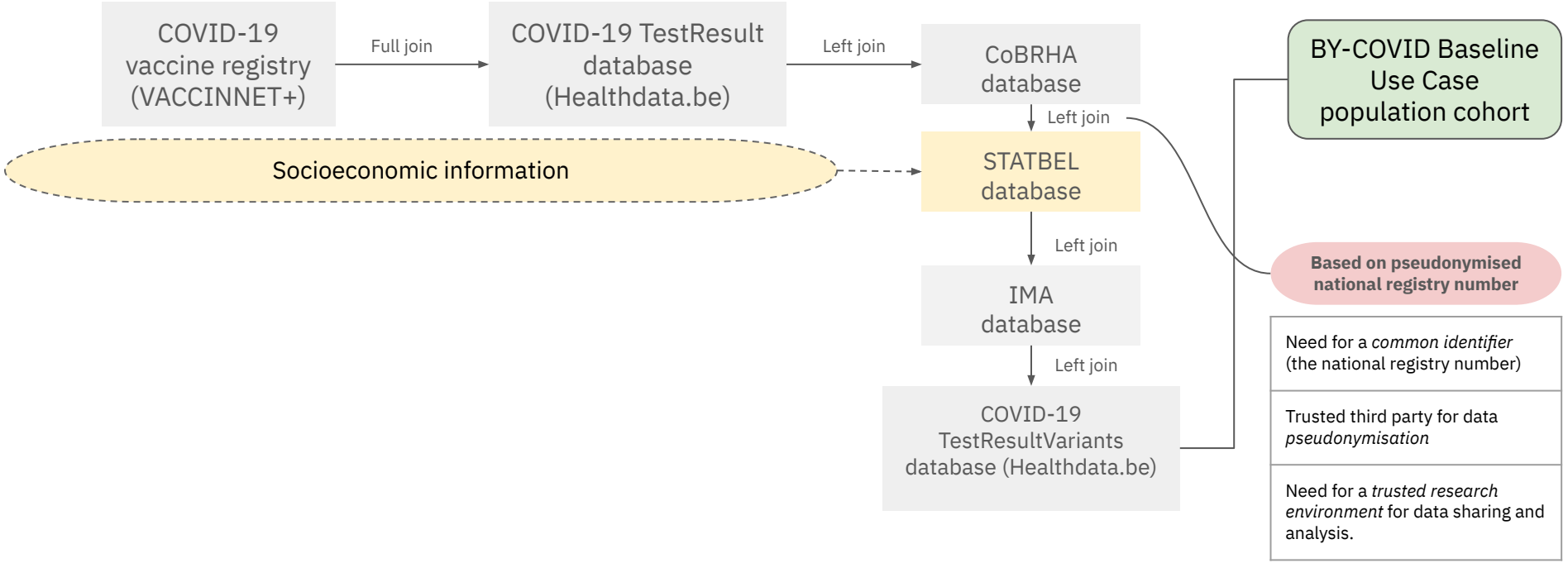
Individual-level (socioeconomic) data availability and integration



Indicators
Education (ISCED)
Civil status
Individual migration background
Employment status
Net taxable income on household level (deciles)

Site: Belgium

Individual-level (socioeconomic) data and integration



Site: Belgium

Real-world effectiveness of primary vaccination in preventing SARS-CoV-infections

Adjustment for Socioeconomic Status (SES), matching (1:1) based on:

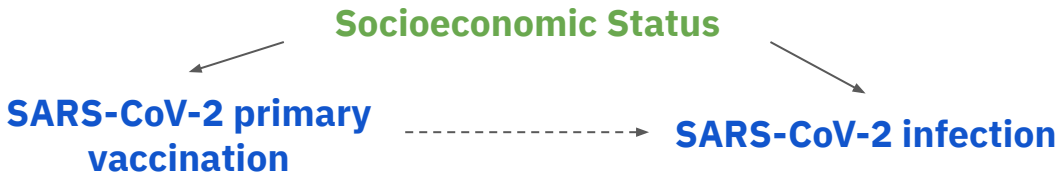
→ Residence area

→ Individual-level SES
Low income (decile 1-4), Middle income (decile 5-7), High income (decile 8-10)

Sensitivity analysis:

→ VE estimate adjusted for all confounder, incl. SES (RMSTD): 59.626 [59.260; 59.991]

→ VE estimate adjusted for all confounder, excl. SESC (RMSTD): 57.734 [57.378; 58.090]



Site: Finland

Individual-level (socioeconomic) data and integration

Data sources used to comply with baseline use case data requirements

- Finnish National Vaccination Register
- Finnish National Infectious Diseases Register (SARS-CoV-2 infection)
- Register of Primary Health Care Visits (SARS-CoV-2 infection, Comorbidities)
- Care Register for Health Care inpatient visits (SARS-CoV-2 infection, Comorbidities)
- Drug Purchase and Reimbursement (Comorbidities)
- Finnish Cancer Registry (Comorbidities)
- Population Register (Baseline data)
- Statistics Finland (Socioeconomic information)

→ Pseudonymised data are kept and linked within the secured ePouta environment

Site: Finland

Individual-level (socioeconomic) data and integration

Availability of **individual-level socioeconomic data**

Data provider	Statistics Finland
Indicator	Occupation
Classification	E-SeC classes: - Higher occupations (E-SeC class 1-3) - Intermediate occupations (E-SeC class 4-6) - Routine and manual occupations (E-SeC class 7-9)