



# Federated causal inference based on real-world observational data sources

## Document Version

Submitted manuscript

[Link to publication record in Manchester Research Explorer](#)

## Citation for published version (APA):

Meurisse, M., Estupiñán-Romero, F., González-Galindo, J., Martínez-Lizaga, N., Royo-Sierra, S., Saldner, S., Dolanski-Aghamanoukjan, L., Degelsegger-Marquez, A., Soiland-Reyes, S., Van Goethem, N., & Bernal-Delgado, E. (2023). Federated causal inference based on real-world observational data sources: Application to a SARS-CoV-2 vaccine effectiveness assessment. Manuscript submitted for publication. <https://s11.no/2023/phd/federated-causal-inference/>

## Published in:

BMC Medical Research Methodology

## Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

## General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

## Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact [uml.scholarlycommunications@manchester.ac.uk](mailto:uml.scholarlycommunications@manchester.ac.uk) providing relevant details, so we can investigate your claim.



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

Marjan Meurisse<sup>1,2\*†</sup>, Francisco Estupiñán-Romero<sup>3†</sup>,  
Javier González-Galindo<sup>3</sup>, Natalia Martínez-Lizaga<sup>3</sup>,  
Santiago Royo-Sierra<sup>3</sup>, Simon Saldner<sup>4</sup>,  
Lorenz Dolanski-Aghamanoukjan<sup>5</sup>,  
Alexander Degelsegger-Marquez<sup>5</sup>, Stian Soiland-Reyes<sup>6,7</sup>,  
Nina Van Goethem<sup>1£</sup>, Enrique Bernal-Delgado<sup>3£</sup>

<sup>1</sup>Department of Epidemiology and Public Health, Sciensano, Brussels, Belgium.

<sup>2</sup>IREC – EPID, Université Catholique de Louvain, Brussels, Belgium.

<sup>3</sup>Data science for Health Services and Policy, Instituto Aragonés de Ciencias de la Salud (IACS), Zaragoza, Spain.

<sup>4</sup>Data Archiving and Networked Services, Royal Netherlands Academy of Arts & Sciences, Amsterdam, The Netherlands.

<sup>5</sup>International Affairs, Policy, Evaluation and Digitalisation, Gesundheit Österreich GmbH (GÖG), Vienna, Austria.

<sup>6</sup>Department of Computer Science, The University of Manchester, Manchester, UK.

<sup>7</sup>Informatics Institute, Universiteit van Amsterdam, Amsterdam, The Netherlands.

\*Corresponding author(s). E-mail(s): [marjan.meurisse@sciensano.be](mailto:marjan.meurisse@sciensano.be);

†These authors contributed equally to this work.

£ These authors contributed equally as senior authors.

## 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.

**Results:** The conceptual and instrumental phase of the proposed methodological framework was successfully demonstrated within the BY-COVID project. Different Findable, Accessible, Interoperable and Reusable (FAIR) research objects were produced, such as a study protocol, a data management plan, a common data model, a synthetic dataset and an interoperable analytical pipeline. **Conclusions:** The framework provides a systematic approach to address federated cross-national policy-relevant causal research questions based on sensitive population, health and care data in a privacy-preserving and interoperable way. The methodology and derived research objects can be re-used and contribute to pandemic preparedness.

## 1 Background

Causal inference, the process of estimating a causal effect of interest (e.g., of a treatment or intervention on a health outcome), is a major interest in public health research. Identifying causal relationships can signal targets for public health policy (e.g., increase exposure to a beneficial determinant or treatment, or reduce exposure to a hazardous one) or allows the evaluation of public health interventions. Estimating causal effects for public health purposes entails the comparison of health outcomes under different treatments or interventions (e.g., comparing the probability of acquiring an infection when vaccinated with the probability of acquiring an infection when not vaccinated).

For inferring causality, randomized controlled trials (RCTs), in which individuals are assigned randomly to one of the intervention groups, are recognized as the “gold standard” [Hernán 2020]. When individuals are randomly assigned to an intervention group, the groups are assumed to be exchangeable or “comparable”, meaning that differences in the outcome can be ascribed solely to the exposure of interest [Hernán 2020] [Greenland 2009]. However, it can be of interest to assess the effect of a treatment or intervention in less controlled real-world settings, considering larger populations, obtained by less restrictive criteria for inclusion, to increase the external validity of the study [Listl 2016].

Further, it is often not ethical or feasible (e.g., because of economic constraints) to perform an RCT. For these purposes, observational studies can be performed, leveraging “real-world” data sources, often obtained through the secondary use of routinely collected health, care and administrative data. When estimating causal effects using observational data, it is essential to consider different potential sources of bias, such as confounding, selection, and information bias, that can appear in natural environments uncontrolled by researchers [Hernán 2020].

The presence of confounders (i.e., variables that influence both the exposure and outcome variable of interest) can result in non-exchangeability of exposure groups, introduce spurious association and, in this way, distort the measured association between exposure and outcome from the causal effect of interest (i.e., differences in the outcome cannot completely be ascribed to the exposure of interest) [Hernán 2020] [Greenland 2009] [Pearce 2016]. Statistical methods, such as confounder adjustment or matching, can be applied to limit confounding bias and to pursue exposure groups that are conditional exchangeable (i.e., comparable) when exchangeability by design (as in an RCT) is not obtained.

These methods generally require the availability of detailed patient information. Alternatively, selection bias represents bias introduced by mechanisms for selecting individuals into the analysis. Selection bias can likewise lead to non-exchangeable exposure groups (i.e., compromised internal validity), as well as impaired generalizability of the study results (i.e., external validity) [Hernán 2020] [Hernán 2004].

Lastly, we refer to information bias as a distortion of the measured association resulting from errors in the measurement or classification of variables, such as the exposure, outcome, or covariates in the analysis. [Hernán 2022] suggested that specifying a hypothetical RCT that would allow the estimation of the causal effect of interest (a target trial) and emulating this target trial using the available observational data is beneficial for maintaining the elements of an RCT. For example, emulating randomization as specified in the target trial during the analysis may help to reduce the risk of confounding and increase the internal validity of the study [Hernán 2020] [Hernán 2006] [Glass 2013].

A treatment or intervention can be applied to populations spanning different regions or countries, with the collected real-world observational data often stored decentralized in isolated environments. Integrating and analyzing these data from different locations and institutions, can support public health decision-making by providing more precise and generalizable estimates. A meta-analysis integrating evidence from different independent studies, for example, as maintained by the International

Vaccine Access Center (IVAC) on the effectiveness of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) vaccine primary series of the online VIEW-hub [IVAC 2023], can be conducted to obtain a pooled effect estimate. However, heterogeneity in the considered confounding factors, criteria for study participant selection, definitions of variables and adopted statistical methods might exist across studies, limiting comparability.

When it is of interest to estimate causal effects and compare interventions or public health programs deployed across (particularly, national) borders, using observational sensitive individual-level data, a federated approach (as opposed to an approach using pooled data) can be used. Such a methodology, implying data visiting, allows to approach causal inference in a privacy-preserving and interoperable way, without sharing sensitive data or gathering them in a centralized location. When conducting federated research, interoperability challenges (i.e., obtaining consistent data from distributed data sources, reproducing an analysis, and comparing the results across the data sources), should be addressed. Different layers of interoperability were defined by the European Interoperability Framework (EIF), namely, Legal, Organizational, Semantic and Technical (LOST) interoperability [EC 2017] [CIPH 2020].

[González-García 2021] presented a methodology and recommendations on how to cope with challenges at the different layers of interoperability when conducting federated research. The current work aims to build upon this pragmatic approach, extending it to a framework amenable to approach causal inference.

Previously, technologies or infrastructures for distributed analysis, such as DataSHIELD [BY-COVID] [Spellman 2018], the Personal Health Train (PHT) [Foster 2017] [Nishikawa-Pacher 2022], and VANTAGE6 [Moncada-Torres 2021], have been proposed. However, to the best of our knowledge, guidance on the full methodological process to approach causal inference, including the specification of data requirements and guaranteeing interoperability when being confronted with a causal research question in federated research, is currently lacking.

The BeYond-COVID (BY-COVID) project (2021-2024) is a Horizon Europe funded project aiming to accelerate access to and linkage of SARS-CoV-2, coronavirus disease 2019 (COVID-19) and patient data, and increase preparedness for future pandemics within Europe [BY-COVID].

The use cases defined within the BY-COVID project are aimed to ensure interoperability across national borders by enabling a federated approach complying with privacy and data protection regulations. This work conceptually describes the proposed methodology and prepares its application to a policy-relevant research question (i.e., investigating the real-world effectiveness of the SARS-CoV-2 primary vaccination program in populations spanning different countries), aiming to facilitate a rapid response in the case of a next pandemic.

## 2 Methods

A methodological framework for federated causal inference research by re-using routinely collected observational data across different regions, was constructed based on the principles of interoperability at Legal (i.e., privacy-by-design), Organizational

(i.e., analysis coordination), Semantic (i.e., built upon a common data model) and Technical (i.e., via the distribution of analyses and a reproducible environment) level [CIPH 2020], Open Science (i.e., transparent and accessible processes and knowledge) [Spellman 2018] [Foster 2017] and international cooperation driven by population-level research questions.

