

Integration of Real-world Heterogeneous Data Sources for Public Health Research Within and Across Jurisdictional Borders

Marjan Meurisse

A thesis submitted in fulfilment of the requirements for the degree of
Doctor in Public Health Sciences
of the
Université Catholique de Louvain

Belgium, January 2025

Marjan Meurisse

Integration of Real-world Heterogeneous Data Sources for Public Health Research Within and Across
Jurisdictional Borders

Front cover created with Midjourney

Jury

President	Prof. Niko Speybroeck Institut de recherche santé et société Faculté de santé publique Université catholique de Louvain, Brussels, Belgium
Promotor	Prof. Annie Robert Pôle Epidémiologie et Biostatistique Institut de Recherche Expérimentale et Clinique Faculté de santé publique Université catholique de Louvain, Brussels, Belgium
Co-promotor	Dr. Nina Van Goethem Public Health and Genome Scientific Directorate of Epidemiology and public health Sciensano, Brussels, Belgium
Secretary	Prof. Miika Vikkula Institut de Duve Faculté de médecine et médecine dentaire Université catholique de Louvain, Brussels, Belgium
Internal members	Dr. Jérôme Ambroise Centre de technologies moléculaires appliquées Institut de Recherche Expérimentale et Clinique Faculté de santé publique Université catholique de Louvain, Brussels, Belgium Prof. Herman Van Oyen Department of Public Health and Primary Care Gent University, Gent, Belgium

**External
members**

Dr. Emilie Cauët

Public Health and Genome

Scientific Directorate of Epidemiology and public health

Sciensano, Brussels, Belgium

Prof. Stijn Vansteelandt

Department of Applied Mathematics, Computer Science, and Statistics

UGent, Gent, Belgium

Prof. Kristel Van Steen

Department of Electrical Engineering and Computer Science

University of Liège, Liège, Belgium

Academic resume

Peer reviewed publications

Meurisse M, Van Oyen H, Blot K, Catteau L, Serrien B, Klamer S, Cauët E, Robert A*, Van Goethem N*. Evaluating methodological approaches to assess the severity of infection with SARS-CoV-2 variants: scoping review and applications on Belgian COVID-19 data. BMC Infectious Diseases. 2022;22:839 *Shared last author.

Meurisse M, Catteau L, van Loenhout J A F, Braeye T, De Mot L, Serrien B, Blot K, Cauët E, Van Oyen H, Cuypers L, Belgian Collaborative Group on COVID-19 Hospital Surveillance, COVID-19 Genomics Belgium Consortium, Robert A*, Van Goethem N* Homologous and heterologous prime-boost vaccination: impact on clinical severity of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients in Belgium. Vaccines. 2023; 11: 378 *Shared last author.

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*, for the BeYond-COVID project. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. BMC Medical Research Methodology. 2023;23:248. [†]Shared first author. *Shared last author.

Meurisse M[†], Estupiñán-Romero F[†], Perola M, Paajanen T, González-Galindo J, Van Goethem N*, Bernal-Delgado E*. Real-world comparative effectiveness of SARS-CoV-2 primary vaccination campaigns against SARS-CoV-2 infections: a federated observational study emulating a target trial in three nations. [†]Equal contribution *Equal contribution as senior authors.

Submitted to PLOS ONE

Non-peer reviewed publications

Estupiñán-Romero F, Van Goethem N, **Meurisse M**, González-Galindo J, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Common Data Model Specification. 2023. <https://zenodo.org/doi/10.5281/zenodo.6913045>.

Meurisse M, Van Goethem N, Estupiñán-Romero F, González-Galindo J, Royo-Sierra S, Martínez-Lizaga N, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Study protocol. 2023. <https://zenodo.org/doi/10.5281/zenodo.7551181>.

Van Goethem N, **Meurisse M**, González-Galindo J, Martínez-Lizaga N, Estupiñán-Romero F, Bernal-Delgado E. Continuously evolving demonstrator - Baseline use case: Prototyping a workflow that is standard to population health research. 2023. <https://zenodo.org/doi/10.5281/zenodo.7993139>.

Meurisse M, Estupiñán-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E. BY-COVID WP5 T5.2 Baseline Use Case. 2023. https://by-covid.github.io/BY-COVID_WP5_T5.2_baseline-use-case/.

Estupiñán-Romero F, González-Galindo J, **Meurisse M**, Martínez-Lizaga N, Royo-Sierra S, Van Goethem N, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment preliminary outputs from Aragon, Spain. 2023. <https://zenodo.org/doi/10.5281/zenodo.7871533>.

Meurisse M, Estupiñán-Romero F, Van Goethem N, Perola M, Paaanen T, González-Galindo J, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Local outputs and comparative analysis. 2024. <https://zenodo.org/doi/10.5281/zenodo.11209362>.

Meurisse M, Saldner S, Van Goethem N, Kalaitzi V, Estupiñán-Romero F, Bernal-Delgado E. BY-COVID Spring 24 Baseline Use Case Workshop. 2024. <https://zenodo.org/doi/10.5281/zenodo.12168494>.

Van Goethem N, **Meurisse M**, Estupiñán-Romero F, Bernal-Delgado E, Saldner S, Kalaitzi V. BY-COVID - WP5 - Baseline Use Case: Integration of socioeconomic data in observational studies using a prototyped workflow standard to federated population health research. 2024. <https://zenodo.org/doi/10.5281/zenodo.12771702>.

Van Goethem N, Bernal-Delgado E, Estupiñán-Romero F, **Meurisse M**. BY-COVID D5.1 Enriched report viral variants and health outcomes. 2024. <https://doi.org/10.5281/zenodo.13870075>

Presentations at international conferences

Meurisse M*, Van Goethem N, Vandromme M, De Mot L, Blot K, Robert A. *Sévérité de l'infection par le variant Delta du SARS-CoV-2 chez les patients hospitalisés en Belgique*. IXème Congrès International d'Epidémiologie «Épidémiologie et santé publique: union des forces en francophonie» (EPITER-ADELFI), poster presentation, 18-19 Aug 2022, Québec, Canada. *Presenting author.

Meurisse M*, Estupiñán-Romero F, Van Goethem N, González-Galindo J, Bernal-Delgado E. *Framework for federated causal inference based on real-world observational data sources (BY-COVID use case)*.

16th European Public Health (EPH) Conference, oral presentation, 10 Nov 2023, Dublin, Ireland.

*Presenting author.

Other presentations

Meurisse M. *Imputation of missing data in federated analysis*. Seminar IACS, oral presentation, 17 Feb 2023, IACS, Zaragoza, Spain.

Meurisse M. *Integration of heterogeneous data sources for public health research across national borders: the intersection of genomics and real-world population and healthcare data*. Seminar IREC-EPID, oral presentation, 14 Sep 2023, UCLouvain, Woluwe, Belgium.

Meurisse M, Catteau L, van Loenhout J A F, Braeye T, De Mot L, Serrien B, Blot K, Cauët E, Van Oyen H, Cuypers L, Belgian Collaborative Group on COVID-19 Hospital Surveillance, COVID-19 Genomics Belgium Consortium, Robert A, Van Goethem N. *Homologous and heterologous prime-boost vaccination: Impact on clinical severity of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients*. IREC PhD day, poster presentation, 22 Sep 2023, UCLouvain, Woluwe, Belgium.

Meurisse M, Catteau L, van Loenhout J A F, Braeye T, De Mot L, Serrien B, Blot K, Cauët E, Van Oyen H, Cuypers L, Belgian Collaborative Group on COVID-19 Hospital Surveillance, COVID-19 Genomics Belgium Consortium, Robert A, Van Goethem N. *Homologous and heterologous prime-boost vaccination: Impact on clinical severity of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients*. Poster lunch (for SBIM-students), poster presentation, 24 Oct 2023, UCLouvain, Woluwe, Belgium.

Attended scientific conferences

IXème Congrès International d'Epidémiologie « Épidémiologie et santé publique : union des forces en francophonie », Québec, Canada, 18-19 Aug 2022.

IREC PhD day, Doctoral School in the Institute of Experimental and Clinical Research (IREC), at UCLouvain, Brussels, 22 Sep 2023.

16th European Public Health Conference 2023: Our Food, Our Health, Our Earth: A Sustainable Future for Humanity, Dublin, Ireland, 8-11 Nov 2023.