The framework expands methodologies to leverage population health data for federated policy-oriented research proposed within the “Information for Action” Joint Action (JA-Infact) [González-García 2021] [Abboud 2018] and Population Health Information Research Infrastructure (PHIRI) project [Bogaert 2021], allowing it to address causal research questions, through applying existing methodologies (e.g., the use of DAGs) and building on literature, experience and expertise

## 2.1 The methodological framework

The developed methodological framework is described in this section. The framework comprises guidelines in the form of the following steps:

1. defining the research question
2. establishing a causal model using Directed Acyclic Graphs (DAGs)
3. translating the causal model into data requirements using a Common Data Model (CDM)
4. generating synthetic data, supporting script development and testing
5. developing an interoperable analytical pipeline using synthetic data
6. extracting, linking, and transforming individual-level data within each node to comply with the CDM specification and information requirements
7. distributed deployment of the analytical pipeline (i.e., federated analysis), and
8. meta-analysis of the local results (see Figure 1).

Step 1 to 5 are part of a ‘conceptual and instrumental phase’ within the framework and can be conducted without access to real-world data, while steps 6 and 7 involve the extraction, transformation and analysis of real-world data within the jurisdiction of each of the participants to reach step 8 and produce comparable results to inform policy. Going through the steps of the conceptual and instrumental phase of the framework requires profound knowledge about real-world data.

The proposed framework requires close collaboration between a coordinating research team (also referred to as the ‘Coordination Node’) and institutions hosting or being able to acquire access to the required sensitive individual-level data (also referred to as ‘Participant Nodes’), to guarantee organizational interoperability (see Figure 1). The Coordination Node is responsible for leading the entire process, promoting the collaboration of the participants in the conceptual and instrumental phase of the framework and producing the documentation (i.e., research objects such as the CDM, the synthetic dataset, the analysis scripts of the interoperable analytical pipeline, etc.) supporting the data linkage, preparation and the deployment of the analytical pipeline. In the sections below, a detailed description of each step is provided.

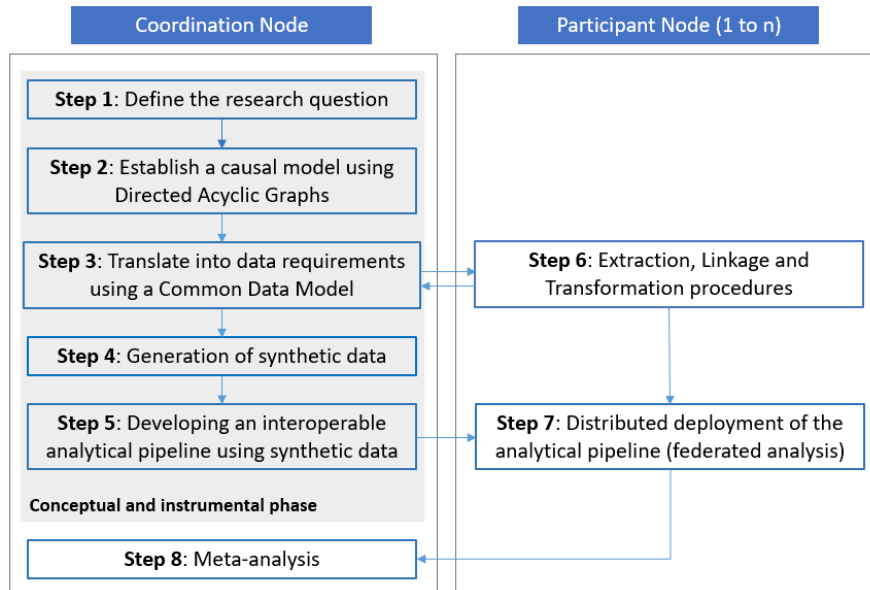


Fig. 1 Visual representation of the proposed methodological framework

### ***Step 1: Defining the research question***

To start, it is essential to precisely define the research question that the study aims to answer. To guide the formulation of the research question, one can follow the PICO(T) strategy frequently used in clinical research, identifying (1) the patient, problem, population, or cohort of study, (2) the intervention or exposure, (3) the comparison or control, (4) the outcome(s) and, optionally (5) the time frame [Nishikawa-Pacher 2022] [Riva 2012] [Lira 2019]. The research question directly feeds into a causal model, study design and information requirements.

### ***Step 2: Establishing a causal model using Directed Acyclic Graphs (DAGs)***

To estimate the causal effect of interest for the defined exposure-outcome relationship, exchangeable exposure groups should be pursued, thereby emulating randomization of the exposure as in an RCT [Hernán 2016]. Approaching exchangeability requires the adjustment for all known factors confounding the relationship between exposure and outcome.

For the identification of these factors, causal models, such as graphical DAGs, can be used as an instrument to collaborate and map conditions to advance towards causal inference. DAGs provide a clear graphical way to identify confounding bias and other potential sources of bias under the described assumptions, and present a way to determine the smallest set of variables to condition on to draw up to a causal association (a ‘minimal sufficient adjustment set’) by using the ‘backdoor criterion’ (i.e., the criterion holds for a set of variables, if all backdoor paths between the exposure and outcome are closed by conditioning on these variables and if none of the variables is

a descendent of the exposure) [Hernán 2020]. They map the knowledge and assumptions of researchers about the causal relationship between the exposure and outcome, and give an explicit view on the assumed relationships [Staplin 2017] [Suzuki 2020] [Tennant 2021].

This way, DAGs increase transparency and facilitate discussion between researchers. The DAGitty web application or corresponding R package *dagitty* can be used to construct and analyze DAGs [Textor 2016]. Guidance on the construction of DAGs and the identification of a minimal sufficient adjustment set can be found elsewhere [Suzuki 2020] [Tennant 2021] [Digitale 2022]. The assumption of ‘no unmeasured confounding’ to identify the causal effect of interest demands appropriately measured confounders and correct statistical inclusion. It is important to consider the possibility of unmeasured or omitted confounding (e.g., due to limitations of surveillance systems or the human understanding of causal relationships respectively) [Kasza 2017].

### ***Step 3: Translating the causal model into data requirements***

Once agreed upon, a causal model should translate the research question into data requirements detailing syntactic and semantic considerations to achieve interoperability and enable sound comparability between the Participant Nodes within the federation. These data requirements are captured in a CDM.

A customizable template for building a CDM is available in Additional File 1. All nodes in the DAG (e.g., variables measuring the exposure, outcome and the minimal sufficient adjustment set, as well as variables required to achieve secondary objectives of the study or to perform supplementary or exploratory analyses) should be captured within the model description of the CDM, irrespective of their inclusion in the minimal sufficient adjustment set. Variable labeling must be consistent and follow a pre-specified convention (e.g., Snake Case, Camel Case) [Broberg 2019] [DiLep 2019]. Variable labeling must not hinder the analysis, and therefore the variables should not start with reserved characters or numbers. Furthermore, variable labeling must include information on variable type for easier identification while interactively exploring the data. For example, the following convention can be used: **cd** for categorical variables, **nm** for numerical variables, **bl** for binary/logical variables, **dt** for date variables, and **id** for the primary (and secondary) key of the entity.

Each of the variables should be characterized in a detailed manner, including:

1. the model entity,
2. the variable label
3. a description of the variable
4. the encoding system
5. the variable format and type
6. the units of measurement
7. the requirement level
8. the variable-level validation rules
9. the transformations at origin
10. the variable properties (observed/calculated)
11. the possible data sources.



The variable format can be expressed differently depending on the data types enabled in each scripting language, however, can commonly be defined as `integer`, `double` or `float` for a number, `string` or `character` for an alphanumeric, `logical` or `binary` for TRUE or FALSE and `date` or `timestamp` for a date.

The requirement level (i.e., required, recommended or optional) denotes the impact of complete absence of information on that variable on achieving the purposes of the study. In studies aiming to approach causal inference, the required variables in the CDM should correspond to those measuring exposure, outcome and the minimal sufficient adjustment set required to close all backdoor paths identified in the DAG. As such, not having any information (complete missingness) on a variable in the minimal sufficient adjustment set, impedes reaching the study objectives by introducing bias and hindering causal interpretation of obtained estimates.

Depending on the context and planned analyses, a variable considered to be required can be allowed to have a certain degree of values missing. Complete missingness of recommended variables could harm the secondary objectives of a study (i.e., planned sensitivity, subgroup analyses, or similar), while complete missingness of optional variables might impede supplementary or explorative analyses. Specifying possible data source(s) and comments are out of the scope of the variable description, but can offer additional information to facilitate the extraction, linkage and transformation procedures, and management of the data at origin during step 6.

If different entities (e.g., person, area, test, vaccination dose) are needed to cover the requirements captured in the DAG, a model description per entity should be provided. Further, a variable capturing information on a certain (co)morbidity might demand the specification of crosswalks (i.e., mapping to different classification systems) to ensure the coverage of the definition within different Participant Nodes using different disease classification systems at origin (i.e., semantic interoperability). The data model specification should additionally contain an unambiguous cohort description, including the specification of eligibility criteria of the study population and the start and end date of the study period. Further, in order to make the data model discoverable for other researchers, a structured metadata file should be provided.

#### ***Step 4: Generation of simulated synthetic data***

The generation of synthetic data, representing the specifications from the CDM, can be instrumental to develop the interoperable analytical scripts and can serve to exemplify the required data for the federated analysis. Synthetic data can be simulated by simply capturing the technical and syntactic requirements as specified in the CDM and using non-informative mathematical distributions, thereby avoiding exposure of the real sensitive data during the conceptual and instrumental phase of the framework and promoting the development and testing of the analytical scripts while managing the data access application process. Nonetheless, simulated data can be enhanced with expert information on the topic to reflect the expected distributions of the actual data based on published healthcare statistics or prior research. Alternatively, when access to real data is possible and a sufficient degree of anonymization can be assured, synthetic data can be modelled based on the real data (i.e., data driven), preserving its

underlying distributions, relationships and statistical properties with the specifications defined in the CDM.

#### ***Step 5: Developing an interoperable analytical pipeline***

Once the data requirements are specified and a synthetic dataset is generated, an analytical pipeline for distributed deployment can be developed. The analytics are dependent on the specified research question and can apply different methods to address biases (e.g., adjusting for identified confounders [Hernán 2020] [Li 2014], controlling for selection bias [Bareinboim 2012]) and handle missing data [Kang 2013] [Haukoos 2007] [Little 2019]. Further, there are various ways to investigate the presence of biases in the results, such as selection and unmeasured or omitted confounder bias, and assess the sensitivity of the results to the applied methods and assumptions in different sensitivity analyses. Nonetheless, irrespective of these specific analytical methods, certain elements common to any federated study should be contained within the pipeline. The first step in the analytical pipeline consists of a comprehensive Data Quality Assessment (DQA), including information on the completeness, uniqueness, and integrity [Cai 2015].

Next, compliance with the CDM specification should be checked, by testing the input data against a set of data validation rules. Further, descriptive statistics can be produced, providing characteristics of the study population. Population characteristics can be used to improve interpretation of the results and detect potential biases, along with the results from the DQA and validation procedure.

Finally, as federated research relies on the distribution of scripts for the analyses and the local deployment and execution of the analyses at each participant’s system, it requires extensive documentation of all functionality and implemented decisions during the development of outputs of the analytical pipeline. All this documentation is required for interpretation of the local outputs, which are later used in the meta-analysis. The analytical pipeline should only produce aggregated results that have lost all sensitive properties, i.e., compliant with disclosure policies.

#### ***Step 6: Extraction, linkage and transformation procedures within the Participant Nodes***

We defined ‘Participant Nodes’ as institutions contributing to the investigation of the research question, hosting or being able to acquire access to individual-level real-world population, health and care data. Each Participant Node is responsible for the data access application process, requesting access to analyze the required data. When access to the data necessary for the research in question is granted, linkage of different data sources needed to comply with the specified data requirements should be performed by the data controllers (i.e., can be the Participant Node or another institution).

The Participant Nodes are responsible for processing the data following the guidelines provided by the CDM specification, in this way preparing the data for the analysis. Perfect adherence to the CDM specification cannot always be achieved with the available data, however should be pursued, particularly for the cohort selection criteria, the syntactic model and the required variables.

### ***Step 7: Distributed deployment of the analytical pipeline***

The interoperable analytical pipeline should subsequently be distributed and deployed within a secured processing environment of each Participant Node. It requires as input the linked and transformed data complying with the CDM specification. Adherence of these input data to the CDM should be informed throughout the analytical pipeline through informative errors (i.e., in the event that the input file format is not as expected, or the input file header does not correspond to the expected variables' names and order), and through the output of the DQA and the validation assessment.

The analytical pipeline can be provided as single or multiple scripts implementing the statistical analysis using auditable open-source software or can be containerized (e.g., using a Docker container [Bashari Rad 2017] [Boettiger 2015]), providing a fixed environment dealing with system and software dependencies, thus ensuring reproducibility by providing a sandbox that can be deployed and run isolated from the Participant Node's systems [González-García 2021] [Piccolo 2016]. Containerization also enables easy pipeline distribution as container images can be published in an open repository facilitating versioning and collaborative improvement. Technologies offered by the Personal Health Train (PHT) [Attema 2021] [Beyan 2020], DataSHIELD [Wolfson 2010] [Gaye 2014], and VANTAGE6 [Moncada-Torres 2021] can alternatively provide a solution to distribute analysis code to different Participant Nodes.

### ***Step 8: Meta-analysis of the local results***

To integrate results across different populations, the aggregated non-sensitive statistics produced as local outputs of the analytical pipeline should be pooled by the Coordination Node and a meta-analysis should be performed. By only sharing non-sensitive aggregated results, compliance with General Data Protection Regulation (GDPR) legislation and legal interoperability is ensured. The type of aggregated statistics (e.g., propensity scores, standardized risks, average treatment effects) that are shared and pooled, and the methodology used to integrate these estimates, will depend on the defined research question and should be detailed in the relevant research object (e.g., Statistical Analysis Plan).

As indicated previously, some fixed outputs common to any federated study (e.g., documentation of the functionality and implemented decisions during the development of outputs, results from a DQA, validation assessment and descriptive analysis) should also be collected by the Coordination Node, thereby improving the interpretation of the results. In addition to the main results, results from several sensitivity analyses, investigating the presence of biases and sensitivity of the results to certain methods, should be shared with the Coordination Node.

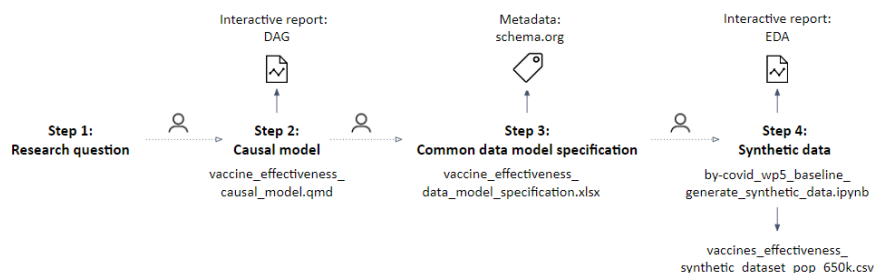
## **2.2 An illustrative example**

The conceptual and instrumental phase (steps 1 to 5) of the proposed methodological framework for federated causal inference (re-)using observational data sources was demonstrated within the BY-COVID project by prototyping a workflow which can be used to assess the real-world effectiveness of SARS-CoV-2 primary vaccination as compared to partial or no vaccination in preventing SARS-CoV-2 infection in populations spanning different countries.

In the current manuscript, we showcase the different steps of this phase of the methodological framework, preparing for the subsequent implementation of the proposed methodology to respond to a policy-relevant research question. For developing the workflow, only open-source software, such as DAGitty [Textor 2016], R and DuckDB [Raasveldt 2019], was used.

### 3 Results

Here, we showcase the conceptual and instrumental phase of the proposed methodological approach as established within the BY-COVID project’s use case. Steps 1 to 4 and the related research objects are presented in Figure 2.



**Fig. 2** Overview of the executed steps and produced research objects. Shown during the implementation of the proposed methodological approach, step 1 to 4, preparing for the assessment of the real-world effectiveness of a primary vaccination schedule as compared to partial or no vaccination in preventing SARS-CoV-2 infection, in populations spanning national borders.

To start, a research question was defined following the PICOT strategy (*step 1*), namely, we aim to assess the real-world effectiveness of a primary vaccination schedule as compared to partial or no vaccination in preventing SARS-CoV-2 infection, in populations spanning national borders [Meurisse 2023a], with the purpose of evaluating the effectiveness of the basic vaccination campaign.

Individuals (age 5 to 115, resident of the participating country) vaccinated with at least one dose of the SARS-CoV-2 vaccine (any of the available brands), or eligible to be vaccinated and with a documented positive diagnosis (irrespective of the type of test) for a SARS-CoV-2 infection during the data extraction period, are eligible for inclusion. Individuals with a documented confirmed infection before completing the primary vaccination schedule (i.e., enrolment), or before January 1, 2021 (SARS-CoV-2 vaccine roll-out) for those not having completed a primary vaccination schedule (controls), will be excluded from the study population.

A DAG corresponding to the research question was produced and a ‘minimal sufficient adjustment’ set was identified using the DAGitty web application [Textor 2016] (*step 2*).

Nodes and edges within the DAG were defined as assumptions based on relationships described in the literature. Once an initial DAG was drafted, field experts participating in the BY-COVID project were invited to discuss and adapt the captured

assumptions where needed. The following DAG nodes present a minimal sufficient adjustment set, conditional on the assumptions that were made: Age, Comorbidities, Country, Essential worker, Foreign, Immune status, Institutionalized people, Pregnancy, Previous infection, Residence area and Sex.

A Quarto notebook was developed (see file `vaccine_effectiveness_causal_model_v.1.1.0.qmd` or a later version as available in the Zenodo publication [Estupiñán-Romero 2023]), generating an interactive report that visualizes the DAG together with information on the research project and the identified minimal sufficient adjustment set. Figure 3, displaying the constructed DAG, illustrates how the digital objects that are consecutively produced during the conceptual and instrumental phase of the proposed methodological approach relate to each other.