Attended Courses

- *Causality and missing data*, Master of Science in Statistical Data Analysis, Ghent University, Ghent, Feb-June 2022.
- *Epicourse*, Sciensano, Brussels, 22-23-24-30-31 Mar 2022.
- *R introduction course (as a co-lecturer)*, Sciensano, Brussels, Mar 2022.
- *R introduction course (as a co-lecturer)*, Sciensano, Brussels, Oct 2022.
- *Celebrating causal inference (in statistics, in medicine and public health)*, Ghent University, Ghent, 11 Oct 2022.
- *Hack4Health R Hackathon*, Sciensano, Brussels, 24 Nov 2022.
- *Statistical analysis of reliability and survival data*, Master in statistiek, Catholic University of Leuven, Leuven, Feb-June 2023.
- *HELICONference (webinar)*, Sciensano, 1 Feb 2023.
- *Open science for pandemic preparedness*, BY-COVID project, Brussels, 6 Dec 2023.
- *Hack4Health R Hackathon*, Sciensano, Brussels, 20 Dec 2023.
- *BY-COVID Fest: Training in data sharing and reuse under GDPR*, BY-COVID project , Athens, 23-24 Jan 2024.
- *Using containers as *SPEs in federated research architectures for health data reuse (as co-lecturer) (webinar)*, BY-COVID project, 27 Feb 2024.
- *R introduction course (as a co-lecturer)*, Sciensano, Brussels, Jun 2024.
- *BY-COVID Spring 2024 Baseline Use Case Workshop: Integration of individual-level socioeconomic data for infectious diseases research and prevention in Europe (as a co-lecturer) (webinar)*, BY-COVID project, 6 Jun 2024.

List of abbreviations

AG	Antigen
AI	Artificial Intelligence
ARDS	Acute Respiratory Distress Syndrome
ATE	Average Treatment Effect
BY-COVID	Beyond-COVID
CDM	Common Data Model
CESSDA	Consortium of European Social Science Data Archives
CHS	Clinical Hospital Survey
CI	Confidence Interval
CInc	Cumulative incidence
CoBRHA	Common Base Registry for Healthcare Actor
COVID-19	Coronavirus disease 2019
Ct	Cycle threshold
DAG	Directed Acyclic Graph
DataSHIELD	Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual-level Databases
DMP	Data Management Plan
DPA	Data Protection Authority
DQA	Data Quality Assessment
EEA	European Economic Area
ECDC	European Centre for Disease Prevention and Control
EDA	Exploratory Data Analysis
EHDS	European Health Data Space
EHR	Electronic Health Records
EIF	European Interoperability Framework
EMA	European Medicines Agency
ESeC	European Socio-economic Classification

EU	European Union
FADP	Federal Act on Data Protection
FAIR	Findable, Accessible, Interoperable, and Reusable
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GDI	Genomic Data Infrastructure
GÖG	Gesundheit Österreich GmbH
HDA	Health Data Agency
HIP	Health Information Portal
HIPAA	Health Insurance Portability and Accountability Act
HR	Hazard Ratio
IACS	Instituto Aragonés de Ciencias de la Salud
ICU	Intensive Care Unit
IMA	Intermutualistic Agency
IPW	Inverse Probability Weighting
ISCED	International Standard Classification of Education
IVAC	International Vaccine Access Center
JA-InfAct	Joint Action Information for Action
KNAW	Royal Netherlands Academy of Arts and Sciences
KNN	K-nearest neighbors
LOST	Legal, Organizational, Semantic and Technical
MAR	Missing at random
MCMC	Markov chain Monte Carlo
ML	Machine Learning
NGS	Next Generation Sequencing
PHIRI	Population Health Information Research Infrastructure
PHT	Personal Health Train
PICO	Patient/Population, Intervention, Comparison/Control, and Outcome
RO-Crate	Research Object Crate
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RCT	Randomized Controlled Trial
RD	Risk difference
REC	Research Ethics Committee
RMST	Restricted Mean Survival Time
RMSTD	Restricted Mean Survival Time Difference
RR	Risk ratio
RT-PCR	Reverse transcription polymerase chain reaction
RWD	Real-world data
SAIL Databank	Secure Anonymized Information Linkage Databank
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SES	Socio-economic status
SGTF	S gene target failure
SMD	Standardized Mean Differences
SNP	Single Nucleotide Polymorphism
SPE	Secured Processing Environment
SSH	Social Sciences and Humanities
T ₀	Time zero
THL	Finnish Institute for Health and Welfare
TTP	Trusted Third Party
VE	Vaccine Effectiveness
VEBIS	Vaccine Effectiveness, Burden and Impact Studies
VOC	Variant Of Concern
VR	Virtual Repository
WGS	Whole Genome Sequencing
WHO	World Health Organization

Table of contents

JURY	V
ACADEMIC RESUME	VII
LIST OF ABBREVIATIONS	XI
 <u>SECTION I: INTRODUCTION, OBJECTIVES AND OVERVIEW OF THE THESIS</u>	 1
 <u>SECTION II: INTEGRATION OF REAL-WORLD DATA FOR PUBLIC HEALTH RESEARCH IN BELGIUM</u>	 29
CHAPTER 1: EVALUATING METHODOLOGICAL APPROACHES TO ASSESS THE SEVERITY OF INFECTION WITH SARS- COV-2 VARIANTS	30
CHAPTER 2: IMPACT OF HOMOLOGOUS AND HETEROLOGOUS PRIME-BOOST VACCINATION ON THE CLINICAL SEVERITY OF SARS-COV-2 OMICRON INFECTIONS	71
 <u>SECTION III: INTEGRATION OF REAL-WORLD DATA FOR PUBLIC HEALTH RESEARCH ACROSS JURISDICTIONAL BORDERS</u>	 112
CHAPTER 3: FEDERATED CAUSAL INFERENCE BASED ON REAL-WORLD OBSERVATIONAL DATA SOURCES	113
CHAPTER 4: REAL-WORLD COMPARATIVE EFFECTIVENESS OF SARS-COV-2 VACCINATION CAMPAIGNS AGAINST SARS-COV-2 INFECTIONS	144
CHAPTER 5: EXPERTS' EXPERIENCES AND PERCEPTIONS TOWARDS THE INTEGRATION OF SOCIO-ECONOMIC AND GENOMIC DATA FOR PUBLIC HEALTH RESEARCH	192
 <u>SECTION IV: GENERAL DISCUSSION AND RECOMMENDATIONS</u>	 291

SECTION I

Introduction, objectives, and overview of the thesis

GENERAL INTRODUCTION

1. POLICY-RELEVANT PUBLIC HEALTH RESEARCH

1.1 Public health

Public health is the field of science that aims to protect and improve the collective health and well-being of populations [1]. Unlike the clinical field, which concentrates on the health of individual patients, public health has a population-level perspective [2]. These populations can range from local neighbourhoods to the population of an entire country or larger region in the world. *Public health research* aims to uncover health patterns, disparities, trends, and determinants of health within and across populations [3]. Improving public health can entail community or policy efforts for the detection, prevention (e.g., by promoting personal hygiene and healthy lifestyles, ensuring healthy environments, immunization), or treatment of diseases, and the prolongation and increase of the quality of life [4, 5]. The term *population health*, although sometimes distinguished as a sub-field of public health that targets groups defined by common characteristics within the entire population [6, 7], is generally used interchangeably, as it is throughout this thesis.

1.2 The role of public health research in shaping public health policy

Evidence-informed public health policy entails the use of scientific evidence and data to guide policies that advance population-wide health and well-being, by informing policy decisions and practice [8]. The spectrum of evidence that can be leveraged to guide policy is broad, ranging from surveillance statistics and results from descriptive studies (e.g., capturing disease outbreaks, monitoring the spread of a disease), to economic evaluations (e.g., cost-benefit analyses), evidence from qualitative studies (e.g., by conducting focus groups or interviews), literature reviews (e.g., systematic reviews), non-interventional observational studies (e.g., cohort studies) and interventional studies like field trials (i.e., conducted out of the context of health care facilities under real-world conditions, with disease-free individuals in the general population) or randomized phase 3 clinical trials (RCTs, including explanatory or pragmatic trials) [8–11].