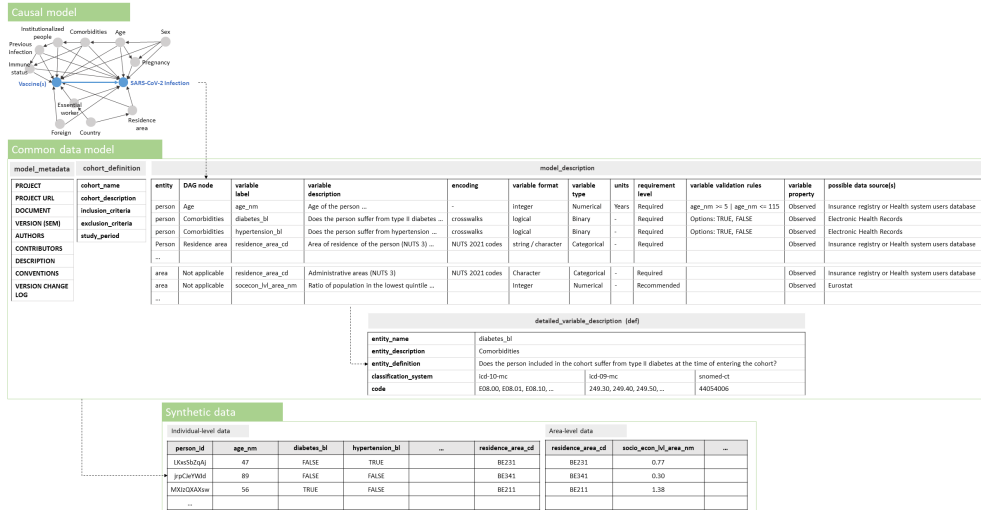
The DAG was translated into data requirements using a CDM (step 3), operationalizing all the nodes in the DAG. Additional individual- and area-level variables were specified to achieve secondary objectives of the study and to perform supplementary or exploratory analysis (e.g., variables `variant_cd` or `socecon_lvl_area_nm`). Variables were labeled following the Snake Case naming convention.

The CDM consisted of an Excel file (see file `vaccine_effectiveness_data_model_specification_v.1.1.0.xlsx` or a later version as available in the Zenodo publication [Estupiñán-Romero 2023]) including a tab with a cohort description, a tab with the model description (characterization of variables), and tabs with a detailed description of certain variables (e.g., comorbidities requiring crosswalks). International classification systems were used when specifying the required encoding of variables and when specifying crosswalks (e.g., IDC-10, IDC-9, and SNOMED-CT for classifying comorbidities).

The structure of the CDM is presented within Figure 3. Compliance to the requirements captured in this CDM can be achieved by performing a full join of the registered individuals in the COVID-19 cases and vaccination datasets and the individual-level linkage to additional data sources, such as patient administrative information (e.g., from insurance registries, health system users-databases, and mortality registration data) and information on patient comorbidities (e.g., from Electronic Health Records) within the Participant Nodes (which is foreseen in *step 6*). A metadata description of the CDM was provided using the Schema.org vocabulary.

A synthetic dataset (see file `vaccines_effectiveness_synthetic_dataset_pop_650k_v.1.1.0.csv` or a later version as available in the Zenodo publication [Estupiñán-Romero 2023]) was subsequently generated (*step 4*), translating the CDM specification into a Python script parameterized to simulate data, considering several known population-level parameters for the COVID epidemic waves (see file `by-covid_wp5_baseline_generate_synthetic_data_v.1.1.0.ipynb` or a later version as available in the Zenodo publication [Estupiñán-Romero 2023]). Within this Python script we made use of the Python package *Faker* [Faraglia 2023]

An exploratory data analysis (EDA) was performed on the synthetic data, exploring different features of the data (i.e., type inference, alerts, uniqueness, outlier values, missing data, univariate analysis) to assess its compliance with the CDM (see file `vaccine_effectiveness_synthetic_dataset_eda_v.1.1.0.html` or a later version



**Fig. 3** The causal model (using a DAG), Common Data Model (CDM) and synthetic data, and how they relate to each other. The DAG, capturing assumptions on factors and relationships when assessing the real-world effectiveness of a primary vaccination schedule as compared to partial or no vaccination in preventing SARS-CoV-2 infection in populations spanning national borders, is visualized. The structure of the CDM and synthetic data, as constructed based on the drafted causal model, is presented.

as available in the Zenodo publication [Estupiñán-Romero 2023]). Based on the EDA, we observe that the generated synthetic data correctly capture the syntactic and technical specifications provided by the CDM. Particularly, the variables in the synthetic data, their labels, encoding, format and type match those specified in the CDM. Variables corresponding to nodes in the minimal sufficient adjustment set have no (e.g., `age_nm` and `sex_cd`) or a limited proportion of (e.g., `residence_area_cd` with 2% missing) missing values.

Subsequently, an analytical pipeline was developed and tested with the support of the synthetic data (*step 5*) using the R statistical programming language as sequential Quarto documents (.qmd files) reflecting and reporting the outputs of different modules:

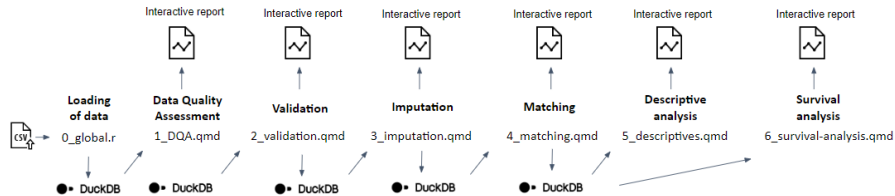
1. DQA of the original input data
2. validation (i.e., applying logic validation rules) of the original input data to check compliance with the CDM
3. imputation of missing data where required
4. iterative matching of the exposed to unexposed individuals and a balance assessment of the matched population
5. a descriptive analysis of the matched and unmatched study population
6. a survival analysis in the matched study population (see the GitHub repository for methodological details [Meurisse 2023b]).

A graphical overview of the analytical pipeline is presented in Figure 4. Each module of the analysis produces an interactive report, including documentation allowing to trace back to decisions made along the way and interpreting the results.

The DQA script (`1_DQA.qmd`) embedded in the analytical pipeline was roughly inspired by the data profiling produced by the `ydata-profiling` Python library [YData 2023a][YData 2023b], usually considered an industry standard for Exploratory Data Analysis in Python, and output contains descriptive dataset statistics and a basic profile of the dataset as a whole (among others containing a variable count, a row count, a basic missing data profile, and some alerts regarding the cardinality, missingness or anomaly of certain variables). The DQA also contains a univariate descriptive analysis of each variable in the dataset, providing summary statistics, information on the categories of the categorical variables, and basic information on the distribution of the continuous variables. Validation of the data, i.e., checking compliance of the data to validation rules captured in the CDM and exclusion of non-compliant data for further analysis, is captured within the script `2_validation.qmd`.

An algorithm capturing decisions on how to deal with missing values in the imported dataset was developed in script `3_imputation.qmd`, implementing the imputation of values, listwise deletion or exclusion of matching variables depending on the characteristics of the data. The script `4_matching.qmd` implements the daily matching of exposed to unexposed individuals on variables corresponding to nodes in the minimal sufficient adjustment set following the causal model, thereby attempting to close non-causal backdoor paths and limit bias. After describing the study population and providing crude estimates in script `5-descriptives.qmd`, a survival analysis is captured within script `6-survival-analysis.qmd`, visualizing survival over time by producing Kaplan-Meier curves and estimating the average treatment effect (ATE). A detailed documentation of the statistical methods, as well as a README file guiding users on the script deployment, accompanies the statistical scripts in the Github repository [Meurisse 2023b].

DuckDB, a lightweight database system, is used to increase the speed of running the workflow by enhancing performance when dealing with large amounts of data and complex analytical queries.



**Fig. 4** Graphical overview of the developed analytical pipeline. Consisting of different subsequent modules, each producing an interactive report. Implementation of step 5 of the proposed methodological approach to assess the real-world effectiveness of a primary vaccination schedule as compared to partial or no vaccination in preventing SARS-CoV-2 infection, in populations spanning national borders.

The DAG, CDM and synthetic dataset, together with all supporting research objects (see Figure 2), were published on Zenodo [Estupiñán-Romero 2023]. Further, the latter plus additional digital research objects produced along the way (i.e., a study protocol [Meurisse 2023a], a data management plan (DMP) [Martínez-Lizaga 2023], and an interoperable analytical pipeline) were collected together in a GitHub Repository [Meurisse 2023b].

A Research Object Crate (RO-Crate) [Sefton 2023] was generated to package these objects together with their metadata and specified relationships, acting in accordance with the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles [Wilkinson 2016]. This way, the digital research objects are persisting and shared with the wider community, and as such remain available for feedback from field experts.

## 4 Discussion

We present a methodological framework, providing a systematic approach to address policy-relevant causal research questions based on federated cross-national sensitive observational data in a privacy-preserving way and addressing challenges at different layers of interoperability. In this way, the current manuscript aims to provide guidance on the full methodological process to approach causal inference in federated research, which is currently lacking.