Empirical evidence from the *public health research* field naturally has a central role in shaping public health policy [8]. Public health research can entail studying correlations between variables, or can be extended to the complex task of inferring causality. While correlational research can be valuable for identifying correlations between various attributes and health outcomes and for determining areas for closer examination (e.g., identifying risk factors of health outcomes), evaluating causal relationships can particularly provide relevant insights for public health policy, enabling the identification of direct

drivers of health outcomes in populations, and thereby enabling the implementation of policies that can reliably improve public health [12, 13]. For example, establishing a causal link between the smoking of tobacco and lung cancer has been crucial in shaping large-scale tobacco control policies in the past, including an increase of taxation on tobacco products, enforcing the presence of warnings on cigarette packages, and the banning of smoking in public indoor areas [14, 15]. As opposed to the clinical field, where RCTs are commonly implemented to causally evaluate the efficacy of treatments or practices in a controlled environment, public health research generally occurs in real-life, uncontrolled settings, e.g., by conducting observational studies that aim to generate insights applicable to an entire population [11,16].

Many examples of evidence-informed public health policy can be found during the coronavirus disease 2019 (COVID-19) pandemic, such as decisions on the enforcement of non-pharmaceutical COVID-19 preventive measures like social distancing measures or regulations enforcing the wearing of face masks by taking into account information such as COVID-19 statistics (e.g., incidence rates, statistics on hospitalization occupation, death rates) or effectiveness estimates from the relationship between non-pharmaceutical interventions and health outcomes from previous epidemiological waves or epidemics [17]. The use of statistics on vaccination coverage or vaccine effectiveness estimates to outline vaccination strategies and address vaccine hesitancy [18, 19] provides another example.

Even though evidence from public health research can provide valuable information, policy-making should never to be ‘determined’ by scientific evidence (i.e., not following it in a strict sense), but simply ‘guided’ in an indirect way [20], accompanied by aspects like clinical expertise and expert opinions, principles of safety, person-centred care, timeliness, efficiency, and quality, and values of individuals receiving healthcare services. Further, it is crucial to consider the perspectives from diverse research fields, and the political and societal context to reach to policy decisions [21]. For example, after the initial COVID-19 waves in Belgium, with restrictive policies like school closures or lockdowns [22, 23], concerns related to the failure to consider the economic and psychological impacts of the crisis were raised [24].

2. DATA FOR PUBLIC HEALTH RESEARCH

Data-driven public health research plays a crucial role in providing evidence to inform public health policy. With continuous technological advancements (e.g., improved methods for database management, data storage capacity) and data becoming more abundant (e.g., routinely collected in many settings), it becomes more and more important to harness the power of data science to address

critical issues in society. A non-exhaustive overview of data for public health research is provided in Figure 1. In the next paragraphs, certain key concepts in this context are described.

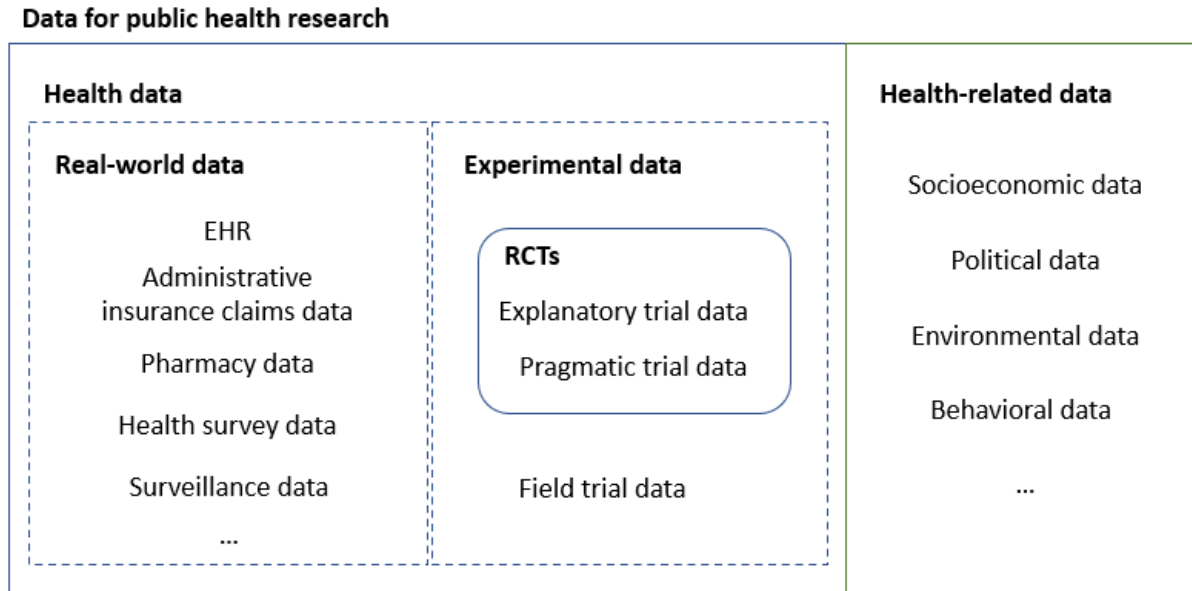


Figure 1. A non-exhaustive overview of data that can be leveraged for public health research. EHR: Electronic Health Records. RCT: Randomized Controlled Trial.

2.1 Health data

Health data are data related to individuals' health or well-being (e.g., health conditions, treatments, quality of life indicators) and have a key role in public health research. Health data include routine digital patient records obtained through interaction of individuals with the healthcare system (i.e., Electronic Health Records or EHR: vital patient statistics, contact information, individuals' medical history, diagnosis and treatment information, clinical visit and admission information, information on allergies, test results) [25, 26], administrative insurance claims data, health survey data, pharmacy data, repurposed data from clinical studies like RCTs, genomic and biomedical data, and census data [27].

2.2 Real-world data

Real-world data (RWD) constitute routinely collected data, like EHRs, administrative patient registries (e.g., billing data) or claims data [28,29]. As opposed to data from RCTs (i.e., interventional data), generally obtained from thoughtfully selected populations and under highly controlled and ideal conditions, RWD are collected in real-life, uncontrolled settings.

2.3 Primary versus secondary use of data

The *primary use of (health) data* entails the utilization of these data for the purpose that they were originally collected, such as the use of data gathered during routine clinical consultations for administration, or informing patient care and management (e.g., establishing a diagnosis, defining a treatment and follow-up plan) [30]. Alternatively, *the secondary use of (health) data* refers to the re-use of RWD for activities like research or policy development, extending the initial scope of their collection [30,31,32]. It involves for instance the repurposing of data generated as by-products of clinical care, without the need for extra data collection, thus leveraging existing data to uncover insights.

2.4 Privacy and data protection

Privacy and data protection are critical concerns in data-driven public health research, where the use of (sensitive) person health information is essential for generating insights that can improve population health. Robust data protection measures, such as encryption, secure data storage, and access control are essential to safeguard the privacy of individuals and prevent data breaches that could compromise the confidentiality of individuals [33,34]. Further, compliance with regulatory frameworks like the General Data Protection Regulation (GDPR) in Europe [35,36], the Federal Act on Data Protection (FADP) in Switzerland [37], or the Health Insurance Portability and Accountability Act (HIPAA) in the United States [38] is imperative to ensure that the right towards individuals' privacy are respected. Besides being compliant with privacy and data protection regulations (which are inherently based on ethical principles), research should also demonstrate compliance with ethical requirements [39,40]. It is not because research is considered legally permissible, that it can in itself be considered ethical [39].

The GDPR, instated in 2018, is a regulatory framework designed to protect natural persons with respect to the processing of personal data [35]. '*Personal data*' is defined as information related to a natural person that allows their identification (directly or indirectly). Only when processing personal data the GDPR applies, and compliance should be enforced [41]. It is also important to consider the category of '*sensitive personal data*', defined under Article 9 of the GDPR ('*Processing of special categories of personal data*') [35]. This can include information related to individuals' health, genetic information, individuals' racial characteristics, religious beliefs, sexual orientation or political opinions. Sensitive data are subject to higher protection levels due to their risk of compromising fundamental human rights [41].

In the context of the '*secondary use of personal data*' compliance with the GDPR has been found to be challenging, with this concept not being defined strictly within the GDPR. The closest related term can be seen as the '*further processing*' of personal data beyond the original purpose for which they were

collected [32,35]. However, as ‘further processing’ cannot unambiguously be considered as the ‘secondary use’ of data, there is a risk of non-compliance due to uncertainty [32]. New regulatory developments like the European Health Data Space (EHDS), enforced in April 2024, aim to clear these compliance uncertainties related to the secondary use of personal data by providing a more straightforward legal framework [42–44].

In the field of public health research, where data sharing across institutions may be required, ensuring data privacy while maintaining the utility of the data remains a complex but important challenge.

3. THE SECONDARY USE OF REAL-WORLD DATA IN OBSERVATIONAL PUBLIC HEALTH RESEARCH

3.1 Causal inference in public health research

As aforementioned, studies investigating causal relationships are particularly valuable to inform public health policies, by for example evaluating public health interventions (e.g., evaluating the effectiveness of past vaccination campaigns) or signaling targets for public health policy (e.g., identifying beneficial or hazardous determinants of health) [12]. The process of uncovering causal relationships is known as causal inference, where one estimates the variation in an outcome that results from changes in an exposure, essentially inferring probabilities of the outcome under changing conditions [45,46]. The central goal of causal inference is to move beyond associations and uncover actionable, cause-effect relationships [47]. In public health research, these investigations entail comparing health outcomes across alternative exposures, treatments, or interventions [48]. Contrasted groups can be referred to as ‘exposure groups’, ‘treatment groups’ or ‘intervention groups’, with these different expressions used interchangeably and depending on the context of the study. In the current section of the thesis, i.e., the general introduction, we will merely refer to ‘exposures’ and ‘exposure groups’ for convenience of use.

Under the *potential outcome framework*, the counterfactual element has a central role in the process of causal inference [47]. This framework defines potential outcomes under different exposure scenarios, with only one of these outcomes observed for a given individual (corresponding to the exposure they actually experienced) [47]. The outcomes under alternative exposures (i.e., ‘*counterfactual exposures*’) remain unobserved, and are commonly referred to as ‘*counterfactual outcomes*’ [47,48]. Conducting causal inference within the potential outcome framework involves counterfactual prediction (i.e., statistical inference of the distribution of an outcome under hypothetical alternative exposures). This approach enables the prediction of the outcome distribution

in hypothetical scenarios where the whole population is either exposed or unexposed to a particular factor. By comparing these predicted outcome distributions, researchers can assess causal effects of the exposure [49]. Making such predictions necessitates an understanding of the underlying data-generation processes, and cannot be derived solely from the data itself, without additional assumptions [45,46,50]. Formulating and explicitly specifying these causal assumptions underlying data analysis is an essential step in causal inference. These assumptions of researchers tend to be based on own experiences and knowledge, as well as existing evidence in literature, and is thus strongly reliant on the quality of such prior knowledge or evidence. Detailed theoretical information on causal inference can be found elsewhere [48].

RCTs, interventional studies where individuals are randomly assigned to contrasted exposure groups (i.e., a design-based method), are traditionally regarded as the ‘gold standard’ to infer causality [51–53]. However, as discussed by Fernainy et al. (2023) [53], both the pros and cons of RCTs should be considered. An advantage of RCTs includes the presence of well-established guidelines, with rigorous methodological and ethical standards [53]. Further, randomly assigning participants to an exposure group in ideal conditions ensures that missing values of counterfactual outcomes occur by chance [48], and results in groups that can be considered comparable or ‘exchangeable’, implying that any difference in the outcome between the groups can be attributed entirely to the exposure of interest and the measured association provides an estimate of the causal effect of interest (i.e., eliminating confounding bias and ensuring internal validity of the study under ideal conditions) [51,53,54]. However, realistically, results may still be inherent to biases such as selection or post-randomization bias (e.g., through the presence of missing data or loss to follow-up), which may be overlooked [47,53]. Further, in certain instances, conducting a RCT can be unfeasible (e.g., due to high costs or long time required to conduct a trial) or unethical, or researchers can be concerned with assessing relationships in realistic, real-world settings, considering larger, inclusive populations and applying less restrictive inclusion criteria, thereby enhancing the study’s external validity (i.e., the generalizability of obtained causal estimates to entire populations) [45,53,55]. The latter is especially relevant when studying relationships in complex systems, such as public health systems, where one needs to account for diverse populations, environmental factors, and societal contexts [53].

In contrast, observational studies, where researchers don’t have an active role in assigning an exposure (i.e., non-interventional), can provide an alternative to estimate causal relationships between exposures or interventions and health outcomes in uncontrolled, realistic settings (i.e., providing real-world evidence with an increased external validity) [27]. The secondary use of existing RWD presents an efficient and cost-effective way to conduct observational public health research [27,28]. In an era of big data, with the extended availability of observational data from diverse sources, observational

studies leveraging existing data have gained increasing popularity to estimate causal relationships [53]. However, since observational studies lack random exposure assignment, bias due the confounding factors (i.e., confounding bias, see section “*Potential biases in observational studies*” for more information) can be expected. Nonetheless, sophisticated data analysis approaches are to an increasing extent available to limit these potential biases, accurately conduct counterfactual prediction, and approach causal relationships of interests [27,45,50].

It is important to recognize that no single study design, neither RCTs or observational studies, is universally suited to address all causal research questions [48,53]. Each design is characterized by its own strengths, methodological challenges, and potential sources of bias [53]. To gain a more comprehensive and reliable understanding of a given phenomenon, it is generally advisable to apply multiple approaches, using alternative study designs, both interventional and observational [50,53]. By synthesizing findings from these complementary study methods, researchers can build a stronger, more nuanced evidence base [50].

Potential biases in observational studies

Observational RWD are subject to different biases that can arise in natural, uncontrolled settings, such as confounding, selection, and measurement bias [48,54]. These biases can influence the internal validity (i.e., the validity of the causal estimate) and external validity (i.e., the extent to which estimates are generalizable beyond the original study conditions and population) of the study results [52,56]. Observational studies lacking a high internal validity cannot readily establish a causal link between an exposure and health outcome [48]. Acknowledging and addressing biases is essential before feeding causal evidence into public health decision-making.

Confounding bias, caused by the presence of a common cause of the exposure and outcome [51], is a considerable challenge when estimating causal relationships in observational studies. Failure to account for confounding factors can result in biased estimates of the causal effect of interest, potentially over- or underestimating the true causal relationship by introducing spurious association (i.e., measured differences in the health outcome cannot entirely be attributed to the considered exposure, influencing the internal validity of the study results) [51,55]. In absence of randomization (i.e., by design), addressing confounding bias using statistical inference techniques can involve conventional stratification-based approaches like stratification or matching, or G-methods like regression standardization (i.e., modeling the outcome in relation to the exposure of interest and confounding covariates, and using this model to simulate the outcome across different levels of the exposure, hence enabling the prediction of unobserved counterfactual outcomes) or inverse probability weighting (IPW, i.e., modeling the probability of being exposed conditional on confounding

factors, with the inverse of this probability referred to as the propensity score, and individuals reweighted according to this propensity score), aiming to balance confounders and hence reach exchangeability between the exposure groups [51,55,57,58]. The large-scale development of machine learning techniques also presents with opportunities in the field of causal inference, offering the potential to improve methods for confounding adjustment, mitigate biases, and enhance the accuracy of causal estimands [45]. These methods assume that we have a set of measured covariates that is sufficient to adjust for confounding (i.e., removing all spurious association from the estimate), known as the assumption of conditional exchangeability [51]. As such, these methods generally rely on detailed individual information. To assess the presence of unmeasured confounding, the E-value can be calculated. This quantitative measure, first proposed by VanderWeele and Ding (2017) [59], expresses the robustness of an observed association to unmeasured confounding, and can be interpreted as the minimal associational strength (Risk Ratio scale) that an unmeasured confounder would need to exhibit with both the exposure and the outcome to completely account for a given exposure-outcome relationship, after adjusting for the measured covariates [53].

Selection bias arises from selection mechanisms for individuals to participate or to be included into the analysis [51,56,60–62], and results in a distinct relationship between exposure and health outcome between those participating and not participating in the study, or included and excluded from the analysis [61]. Hence, selection bias occurs when the study participants are not representative of the target population. Selection bias can be considered a form of *collider bias* when introduced by (inappropriately) conditioning on common effects of the exposure and outcome, i.e., a collider [63–65]. Selection bias can influence both the internal validity (by compromising exchangeability of exposure group) and external validity (by impairing generalizability of the results) of studies [61,62]. Examples of selection bias are among others non-response bias, loss-to-follow-up bias, and volunteer bias [48,56]. In cases where selection bias is unavoidable by design or at the stage of data collection, it can potentially be dealt with in the statistical analysis, correcting for it by using methods such as IP weighting [48,56].