This approach facilitates the comparison and integration of causal estimates obtained from distributed analyses through approaching homogeneity in the considered confounders, criteria for study participant selection, definitions of variables and adopted statistical methods.

The conceptual and instrumental phase of the proposed methodological framework, consisting of different consecutive steps, was successfully demonstrated within a Use Case of the European BY-COVID project, thereby preparing the subsequent assessment of SARS-CoV-2 vaccine effectiveness in preventing infection in a population crossing national borders and prototyping a workflow that is standard for causal population health research. By going through these steps, different challenges and limitations of the proposed methodology were identified. These challenges, together with recommendations on how to address them, are described in the following sections.

### 4.1 Challenges and recommendations in addressing the emulation of an RCT using federated sensitive cross-border observational data

Estimation of the causal effect of interest can be approached by emulating an RCT, as suggested by [Hernán 2022], thereby mimicking randomization by conditioning on confounders and controlling for other potential biases. Crucial in this process is the identification of the set of confounders for which adjustment during the analysis is required to approach causal association under the assumptions captured in the causal model. However, this is a non-trivial task given that not all underlying mechanisms may be known [Lee 2016].

Theoretical identification of confounding paths and other biases was based on the assumptions of the researchers on the data-generating processes (e.g., method of



participant selection and variable measurement, or relationships between variables), which were in turn built upon available theory and expert knowledge at the time of analysis. Further, detailed patient information is required to adjust for the identified confounders and the necessary data might not be available in all Participant Nodes. When this is the case, it may result in the presence of residual confounding bias in the obtained causal effect estimates, hampering both the internal and external validity of the study [Andrade 2018] [Findley 2021] [Grimes 2002].

Moreover, taking into account potential biases when conducting causal inference using observational data results in complex analytical procedures which generally demand several human decision points. This raises the question whether analytical procedures to approach a complex causal research question can be fully automated. Sensitivity analyses can be implemented to assess the presence of biases, and sensitivity of the results to the methods implemented for confounder adjustment.

**Recommendation 1** The use of DAGs is recommended to build causal models, transparently displaying assumptions, aiding the identification of a minimal sufficient adjustment set (i.e., set variables for which you are required to adjust to estimate the causal effect of interest under the described assumptions), and in this way feeding into the specification of data requirements. DAGs are easy-to-use graphical instruments, facilitating explicit specification of assumptions. The DAGitty web application provides a practical tool to create and edit DAGs, and to identify the minimal sufficient adjustment set.

**Recommendation 2** By enabling field experts to provide feedback on the assumptions and iteratively updating the DAG when new information becomes available (i.e., building the causal model as a collaborative effort), theoretical identification of confounders and other sources of bias can be optimized. As such, we encourage making the constructed causal model publicly available, working on collaborative platforms (e.g., GitHub), and providing occasions (e.g., workshops) for field experts to evaluate it.

**Recommendation 3** Acknowledging the limitations to a causal interpretation of the results can be addressed by including an assessment of the data quality at each step of the analytical process. In addition to a general DQA and validation assessment, quality evaluations specific to the performed analytics can be performed. For example, when matching based on identified confounders is performed to obtain comparable intervention groups, an assessment of covariate balance is conducted and reported thereafter. In addition, every point of automated decision making should be documented in the output of the analytical pipeline to allow for a meaningful interpretation of the obtained results. For instance, automated decisions on whether or not to impute missing values or perform listwise deletion, are reported in an interactive report. Quarto [Quarto], an open-source tool, provides an efficient way to produce rich interactive outputs, registering the latter information.

## 4.2 Challenges and recommendations on the different layers of interoperability (LOST) when conducting federated causal research

### *Legal interoperability*

Legal constraints based on privacy and data protection regulations (GDPR) can block the re-use of sensitive personal data for population health research across (national) borders. Implementing a federated analysis approach where sensitive data stays under the jurisdiction and governance of data holders (i.e. data visiting principle) offers a solution. However, some steps in the proposed methodology may still pose legal challenges. Data-driven methods for developing synthetic data [Dube 2014] might give rise to concerns of re-identification and require access of the Coordination Node to real data.

Further, for individual Participant Nodes to comply with the CDM specification, linkage of and access to the required data sources for the research in question should be authorized. Access to sensitive real-world health data is in many European countries granted by an authorizing body, such as the national Data Protection Authority (DPA) or Research Ethics Committee (REC), based on the evaluation of a study protocol and data management plan. However, no standard process for applying for data access is available at this time in Europe, preventing the use of a uniform approach.

**Recommendation 4** Building a CDM, specifying data requirements, is recommended to comply with the principles of data minimization. In this way, we can limit the collection of sensitive information to what is strictly relevant and necessary for the purpose. By including a requirement level in the description of variables specified in the CDM, it can be indicated which variables are essential to close backdoor paths identified in the DAG.

**Recommendation 5** To facilitate a rapid data access application process, it is recommended for the Coordination Node to provide the Participant Nodes with the necessary tools, i.e., produce and share a comprehensive study protocol and give guidance for the development of a research DMP. A study protocol provides a plan of action and contains among others the study objectives and planned methodology for conducting the study. The study protocol additionally facilitates compliance with purpose minimization principles. Guidelines for writing a scientific study protocol can be consulted elsewhere [Al Jundi 2016]. For the development and publishing of a DMP, use of Argos' services [OpenAIRE 2023] [Papadopoulou 2021] can be recommended.

**Recommendation 6** As data-driven methods (i.e., based on real-world data) for developing synthetic data might have some weaknesses in terms of legal interoperability, we advise the manual development of synthetic data, without requiring access to real data [Dube 2014], capturing the structure, syntactic and semantic requirements as specified in the CDM and reflecting true distributions by using known population-level parameters.

### *Organizational interoperability*

To achieve organizational interoperability and reach common goals, organizations should define and align responsibilities, processes, and expectations [Margariti 2022]

[Weichhart 2014]. A diversity of theoretical backgrounds of researchers in the federated network was observed, resulting in the need for building a common ground. Further, unambiguously defining data requirements, which is essential to obtain uniform data across different federations, was additionally found to be a difficult task. An approach to collaboratively address the causal research question was required.

**Recommendation 7** Clearly assigning responsibilities (i.e., allocating the role of Coordination versus Participant Node), documenting processes and exchanging relevant information (i.e., publishing and sharing a research protocol, DMP, and digital research objects), and facilitating interactions within the federated network (e.g., using a collaboration platform) is recommended to achieve *organizational interoperability*. It is the role of the Coordination Node to supervise and synchronize the activities executed by the Participant Nodes, and to provide information and support to establish a common knowledge on the process and required involvement.

For example, in the demonstrated phase of the framework a theoretical overview on DAGs was given by the Coordination Node to participating partners in the form of a workshop. Regular contact with the Participant Nodes is required throughout the process, but more intensely when agreeing on the research question and defining the CDM specification. This allows the Participant Nodes to put forward ambiguities and requests for clarification. Transparency of the process, making research objects produced in every step openly available, is recommended to enhance trust and allow for providing expert-knowledge and user-based feedback. Use of a collaborative and version control platform, such as GitHub, enables collaboration with several partners.

#### ***Semantic interoperability***

Semantic interoperability indicates the consistency in meaning of exchanged data among organizations [de Mello 2022], enabling the interpretation of data independently of the partner involved. When working across borders, different Participant Nodes can have distinct codebooks and use different classification systems. Mapping a definition to different classification systems and identifying intersections between these classification systems (i.e., specifying crosswalks), is not always straightforward. Further, definitions of variables and cohorts can be ambiguous and open to interpretation.

**Recommendation 8** The construction of a CDM is recommended to ensure a uniform syntactic structure (i.e., format and grammar) and meaning (i.e., semantics) of elements of the distributed data used to address the specified cross-border research question. To improve compliance with data requirements and consistency between distributed datasets, involving collaborators within the nodes in reviewing the specifications captured in the CDM is recommended. Based on this evaluation, ambiguities can be eliminated and the specification of crosswalks, mapping definitions to different classification systems, can be optimized.

#### ***Technical interoperability***

A critical part of deploying a reproducible analytical pipeline, is dealing with dependencies of the pipeline, ensuring consistent deployment independent from the system in which it is executed, and in this way ensuring *technical interoperability*. Packaging the analytical pipeline created in the prototyped workflow within a container

(i.e. an isolated portable execution environment including analytical code), such as a Docker container, presents a way for easy transmission of scripts, easy management of dependencies and allows for consistent execution of the analyses in different premises, decoupled from the local system. However, use of container technology might not be feasible due to a lack of support for Docker within certain operating systems or other organizational barriers to deployment.

Further, the functioning of the analytical pipeline code relies heavily on the ability of the Participant Nodes to comply with the CDM specification. When the input data of the analytical pipeline does not conform with the specified semantic and syntactic requirements, the process will fail before generating the required output statistics. Efficiency challenges were also encountered when inferring causality for the entire population of a country or region, requiring the handling of large volumes of data.