Measurement bias (or information bias) refers to systematic errors introduced by incorrect measurement or classification of study data (e.g., the exposure variable, outcome variable, confounders) [48,61,66], thereby distorting the measured association and affecting the internal validity of the results. Different types of measurement error can be distinguished, according to the properties of independence and nondifferentiality, with most structures resulting in bias and with the type of error determining the potential method for correction [48,67]. Commonly, methods like using validation samples (i.e., subsets in which data are measured with no or limited error) and regression calibration are used to correct for measurement error [66,68].

Selection bias and measurement bias are not only of concern in observational studies, but should also be recognized and addressed in interventional designs (e.g., explanatory trials, pragmatic trials) [48].

Directed Acyclic Graphs

Directed Acyclic Graphs (DAGs), representing a widely used type of causal diagram, provide a valuable tool to visually represent complex hypothesized relationships between variables in causal pathways [69,70]. They can be constructed based on prior knowledge (e.g., about biological, political, or behavioral processes that are related to the considered research question) and explicitly display assumptions of researchers, thereby increasing transparency of studies approaching causal inference [71]. DAGs are composed of nodes that represent variables (i.e., exposure and outcome variables, potential confounders, selection factors) and edges (i.e., arrows) indicating the direction of the assumed causal relationships [71,72]. An arrow between two variables represents a hypothesis, but does not necessarily guarantee the observation of a relationship in the data. Variables of interest should be represented in the DAG, irrespective of whether they are measured [71]. Due to their reliance on prior knowledge and assumptions of researchers, there is a possibility of constructing multiple alternative DAGs, potentially resulting in meaningful differences in the obtained causal estimates.

Following some elementary principles, DAGs can assist researchers with the identification of different biases [73]. DAGs can play an important role in detecting potential confounding variables as common causes of the exposure and outcome, and determining a minimal set (or multiple possible sets) of variables required to control for in order to block confounding backdoor paths [48] that are potentially introducing spurious association between the exposure and outcome (i.e., a '*minimal sufficient adjustment set*', obtained by applying the *backdoor criterion*) [48,71]. Further, DAGs can help to represent and identify potential selection bias (introduced by conditioning on common effects of the exposure and outcome) [62,63] and measurement bias [67].

Target trial specification and emulation

As a way to minimize avoidable biases and enhance the validity of (causal) estimates from observational studies, Hernán et al. (2022) [74] suggested to design observational studies by emulating a target trial (i.e., a hypothetical pragmatic RCT that could alternatively have been conducted to estimate the causal relationship of interest). When conducting an observational study using the target trial emulation framework, the target trial is expected to be specified in a protocol, including different elements: eligibility criteria (specifying who would be included in the study), exposure strategy and assignment (i.e., random assignment of eligible individuals to the exposure arms), follow-up period

(starting at randomization up to a specified end point), defined outcome(s), causal estimand, analysis strategy [74–76]. The goal of researchers is to subsequently set up a properly designed observational study to emulate each element of the trial (e.g., defining time zero, the follow-up period, and inclusion/exclusion criteria, emulate randomization using statistical techniques to adjust for confounding) [74]. By providing these details in a target trial protocol, researchers increase transparency regarding their assumptions and decisions, and enhance confidence in the obtained results.

The specification of a target trial directs the data requirements and analytical procedures of the study. Mimicking randomized exposure assignment as in the target trial specification, generally requires the availability of detailed individual information to sufficiently correct for confounders using statistical procedures [76].

3.2 Integration of heterogeneous data sources

With RWD for public health research often scattered across diverse sources and different parties having distinct features of the same individual, i.e., a setting of *vertically-partitioned data* (see Figure 2), the integration of heterogeneous data sources from diverse domains is generally required to obtain detailed information on individuals, such as demographics, health, socio-economic status, and environmental exposure [77,78]. Integrating data from different sources would in this case serve to increase the number of features of individuals in the study population which are available for analysis. Combining these features is often required to gain a comprehensive understanding of health outcomes. Health outcomes tend to be multifactorial, with risks determined by diverse factors ranging from genetic predispositions to lifestyle factors, socio-economic position, and environmental conditions [79,80]. In addition, interactions between these determinants can exist. To have a comprehensive view on health outcomes, combining data on diverse health determinants is required.

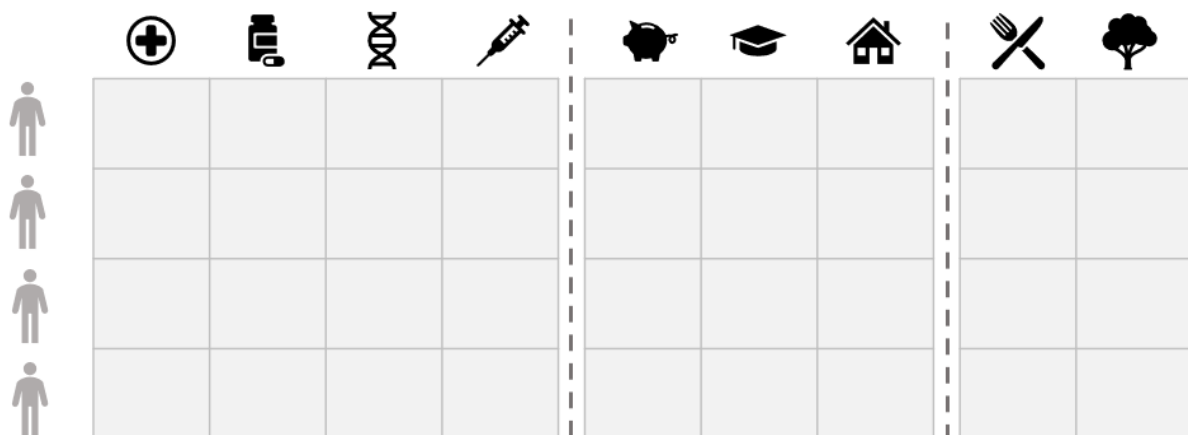


Figure 2. Graphical representation of vertical-partitioned data in the context of public health research, i.e., different parties having distinct features of the same individual.

RWD for public health research can also be horizontally partitioned, i.e., a setting with different parties having equivalent features of different individuals (see Figure 3) [77]. For example, this can occur when regional health departments or healthcare facilities maintain their own independent databases, containing data from distinct populations. Integrating *horizontally-partitioned data* can enlarge the study population (thereby enhancing statistical power), and can improve the diversity and representativeness of the dataset (by including individuals from diverse geographic locations, healthcare systems, or demographic groups) [78].

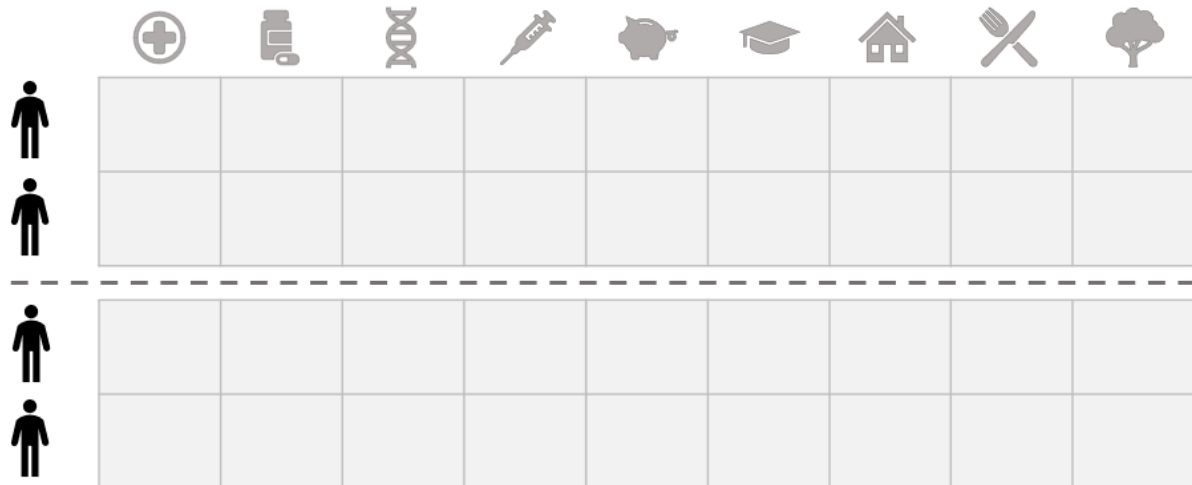


Figure 3. Graphical representation of horizontally-partitioned data in the context of public health research, i.e., different parties having equivalent features of the different individuals.