**Recommendation 9** When transferring and deploying an interoperable analytical pipeline, the use of existing technological solutions to deal with the required dependencies and allowing deployment of the analytical process consistently and independently from the local execution system (e.g., Docker containers), is recommended. When deploying a container is not feasible within a Participant Node’s system, several alternative strategies can be adopted, such as deploying the Docker within a virtual machine, deploying the Docker container within a research environment provided by a trusted third party, or manual installation of the required dependencies and manual execution of the analysis scripts.

**Recommendation 10** To facilitate rapid deployment in the Participant Nodes, it is recommended to provide users deploying the analytical pipeline with feedback based on error logs when compliance with the structure and syntactic requirements captured in the CDM is lacking for the input data and the process is failing. Further, it is recommended to check the distributed input data against a set of logical validation rules, examining compliance to the specifications captured in the CDM. This can be implemented as one of the sequential steps in the analytical pipeline.

**Recommendation 11** To deal with efficiency challenges when handling large volumes of data, it can be recommended to implement efficient programming strategies [Gillespie 2016] [Lutz 2013], to parallelize heavy and repetitive computations where this increases throughput, and work with data management solutions, such as DuckDB [Raasveldt 2019].

### *Challenges and recommendations related to the reuse of digital objects*

There is an increasing demand for researchers to document and share the data and research objects supporting their scientific conclusions, to increase transparency, facilitate collaboration, and allow subsequent replication, integration and reuse by the community [Wilkinson 2016]. In public health, this can be essential in the response to emerging public health threats (e.g., a pandemic). However, processes might not always be well documented, researchers might not be aware of the benefits of sharing their data or research objects (e.g., avoiding duplication, greater visibility), or potentially can’t locate appropriate repositories.

**Recommendation 12** To facilitate the exchange and reuse of the digital research objects of the workflow and in this way enabling an accelerated response in the case

of a new pandemic, it is recommended to publish these objects following the FAIR principles. To make your object Findable, rich metadata should be provided. Making the digital objects Accessible, means that they have to be retrievable when access is allowed. For this purpose, the objects can be shared with rich metadata in open repositories like Zenodo. In the context of an interconnected workflow, RO-Crate [Sefton 2023] provides an alternative approach to package research objects together with their metadata, allowing the indication of relationships between entities. Further, the objects should be made Interoperable, by using standards and controlled vocabularies, and Reusable, by providing clear documentation (e.g., a README) [Wilkinson 2016].

## 5 Conclusion and future perspectives

The proposed methodological framework provides guidance in the form of a systematic approach to address federated cross-national causal research questions in a privacy-preserving way, while tackling challenges at different layers of interoperability. Additionally, the conceptual and instrumental phase of the methodological framework was demonstrated in the current work, thereby prototyping a standard workflow for causal population health research.

Describing the methodological framework, publishing the produced research objects (e.g., causal model, CDM and synthetic data) and prototyping a workflow using open-source tools available for reuse, allows researchers to respond more rapidly to newly emerging public health research questions and in this way contributes to pandemic preparedness.

Future planned work in the context of the BY-COVID project entails the implementation of the proposed methodology and the actual assessment of SARS-CoV-2 vaccine effectiveness in preventing infection in a population crossing national borders. This proof-of-concept will evaluate the value of the proposed framework in terms of drawing conclusions on causality for the specified research question, the linkage of heterogeneous data sources and data transformation by the Participant Nodes to comply with the specified data requirements, the deployment of the developed analytical pipeline in a distributed manner across different Participant Nodes, and the pooling of these results for a meta-analysis.

Further research is needed to test the implications of the implementation of alternative statistical methods for causal inference using a federated research approach, although the framework enables the use of any method currently available. Upgrading the proposed methodological framework and applying it to new policy-relevant questions in emerging public and population health issues can be considered an important research priority in the field of federated causal inference.

## 6 List of abbreviations

- ATE: Average Treatment Effect
- BY-COVID: BeYond-COVID
- COVID-19: coronavirus disease 2019
- CDM: Common Data Model
- DAG: Directed Acyclic Graphs

- DMP: Data Management Plan
- DPA: Data Protection Authority
- DQA: Data Quality Assessment
- EDA: Exploratory Data Analysis
- EIF: European Interoperability Framework
- FAIR: Findable, Accessible, Interoperable, and Reusable
- GDPR: General Data Protection Regulation
- IACS: Instituto Aragonés de Ciencias de la Salud
- IPW: Inverse Probability Weighting
- IVAC: International Vaccine Access Center
- LOST: Legal, Organizational, Semantic and Technical
- PICO: Patient/Population, Intervention, Comparison/Control, and Outcome
- RCT: Randomized Controlled Trial
- REC: Research Ethics Committee.
- RO-Crate: Research Object Crate
- SARS-CoV-2: severe acute respiratory syndrome coronavirus 2

## Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable.

### Availability of data and materials

The research objects generated during the current study (i.e., a study protocol, a data management plan, a causal model, a common data model, a synthetic dataset, an interoperable analytical pipeline) are available from Zenodo [Martínez-Lizaga 2023] and GitHub <https://w3id.org/ro/doi/10.5281/zenodo.6913045>

### Competing interests

The authors declare no competing interests.

### Funding

This research was supported by BY-COVID Project funds. BY-COVID (BeYond COVID) is a Horizon Europe funded project from the European Union (grant agreement HORIZON-INFRA-2021-EMERGENCY-01 101046203), launched in October 2021.

## Authors' contributions

FER, EBD, NVG and MM conceptually developed the study and methodology. FER, NVG, EBD, MM, JGG, NML, SRS, SS, LDA, ADM, and SSR contributed to the development of the research objects produced during the study. SS, LDA, and ADM critically reviewed the intellectual content of the study. JGG, MM, SRS, FER, and SSR contributed to the development of data generation and analysis code and to the maintenance of code dependencies. NVG and EBD managed and coordinated the research activities and took leadership responsibility for the planning and execution. MM drafted the manuscript. FER, NVG, EBD, and LDA reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

## Consortia

BeYond-COVID project contributors: Vasso Kalaitzi (KNAW/DANS), Claudia Habl (GÖG), Gunter Maier (GÖG), Mirjam Knol (RIVM), Chantal Reusken (RIVM), Mariken Tjihuis (RIVM), Leon Schutte (RIVM), Markus Perola (THL), Teemu Paajanen (THL), Kati Kristiansson (THL), Pekka Jousilahti (THL), Jostein Starrfelt (NIPH), Hinta Meijerink (NIPH).

## References

- [Abboud 2018] L A Abboud, P Bogaert, A Fehr, D Urbanski, H Tolonen, I Noguier-Zambran, A Padron-Monedero, L Palmieri, H Van Oyen (2018):  
**The new Joint Action on Health Information: information for action (InfAct)!**  
*European Journal of Public Health* **28**(suppl.4)  
<https://doi.org/10.1093/eurpub/cky212.651>
- [Al Jundi 2016] Azzam Al Jundi, Salah Sakka (2016):  
**Protocol Writing in Clinical Research.**  
*Journal of Clinical and Diagnostic Research* **10**(1)  
<https://doi.org/10.7860/jcdr/2016/21426.8865>
- [Andrade 2018] Chittaranjan Andrade (2018):  
**Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation.**  
*Indian Journal of Psychological Medicine* **40**(5)  
[https://doi.org/10.4103/ijpsym.ijpsym\\_334\\_18](https://doi.org/10.4103/ijpsym.ijpsym_334_18)
- [Attema 2021] T. Attema T, D. Worm (2021):  
**Technologische doorbraak: eindelijk een privacyvriendelijke manier om data te benutten.**  
**Technological breakthrough: Finally, a privacy-friendly way to harness data.**  
TNO whitepaper, 953235.  
<http://resolver.tudelft.nl/uuid:8002b966-7bba-427c-b343-56326c1a587b>
- [Bareinboim 2012] Elias Bareinboim, Judea Pearl (2012):  
**Controlling Selection Bias in Causal Inference.**  
*Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*  
*Proceedings of Machine Learning Research* **22**  
<https://proceedings.mlr.press/v22/bareinboim12.html>
- [Bashari Rad 2017] Babak Bashari Rad, Harrison John Bhatti, Mohammad Ahmadi (2017):  
**An Introduction to Docker and Analysis of its Performance.**