Integrating data from different sources could thus serve to increase the number of features of individuals in the study population which are available for analysis, but also the volume of individuals available for analysis [77]. By integrating these diverse data sources and co-analyzing them by adopting complex statistical methods, researchers can explore complex relationships in a more precise and accurate way. More specifically, it allows for the estimation of causal relationships between exposures and health outcomes by being able to account for potential biases (e.g., by correcting for various confounding factors) [81]. However, achieving the integration of heterogeneous real-world data sources for research requires addressing multiple challenges, like those related to privacy, data security, access restrictions, data quality, interoperability, and standardization [82,83]. In addition, approaching causal inference using observational data in an accurate way is challenging, requiring a rigorous correction for biases [48,84].

4. THE SECONDARY USE OF HETEROGENEOUS REAL-WORLD DATA FOR CROSS-JURISDICTIONAL BORDER RESEARCH

The secondary use of heterogeneous RWD for cross-jurisdictional border research, leveraging data from diverse regions and sources (i.e., a setting of both *horizontally-partitioned*— see Figure 3, and *vertically-partitioned data* – see Figure 2) [85], plays a pivotal role in addressing cross-border, European or global health challenges, by allowing researchers to generate new insights at a broader scale, producing more precise and generalizable results. When working across jurisdictional borders, with siloed data sources distributed across these borders, dealing with challenges related to standardization and interoperability, and careful navigation to ensure compliance with the GDPR becomes even more challenging [85–88].

4.1 Interoperability

Interoperability refers to capability of different systems to connect and exchange information, enabled by the implementation of standards, protocols, technologies and mechanisms to allow the flow between systems with minimal human intervention [89,90]. In public health research this more specifically entails the capacity of organizations, like national health care systems or hospital information systems (potentially located in different jurisdictions), to collaborate effectively in pursuit of shared objectives, involving the straightforward and secure exchange of information and knowledge across health systems and platforms, ensuring that health and other types of data can be accessed, shared, and integrated effectively [85]. The European Interoperability Framework (EIF) has been set up as a strategy to provide European public services in a seamless and interoperable way, giving direction through outlining fundamental interoperability guidelines as shared principles and recommendations to improve interoperability efforts, to set up connections across organizations and to streamline processes [91,92]. The EIF categorizes elements of interoperability into four layers, each of which must be addressed to achieve successful interoperable public service: legal, organizational, semantic, and technical (LOST) interoperability [91,93]. *Legal interoperability* refers to the alignment and collaboration across organizations operating under alternative regulations and policies (e.g., across jurisdictional borders of countries). *Organizational interoperability* relates to the adoption of aligned processes, governance, and responsibilities. *Semantic interoperability* represents the semantic and syntactic attributes of data, guaranteeing the preservation of the meaning of data and assuring a uniform understanding and correct interpretation across interested parties (e.g., through standardized formats and vocabularies). *Technical interoperability* relates to the technical aspects required for different systems to communicate, e.g., the use of standards, secure connections, and common technical specifications.

Proper health research integrating diverse data sources, distributed across systems and jurisdictional borders. interoperability governance according to the interoperability principles, following guidelines of the EIF, will facilitate public

4.2 The FAIR data principles

The *FAIR – Findable, Accessible, Interoperable, Reusable – data principles*, complementing the EIF, were formulated to promote standards that enhance data sharing and reuse in scientific research, encouraging more open, collaborative, and impactful public health research [94]. They are promoting to make data organized, and easily discoverable (Findable), accessible by means of standardized protocols (Accessible), compatible across sources and systems (Interoperable), well-documented, machine-readable and reusable across diverse systems (Reusable).

Various self-assessment tools or checklists are available to assist researchers in understanding the FAIR data principles and improving their adherence to these guidelines [95–97]. For instance, FAIR-aware offers interactive questions to assess researchers' knowledge of the FAIR data principles and provides practical information and tips for improvement [95,98]. Additionally, (semi-)automated tools like F-UJI [96,99], FAIR-Checker [100] or the FAIR Evaluation Services [101], are designed to programmatically evaluate the FAIRness of published research objects, to pinpoint areas where improvements can be made, and guide researchers in making their research objects FAIR-compliant [95].

4.3 Data integration across jurisdictional borders

Jurisdictional borders are characterized by laws, regulations, enforcement mechanisms and governance rules, which can differ from one jurisdiction to another [102]. In Europe, the centralization of personal data distributed across jurisdictional borders is generally precluded. Reasons behind this include the strict GDPR regulation and the existence of additional local laws and regulations, combined with concerns about data sovereignty (i.e., countries aspiring to preserve control over data that are generated within their borders, thereby being able to ensure that they are stored and processed in compliance with local laws) [103]. Further, the centralization of large volumes of personal data poses a security risk, potentially exposing (sensitive) personal information on a broad scale in the case of a data breach [104,105]. Transfer of personal data across jurisdictional borders must adhere to strict security frameworks, which make direct centralization challenging and legally complex [103]. In addition, aspects like compliancy uncertainty, lack of standardization, and limitations in data storage capacity can present barriers towards data centralization [106].

In the setting of data integration within and/or across jurisdictional borders, solutions alternative to a data pooling approach should be explored. Statistical meta-analyses, pooling estimates from

independent observational studies, can be performed [107–110], however, original observational studies might consider heterogeneous confounding factors, participant selection criteria, variable definitions, and statistical approaches, which may limit their comparability [86]. Alternatively, analyses to obtain estimates for meta-analysis can be guided by a common study protocol, thereby aiming to overcome challenges related to contrasting study characteristics, as demonstrated in the Vaccine Effectiveness, Burden and Impact Studies (VEBIS) study by Fontán-Vela et al. (2024) [111]. Federated analysis represents an alternative approach for the cross-border assessment of relationships in observational studies, ensuring compliance with privacy and data protection regulations and addressing interoperability challenges by adopting a reproducible methodology [85].

4.4 Federated analysis

Federated analysis enables concurrent data analysis across multiple sites or institutions, without the need to physically pool data into a centralized location [85]. This method entails data visiting, with queries or code for local reproducible analysis transferred to the premises of data controllers, analysis of data executed within local secured infrastructures, followed by the centralization of aggregated non-sensitive results [112,113]. This allows individual institutions to retain control of their (sensitive) personal data, thereby addressing privacy and data security concerns and instating compliance to the regulatory frameworks (e.g., GDPR) by design. By adopting a decentralized federated framework, researchers can gain important insights in public health matters, without migrating data or compromising data privacy and legal compliance.

Federated analysis can be distinguished from federated learning by differing in purposes and outcomes. Federated learning implies the training of a Machine Learning (ML) algorithm across distributed data, sharing local model parameters (e.g., gradients or weights) to generate an aggregated model [114,115]. Alternatively, federated analysis focusses on decentralized statistical inference, where analytical results are shared to enable drawing conclusions [115,116]. This can be compared to the ML inference phase, making calculations using the trained ML model. This thesis focuses on the federated analysis approach, with federated learning falling outside its scope.

When conducting federated analyses, interoperability challenges should be addressed. Interoperability in the context of federated public health research is regarded as the ability to mobilize coherent data from different sites, and the capacity to reproduce analyses across these sites, allowing the allocation of comparable analysis results [85]. In previous work, González-García et al. (2021) [85] proposed a methodology and offered recommendations on how to address challenges that arise at the different layers of interoperability (LOST) in federated research, such as the use of a Common Data

Model (CDM, i.e., a standardized schema in which data entities, structure and format are well-defined) and established ontologies to ensure semantic interoperability.

Although federated analysis offers a privacy-preserving alternative to centralized data analysis, it also presents several challenges. This approach enables the execution of consistent analytical procedures across multiple sites, accounting for shared confounding factors. However, its success is contingent on the availability and quality of the necessary data within local sites, where variations in data availability and data quality of confounding variables can complicate the comparability of results. Effect sizes may differ in magnitude or even in direction across sites due to factors such as local population characteristics, data quality disparities, or contextual influences. Consequently, federated analysis requires careful interpretation to account for these site-specific factors. In addition, privacy risks remain a concern, such as the potential for privacy attacks where unauthorized access to private local data occurs. Cryptographic, anonymization or pseudonymization methods can be adopted to ensure privacy preservation in the care of data leakage [117]. Further, in these collaborative federated efforts ‘fairness’ problems can occur, e.g., when different partners gain more benefits from participating in the collaboration than others (i.e., challenges in collaboration fairness) or the lack of representation of population groups with certain characteristics in local populations (for example due to local data availability), introducing bias and limiting generalizability (i.e., challenges in performance fairness) [117].

Several technologies and infrastructures have been developed to support federated analysis [113]. *Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual-level Databases (DataSHIELD)* is an innovative technological solution that allows parallel analysis of distributed datasets [118]. The *Personal Health Train (PHT)* presents an alternative framework designed for federated data analysis, using the metaphor of ‘trains’ (constituting workflows maintained by researchers) traveling between ‘stations’ (constituting FAIR data sources maintained by trusted entities) on ‘tracks’ (constituting the interface between the train and stations maintained by experts in health informatics; determining, maintaining and verifying connection rules and technical requirements) to allow the analysis of distributed data without data migration [113,119]. Privacy preserving federated learning infrastructure for secure insight exchange (VANTAGE6) is another open source decentralized platform inspired by the PHT, providing tools for federated learning and analysis, dealing with both horizontally- and vertically-partitioned data and allowing for client-server communication in various programming languages [17,113]. These technologies collectively enhance the ability to conduct large-scale, multi-center studies, while addressing privacy concerns and regulatory requirements.

4.5 European initiatives towards the secondary use of real-world data

Several key European initiatives advanced or are actively advancing the secondary use of RWD to enhance public health research and evidence-informed decision-making in Europe. The *Joint Action Information for Action! (JA-InfAct) on Health Information* (2018-2021) aimed to build and strengthen a health information system infrastructure in Europe, by building capacity, developing tools, and providing political support to achieve improved data interoperability and cross-border collaboration [85,120]. JA-InfAct laid the groundwork for subsequent initiatives by identifying gaps and proposing solutions for improved collection and accessibility of data across Europe. Building on this foundation, the *Population Health Information Research Infrastructure (PHIRI)* (2020-2023) continued these efforts by implementing proposed solutions and laying the foundation for a sustainable research infrastructure for the structured exchange of health data for public health research and further enhancing collaboration between European stakeholders to generate data-driven insights into public health issues [121]. The European *BeYond COVID (BY-COVID) project* (2021-2024) aimed to improve pandemic preparedness by making data on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and other diseases openly accessible and useable for research and has put efforts into the standardization and integration of data from diverse sources [122], complementing the PHIRI project by addressing aspects of data reuse and infrastructure development.

The recently adopted *EHDS regulation* (2024) [42–44] further advances the secondary use of real-world health data for public health research, by defining procedures for standardized, secure, and regulatory-compliant data sharing among European Union or European Economic Area (EU/EEA) member states [32]. The ‘Towards the European Health Data Space’ (TEHDAS) joint action (2021-2023) has played a significant role in laying the groundwork for the EHDS by mapping countries’ preparedness for and expectations from the EHDS, and by raising stakeholder awareness through active dialogue [123-125]. Building on these efforts, the TEHDAS2 joint action (2024-2026) focuses on preparing for the harmonized adoption of mechanisms to enable the secondary use of health data across countries [126]. The HealthData@EU Pilot project (2022-2024) further supports these efforts by prototyping the EHDS infrastructure, testing the technical and governance framework for the secondary use of health data and cross-border data sharing [124,127,128].

These and other European initiatives collectively represent Europe's commitment towards enabling the leveraging of RWD for large-scale multi-center observational studies in a privacy-presenting and interoperable way, and more robust, data-driven public health strategies.

OBJECTIVES

RWD, collected routinely in vast amounts, are revolutionizing public health research and advancing evidence-informed public health policy. In a complex landscape, with data distributed across isolated data sources (i.e., vertically-partitioned data) and jurisdictional borders (i.e., horizontally-partitioned data), researchers are expected to provide actionable evidence on public health concerns in a timely and accurate manner. This often entails the inference of causal estimates in observational studies. To successfully conduct observational public health research and to obtain valid (causal) evidence, they need to find their way through legislations, take into consideration ethical and legal concerns related to privacy and data security, tackle interoperability challenges, and effectively address biases.

The main objective of this thesis is to **develop, describe, demonstrate, and evaluate methodological approaches to integrate heterogeneous real-world data sources at the Belgian and cross-jurisdictional border level**, enabling researchers to respond to emerging (causal) public health research questions in an accurate and timely way. This includes the detailed description of a methodological framework to develop federated cross-border studies attempting causal inference using distributed (sensitive) personal data. More specific objectives include: (1) to describe biases present in observational health and health-related data and propose methodological approaches to address them, (2) to demonstrate the integration on heterogeneous data sources for policy-relevant observational public health research in Belgium and Europe, and (3) to map experts' experiences and perceptions towards the integration of heterogeneous data sources for public health research. By addressing these aspects, this thesis pursues to provide insights that can enhance elements of pandemic preparedness and response strategies for potential future public health crises.

OVERVIEW OF THE THESIS

Conducting policy-relevant public health research based on (sensitive) individual-level routine observational data is particularly challenging in contexts with vertically-partitioned data, where the required information is distributed across different systems or institutions. [Chapter 1](#) and [Chapter 2](#) focus on demonstrating the integration of scattered data sources within Belgium's jurisdictional borders to generate causal evidence that can guide public health policy. Accuracy and generalizability of the evidence strongly relies on the quality of the leveraged data, assumptions that were formed, and the adopted methodological procedures. These first chapters can be viewed as case studies, conducted to evaluate adopted methodological approaches and addressing policy-relevant research questions as exemplary illustrations. In [Chapter 1](#), various methodological approaches to assess the severity of SARS-CoV-2 infections across different variants were evaluated, with an emphasis on identifying and mitigating biases arising from the study design, data analysis methods, underlying surveillance strategies, and established data infrastructure. [Chapter 2](#) builds on this by showcasing the effective integration of heterogeneous data sources to address another policy-relevant research question, namely assessing the impact of homologous and heterologous prime-boost vaccination on clinical severity of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients in Belgium. This analysis additionally reveals challenges and limitations, highlighting areas for future improvement.

When addressing policy-relevant public health research questions that cross jurisdictional borders, the complexity increases. In these situations, horizontally-partitioned data, where data sources are isolated within the jurisdictional borders of different regions or countries, further complicate the research process. [Chapter 3](#) outlines a methodological framework for federated causal inference, offering a step-by-step procedure to conduct observational studies approaching causal inference using Real-World Data (RWD) without compromising data privacy. Subsequently, as reported in [Chapter 3](#) and [Chapter 4](#), the methodological approach was showcased, by investigating the real-world effectiveness of primary vaccination campaigns in preventing SARS-CoV-2 infections. By finalizing this demonstration, we were able to highlight the feasibility of the process, while also exposing important limitations. Stemming from difficulties encountered in integrating pathogen genomic and socio-economic data for the demonstrated research question, we gathered experts' experiences and perceptions towards integrating socio-economic and genomic data for public health research in [Chapter 5](#), by conducting qualitative analyses. This work shed light on the barriers and opportunities surrounding integrating these data sources with other health-related data sources, offering practical recommendations for improving the process for cross-border public health research.

Table 1. Overview of the thesis objectives, expected outcomes and potential impact.