- IJCSNS International Journal of Computer Science and Network* **17**(3)  
[http://paper.ijcsns.org/07\\_book/201703/20170327.pdf](http://paper.ijcsns.org/07_book/201703/20170327.pdf)
- [Beyan 2020] Oya Beyan, Ananya Choudhury, Johan van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Md. Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino da Silva Santos, Andre Dekker (2020):  
**Distributed Analytics on Sensitive Medical Data: The Personal Health Train.**  
*Data Intelligence* **2**(1–2) [https://doi.org/10.1162/dint\\_a.00032](https://doi.org/10.1162/dint_a.00032)
- [Boettiger 2015] Carl Boettiger (2015):  
**An introduction to Docker for reproducible research.**  
*ACM SIGOPS Operating Systems Review* **49**(1)  
<https://doi.org/10.1145/2723872.2723882>
- [Bogaert 2021] P Bogaert, N Schutte (2021):  
**Towards a Population Health Information Research Infrastructure.**  
*European Journal of Public Health* **31**(Supplement\_3)  
<https://doi.org/10.1093/eurpub/ckab164.572>
- [BY-COVID] **BY-COVID.**  
<https://by-covid.org/> [archived 2023-09-17]
- [Broberg 2019] Pontus Broberg, Shapour Jahanshahi (2019):  
**Using eye tracking to study variable naming conventions and their effect on code readability.**  
 Dissertation, Master in Computer Science.  
 KTH, School of Electrical Engineering and Computer Science (EECS).  
 TRITA-EECS-EX 2019:323  
<https://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Aakth%3Adiva-255048>
- [Cai 2015] Li Cai, Yangyong Zhu (2015):  
**The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.**  
*Data Science Journal* **14**  
<https://doi.org/10.5334/dsj-2015-002>
- [CIPH 2020] Croatian Institute of Public Health (2020):  
**LOST\* and found: Report on interoperability landscape in Europe.**  
*Joint Action on Health Information* (INFACT)  
<https://www.inf-act.eu/sites/inf-act.eu/files/2020-11/D10.1.pdf>
- [Digitale 2022] Jean C. Digitale, Jeffrey N. Martin, Medellena Maria Glymour (2022):  
**Tutorial on directed acyclic graphs.**  
*Journal of Clinical Epidemiology* **142**  
<https://doi.org/10.1016/j.jclinepi.2021.08.001>
- [DiLep 2019] Carleton DiLeo (2019):  
**Naming Things.**  
*Clean Ruby: A Guide to Crafting Better Code for Rubyists*  
 ISBN 978-1-4842-5546-9  
[https://doi.org/10.1007/978-1-4842-5546-9\\_2](https://doi.org/10.1007/978-1-4842-5546-9_2)
- [Dube 2014] Kudakwashe Dube, Thomas Gallagher (2014):  
**Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use.**  
*Foundations of Health Information Engineering and Systems* (FHIES 2013), Third International Symposium, 2013-08-21/-23, Macau, China.  
[https://doi.org/10.1007/978-3-642-53956-5\\_6](https://doi.org/10.1007/978-3-642-53956-5_6)
- [EC 2017] Directorate-General for Informatics (European Commission) (2017):  
**New European interoperability framework: promoting seamless services and data flows**



- for European public administrations.  
*Publications Office of the European Union*  
<https://doi.org/10.2799/78681>
- [Estupiñán-Romero 2023] Francisco Estupiñán-Romero, Nina Van Goethem, Marjan Meurisse, Javier González-Galindo, Enrique Bernal-Delgado (2023):  
**BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Common Data Model Specification.**  
*Zenodo*  
<https://doi.org/10.5281/zenodo.7572373>
- [Faraglia 2023] Daniele Faraglia (2023):  
**Welcome to Faker’s documentation!**. Faker 18.13.0 documentation.  
*Read the Docs*  
<https://faker.readthedocs.io/>
- [Findley 2021] Michael G. Findley, Kyosuke Kikuta, Michael Denly (2021):  
**External Validity.**  
*Annual Review of Political Science* **24**(1)  
<https://doi.org/10.1146/annurev-polisci-041719-102556>
- [Foster 2017] Erin D. Foster, Ariel Deardorff (2017):  
**Open Science Framework (OSF).**  
*Journal of the Medical Library Association* **105**(2):  
<https://doi.org/10.5195/jmla.2017.88>
- Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, Rebecca Wilson, Oliver Butters, Barnaby Murtagh, Ipek Demir, Dany Doiron, Lisette Giepmans, Susan E Wallace, Isabelle Budin-Ljøsne, Carsten Oliver Schmidt, Paolo Boffetta, Mathieu Boniol, Maria Bota, Kim W Carter, Nick deKlerk, Chris Dibben, Richard W Francis, Tero Hiekkalinna, Kristian Hveem, Kirsti Kvaløy, Sean Millar, Ivan J Perry, Annette Peters, Catherine M Phillips, Frank Popham, Gillian Raab, Eva Reischl, Nuala Sheehan, Melanie Waldenberger, Markus Perola, Edwin van den Heuvel, John Macleod, Bartha M Knoppers, Ronald P Stolck, Isabel Fortier, Jennifer R Harris, Bruce HR Woffenbuttel, Madeleine J Murtagh, Vincent Ferretti, Paul Burton (2014):  
**DataSHIELD: taking the analysis to the data, not the data to the analysis.**  
*International Journal of Epidemiology* **43**(6)  
<https://doi.org/10.1093/ije/dyu188>
- [Gillespie 2016] Colin Gillespie, Robin Lovelace (2016):  
**Efficient R Programming.** First Edition.  
O’Reilly Media, Inc.  
<https://identifiers.org/isbn/9781491950784>
- [Glass 2013] Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, Jonathan M. Samet (2013):  
**Causal Inference in Public Health.**  
*Annual Review of Public Health* **34**(1)  
<https://doi.org/10.1146/annurev-publhealth-031811-124606>
- [González-García 2021] Juan González-García, Francisco Estupiñán-Romero, Javier González-Galindo, Carlos Telleria-Orriols, Luigi Palmieri, Andrea Faragalli, Ivan Pristās, Jakov Vuković, Jānis Misiņš, Irisa Zile, Enrique Bernal-Delgado, InfAct Consortium (2021):  
**Coping with Interoperability in the Development of A Federated Research Infrastructure:** Achievements, Challenges and Recommendations from the JA Infact.  
*Archives of Public Health* **79**:221  
<https://doi.org/10.1186/s13690-021-00731-z>

- [Greenland 2009] Sander Greenland, James M Robins (2009):  
**Identifiability, exchangeability and confounding revisited.**  
*Epidemiologic Perspectives & Innovations* **6**(1)  
<https://doi.org/10.1186/1742-5573-6-4>
- [Grimes 2002] David A Grimes, Kenneth F Schulz (2002):  
**Bias and causal associations in observational research.**  
*The Lancet* **359**(9302)  
[https://doi.org/10.1016/s0140-6736\(02\)07451-2](https://doi.org/10.1016/s0140-6736(02)07451-2)
- [Haukoos 2007] Jason S. Haukoos, Craig D. Newgard (2007):  
**Advanced Statistics: Missing Data in Clinical Research-Part 1: An Introduction and Conceptual Framework.**  
*Academic Emergency Medicine* **14**(7)  
<https://doi.org/10.1111/j.1553-2712.2007.tb01855.x>
- [Hernán 2004] Miguel A. Hernán, Sonia Hernández-Díaz, James M. Robins (2004):  
**A Structural Approach to Selection Bias.**  
*Epidemiology* **15**(5)  
<https://doi.org/10.1097/01.ede.0000135174.63482.43>
- [Hernán 2006] Miguel A Hernán, James M Robins (2006):  
**Estimating causal effects from epidemiological data.**  
*Journal of Epidemiology & Community Health* **60**(7)  
<https://doi.org/10.1136/jech.2004.029496>
- [Hernán 2016] Miguel A. Hernán, James M. Robins (2016):  
**Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.**  
*American Journal of Epidemiology* **183**(8)  
<https://doi.org/10.1093/aje/kwv254>
- [Hernán 2020] Miguel A. Hernan, James M. Robins (2020):  
**Causal Inference: What If.**  
 1st edition. Boca Raton: Chapman & Hall/CRC  
<https://identifiers.org/isbn/9781420076165> [preprint]
- [Hernán 2022] Miguel A. Hernán, Wei Wang, David E. Leaf (2022):  
**Target Trial Emulation: A Framework for Causal Inference From Observational Data.**  
*JAMA* **328**(24)  
<https://doi.org/10.1001/jama.2022.21383>
- [IVAC 2023] International Vaccine Access Center (IVAC) (2023):  
**VIEW-hub.**  
<https://view-hub.org/covid-19/effectiveness-studies> [archived 2023-09-17]
- [Kang 2013] Hyun Kang (2013):  
**The prevention and handling of the missing data.**  
*Korean Journal of Anesthesiology* **64**(5)  
<https://doi.org/10.4097/kjae.2013.64.5.402>
- [Kasza 2017] Jessica Kasza, Rory Wolfe, Tibor Schuster (2017):  
**Assessing the impact of unmeasured confounding for binary outcomes using confounding functions.**  
*International Journal of Epidemiology* **46**(4)  
<https://doi.org/10.1093/ije/dyx023>
- [Lee 2016] Paul H. Lee, Igor Burstyn (2016):  
**Identification of confounder in epidemiologic data contaminated by measurement error in covariates.**

- BMC Medical Research Methodology* **16**(1)  
<https://doi.org/10.1186/s12874-016-0159-6>
- [Li 2014] L. Li, K. Kleinman, M. W. Gillman (2014):  
**A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity.**  
*Journal of Developmental Origins of Health and Disease* **5**(6)  
<https://doi.org/10.1017/s2040174414000415>
- [Lira 2019] Authors: Rodrigo Pessoa Cavalcanti Lira, Eduardo Melani Rocha (2019):  
**PICOT: Imprescriptible items in a clinical research question.**  
*Arquivos Brasileiros de Oftalmologia* **82**(2)  
<https://doi.org/10.5935/0004-2749.20190028>
- [Listl 2016] Stefan Listl, Hendrik Jürges, Richard G. Watt (2016):  
**Causal inference from observational data.**  
*Community Dentistry and Oral Epidemiology* **44**(5)  
<https://doi.org/10.1111/cdoe.12231>
- [Little 2019] Roderick Little, Donald Rubin (2019):  
**Statistical Analysis with Missing Data.** Third Edition.  
 John Wiley & Sons.  
<https://doi.org/10.1002/9781119482260>
- [Lutz 2013] Mark Lutz (2013):  
**Powerful Object-Oriented Programming.** Fifth Edition.  
 O'Reilly Media, Inc.  
<https://identifiers.org/isbn/9781449355739>
- [Margariti 2022] Vasiliki Margariti, Teta Stamati, Dimosthenis Anagnostopoulos, Mara Nikolaidou, Anastasia Papastilianou (2022):  
**A holistic model for assessing organizational interoperability in public administration.**  
*Government Information Quarterly* **39**(3)  
<https://doi.org/10.1016/j.giq.2022.101712>
- [Martínez-Lizaga 2023] Natalia Martínez-Lizaga; Marjan Meurisse; Francisco Estupiñán-Romero; Nina Van Goethem; Enrique Bernal-Delgado (2023):  
**BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Data Management Plan.**  
*Zenodo*, Data management plan.  
<https://doi.org/10.5281/zenodo.7625784>
- [de Mello 2022] Blanda Helena de Mello, Sandro José Rigo, Cristiano André da Costa, Rodrigo da Rosa Righi, Bruna Donida, Marta Rosecler Bez, Luana Carina Schunke (2022):  
**Semantic interoperability in health records standards: a systematic literature review.**  
*Health and Technology* **12**(2)  
<https://doi.org/10.1007/s12553-022-00639-w>
- [Meurisse 2023a] Marjan Meurisse, Nina Van Goethem, Francisco Estupiñán-Romero, Javier González-Galindo, Santiago Royo-Sierra, Natalia Martínez-Lizaga, Enrique Bernal-Delgado (2023):  
**BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Study protocol.**  
*Zenodo*  
<https://doi.org/10.5281/zenodo.7825979>
- [Meurisse 2023b] Marjan Meurisse, Francisco Estupiñán-Romero, Nina Van Goethem, Javier González-Galindo, Santiago Royo-Sierra, Enrique Bernal-Delgado (2023):  
**BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness**

- assessment.**  
*BY-COVID Project*, RO-Crate.  
<https://w3id.org/ro/doi/10.5281/zenodo.6913045>
- [Moncada-Torres 2021] Arturo Moncada-Torres, Frank Martin, Melle Sieswerda, Johan Van Soest, Gijs Geleijnse (2021):  
**VANTAGE6**: an open source priVAcY preserviNg federaTed leArninG infrastruCTurE for Secure Insight eXchange.  
*AMIA Annual Symposium Proceedings 2020*  
<https://identifiers.org/pmc/PMC8075508>
- [Nishikawa-Pacher 2022] Andreas Nishikawa-Pacher (2022):  
**Research Questions with PICO**: A Universal Mnemonic.  
*Publications 10*(3)  
<https://doi.org/10.3390/publications10030021>
- [OpenAIRE 2023] OpenAIRE (2023):  
**Argos**.  
<https://argos.openaire.eu/> [archived 2023-09-17]
- [Papadopoulou 2021] Elli Papadopoulou (2021):  
**ARGOS: plan and follow your data**.  
*OpenAIRE National Workshop in Malta*, 2021-02-24.  
[https://www.um.edu.mt/library/oar/bitstream/123456789/70269/1/ARGOS\\_plan\\_and\\_follow\\_your\\_data\\_2021.pdf](https://www.um.edu.mt/library/oar/bitstream/123456789/70269/1/ARGOS_plan_and_follow_your_data_2021.pdf)
- [Pearce 2016] Neil Pearce, Debbie A Lawlor (2016):  
**Causal inference—so much more than statistics**.  
*International Journal of Epidemiology 45*(6)  
<https://doi.org/10.1093/ije/dyw328>
- [Piccolo 2016] Stephen R. Piccolo, Michael B. Frampton (2015):  
**Tools and techniques for computational reproducibility**.  
*GigaScience 5*(1)  
<https://doi.org/10.1186/s13742-016-0135-4>
- [Quarto] **Quarto**.  
<https://quarto.org/> [archived 2023-09-17]
- [Raasveldt 2019] Mark Raasveldt, Hannes Mühleisen (2019):  
**DuckDB**: an Embeddable Analytical Database.  
*Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*.  
<https://doi.org/10.1145/3299869.3320212>
- [Riva 2012] John J. Riva, Keshena M.P. Malik, Stephen J. Burnie, Andrea R. Endicott, Jason W. Busse (2012):  
**What is your research question?** An introduction to the PICOT format for clinicians.  
*The Journal of the Canadian Chiropractic Association 56*(3)  
<https://identifiers.org/pmc/PMC3430448>
- [Sefton 2023] Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes, Oscar Corcho, Daniel Garijo, Raul Palma, Frederik Coppens, Carole Goble, José María Fernández, Kyle Chard, Jose Manuel Gomez-Perez, Michael R Crusoe, Ignacio Eguinoa, Nick Juty, Kristi Holmes, Jason A. Clark, Salvador Capella-Gutierrez, Alasdair J. G. Gray, Stuart Owen, Alan R. Williams, Giacomo Tartari, Finn Bacall, Thomas Thelen, Hervé Ménager, Laura Rodríguez-Navas, Paul Walk, brandon whitehead, Mark Wilkinson, Paul Groth, Erich Bremer, LJ Garcia Castro, Karl Sebby, Alexander Kanitz, Ana Trisovic, Gavin Kennedy, Mark Graves, Jasper Koehorst, Simone Leo, Marc Portier, Paul Brack, Milan Ojsteršek, Bert Droesbeke, Chenxu Niu, Kosuke Tanabe, Tomasz Miksa, Marco La Rosa, Cedric Decruw, Andreas Czerniak, Jeremy Jay, Sergio Serra, Ronald Siebes, Shaun de Witt, Shady El Damaty, Douglas Lowe,

- Xuanqi Li, Sveinung Gundersen, Muhammad Radifar (2023):  
**RO-Crate Metadata Specification 1.1.3.**  
*ResearchObject.org / Zenodo*  
<https://doi.org/10.5281/zenodo.7867028>  
<https://w3id.org/ro/crate/1.1>
- [Spellman 2018] Barbara A. Spellman, Elizabeth A. Gilbert, Katherine S. Corker (2018):  
**Open Science.**  
*Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, John T. Wixted (ed.)  
<https://doi.org/10.1002/9781119170174.epcn519>
- [Staplin 2017] Natalie Staplin, William G. Herrington, Parminder K. Judge, Christina A. Reith, Richard Haynes, Martin J. Landray, Colin Baigent, Jonathan Emberson (2017):  
**Use of Causal Diagrams to Inform the Design and Interpretation of Observational Studies: An Example from the Study of Heart and Renal Protection (SHARP).**  
*Clinical Journal of the American Society of Nephrology* **12**(3)  
<https://doi.org/10.2215/cjn.02430316>
- [Suzuki 2020] Etsuji Suzuki, Tomohiro Shinozaki, Eiji Yamamoto (2020):  
**Causal Diagrams: Pitfalls and Tips.**  
*Journal of Epidemiology* **30**(4)  
<https://doi.org/10.2188/jea.je20190192>
- [Tennant 2021] Peter W G Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lysie R Ranker, Johannes Textor, Georgia D Tomova, Mark S Gilthorpe, George T H Ellison (2021):  
**Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations.**  
*International Journal of Epidemiology* **50**(2)  
<https://doi.org/10.1093/ije/dyaa213>
- [Textor 2016] Johannes Textor, Benito van der Zander, Mark S. Gilthorpe, Maciej Liškiewicz, George T.H. Ellison (2017):  
**Robust causal inference using directed acyclic graphs: the R package ‘dagitty’.**  
*International Journal of Epidemiology* **45**(6)  
<https://doi.org/10.1093/ije/dyw341>
- YData 2023a. YData (2023):  
**ydata-profiling.** *GitHub*  
<https://github.com/ydataai/ydata-profiling>
- YData 2023b YData (2023):  
**Welcome - YData Profiling 4.5.** *YData* <https://docs.profiling.ydata.ai/4.5/>
- [Weichhart 2014] Georg Weichhart (2014):  
**Learning for Sustainable Organisational Interoperability.**  
*IFAC Proceedings Volumes* **47**(3)  
<https://doi.org/10.3182/20140824-6-za-1003.01590>
- [Wilkinson 2016] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei,

Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, Barend Mons (2016): **The FAIR Guiding Principles for scientific data management and stewardship.**

*Scientific Data* **3**(1):160018

<https://doi.org/10.1038/sdata.2016.18>

[Wolfson 2010] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, Isabel Fortier, Bartha M Knoppers, Paul R Burton (2010):

**DataSHIELD: resolving a conflict in contemporary bioscience**—performing a pooled analysis of individual-level data without sharing the data.

*International Journal of Epidemiology* **39**(5)

<https://doi.org/10.1093/ije/dyq111>