Objective:	Develop and describe methodological approaches to leverage, integrate and analyze heterogeneous RWD sources for public health research within and across jurisdictional borders	
Expected outcomes	Working methodologies	Prototyped workflows
Chapter(s)	Chapter 1, 3	Chapter 3
Expected impact	Provide insights and recommendations that can enhance elements of pandemic preparedness and response strategies for potential future public health crisis	
Objective:	Demonstrate and evaluate the methodological approaches to leverage, integrate and analyze heterogeneous RWD sources for observational public health research at the Belgian and cross-jurisdictional border level	
Expected outcomes	Scientific evidence	Preparedness
Chapter(s)	Chapter 1, 2, 4	Chapter 1-5
Expected impact	Expanding the evidence pool which can be consulted for decisive action	Identify targets for future action

REFERENCES

1. Azari R, Borisch B. What is public health? a scoping review. *Archives of Public Health*. 2023;81:86.
2. Choi BCK, King AS, Graham K, Bilotta R, Selby P, Harvey BJ, et al. Clinical public health: harnessing the best of both worlds in sickness and in health. *Health Promot Chronic Dis Prev Can*. 2022;42(10):440–444.
3. Tulchinsky TH, Varavikova EA. Chapter 2 - Expanding the Concept of Public Health. In: Tulchinsky TH, Varavikova EA, editors. *The New Public Health*. Third Edition. San Diego: Academic Press; 2014. p. 43–90.
4. Schneider M-J. *Introduction to Public Health*. Sixth edition. Burlington: Jones & Bartlett Learning; 2020.
5. Saranya R, Kathirvel S. Chapter 1 - Principles and approaches in public health practice. In: Kathirvel S, Singh A, Chockalingam A, editors. *Principles and Application of Evidence-based Public Health Practice*. Academic Press; 2024. p. 3–21.
6. Kindig D, Stoddart G. What Is Population Health? *Am J Public Health*. 2003;93(3):380–383.
7. Kindig DA. Understanding Population Health Terminology. *The Milbank Quarterly*. 2007;85:139–161.
8. Brownson RC, Fielding JE, Maylahn CM. Evidence-Based Public Health: A Fundamental Concept for Public Health Practice. *Annual Review of Public Health*. 2009;30:175–201.
9. Brownson RC, Chiqui JF, Stamatakis KA. Understanding Evidence-Based Public Health Policy. *Am J Public Health*. 2009;99(9):1576–1583.
10. Victora CG, Habicht J-P, Bryce J. Evidence-Based Public Health: Moving Beyond Randomized Trials. *Am J Public Health*. 2004;94(3):400–405.
11. Rothman K, Greenland S, Lash TL. Types of Epidemiological Studies. In: Rothman K, Greenland S, Lash TL, editors. *Modern epidemiology*. Third Edition. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 87-99.
12. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. *Annual Review of Public Health*. 2013;34:61–75.
13. Kühne F, Schomaker M, Stojkov I, Jahn B, Conrads-Frank A, Siebert S, et al. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. *Ger Med Sci*. 2022;20:Doc12.
14. Ruegg TA. Historical Perspectives of the Causation of Lung Cancer. *Glob Qual Nurs Res*. 2015;2:2333393615585972.
15. Bertollini R, Ribeiro S, Mauer-Stender K, Galea G. Tobacco control in Europe: a policy review. *Eur Respir Rev*. 2016; 25:151-7.
16. Barrett DH, Ortmann LW, Brown N, DeCausey BR, Saenz C, Dawson A. Public Health Research. In: H. Barrett D, W. Ortmann L, Dawson A, Saenz C, Reis A, Bolan G, editors. *Public Health Ethics: Cases Spanning the Globe*. Cham: Springer International Publishing; 2016. p. 285–318.

17. Lison A, Banholzer N, Sharma M, Mindermann S, Unwin HJT, Mishra S, et al. Effectiveness assessment of non-pharmaceutical interventions: lessons learned from the COVID-19 pandemic. *The Lancet Public Health*. 2023;8(4):e311–317.
18. Weintraub RL, Subramanian L, Karlage A, Ahmad I, Rosenberg J. COVID-19 Vaccine To Vaccination: Why Leaders Must Invest In Delivery Strategies Now. *Health Affairs*. 2021;40:33–41.
19. Finney Rutten LJ, Zhu X, Leppin AL, Ridgeway JL, Swift MD, Griffin JM, et al. Evidence-Based Strategies for Clinical Organizations to Address COVID-19 Vaccine Hesitancy. *Mayo Clinic Proceedings*. 2021;96(3):699–707.
20. Lohse S, Canali S. Follow *the* science? On the marginal role of the social sciences in the COVID-19 pandemic. *Euro Jnl Phil Sci*. 2021;11(4):99.
21. Head BW. Three Lenses of Evidence-Based Policy. *Australian Journal of Public Administration*. 2008;67:1–11.
22. Luyten J, Schokkaert E. Belgium’s response to the COVID-19 pandemic. *Health Econ Policy Law*. 7:37–47.
23. van Loenhout JAF, Vanderplanken K, Van den Broucke S, Aujoulat I. COVID-19 measures in Belgium: how perception and adherence of the general population differ between time periods. *BMC Public Health*. 2022;22:245.
24. Van den Cruyce N, Van Hoof E, Godderis L, Gerard S, Van Leuven F. The impact of Covid-19 on Belgian mental health care: A Delphi study among psychosocial health professionals, patients, and informal caretakers. *Humanit Soc Sci Commun*. 2021;8:1–9.
25. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. *J Med Syst*. 2018;42(11):214.
26. Cowie MR, Blomster JL, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol*. 2017;106:1–9.
27. Näher A-F, Vorisek CN, Klopfenstein SAI, Lehne M, Thun S, Alsalamah S, et al. Secondary data for global health digitalisation. *The Lancet Digital Health*. 2023;5(2):e93–101.
28. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*. 2022;22:287.
29. Zisis K, Pavi E, Geitona M, Athanasakis K. Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. *J Pharm Pharm Sci*. 2024;27:12302.
30. Raab R, Küderle A, Zakreuskaya A, Stern AD, Klucken J, Kaissis G, et al. Federated electronic health records for the European Health Data Space. *The Lancet Digital Health*. 2023;5(11):e840–847.
31. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14:1–9.
32. Becker R, Chokoshvili D, Comandé G, Dove ES, Hall A, Mitchell C, et al. Secondary Use of Personal Health Data: When Is It “Further Processing” Under the GDPR, and What Are the Implications for Data Controllers? *European Journal of Health Law*. 2022; 30(2):129-157.

33. Shi M, Jiang R, Hu X, Shang J. A privacy protection method for health care big data management based on risk access control. *Health Care Manag Sci.* 2020;23(3):427–442.
34. Vora J, Italiya P, Tanwar S, Tyagi S, Kumar N, Obaidat MS, et al. Ensuring Privacy and Security in E-Health Records. *2018 International Conference on Computer, Information and Telecommunication Systems (CITS)*. Colmar. 2018. p. 1–5.
35. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Apr 27 2016 [cited 5 Sep 2024]. Available: <http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng>
36. Lincke S. Complying with the European Union General Data Protection Regulation (GDPR). In: Lincke S, editor. *Information Security Planning: A Practical Approach*. Cham: Springer International Publishing; 2024. p. 311–22.
37. SR 235.1 - Federal Act of 25 September 2020 on Data Protection. Fedlex. [cited 5 Sep 2024]. Available: <https://www.fedlex.admin.ch/eli/cc/2022/491/en>
38. Edemekong PF, Annamaraju P, Haydel MJ. Health Insurance Portability and Accountability Act. StatPearls. Treasure Island (FL): StatPearls Publishing; 2024 [cited 5 Sep 2024]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK500019/>
39. Hayes B, Kuyumdzhieva A. Ethics and data protection [Internet]. European Commission (DG Research and Innovation); 2021 [cited 29 Sep 2024]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf
40. Rochel J. Ethics in the GDPR: A Blueprint for Applied Legal Theory. *International Data Privacy Law*. 2021;11(2):209–223.
41. Finck M, Pallas F. They who must not be identified—distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law*. 2020;10:11–36.
42. European Parliament. EU Health Data Space: more efficient treatments and life-saving research. 2024 [cited 26 Sep 2024]. Available: <https://www.europarl.europa.eu/news/en/press-room/20240419IPR20573/eu-health-data-space-more-efficient-treatments-and-life-saving-research>
43. European Commission. European Health Data Space. 2024 [cited 15 Aug 2024]. Available: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
44. Horgan D, Hajdich M, Vrana M, Soderberg J, Hughes N, Omar MI, et al. European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare. *Healthcare*. 2022;10(9):1629.
45. Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A. A Survey on Causal Inference. *ACM Trans Knowl Discov Data*. 2021;15(5):1–46.
46. Pearl J. An Introduction to Causal Inference. *Int J Biostat*. 2010;6(2):7.
47. Höfler M. Causal inference based on counterfactuals. *BMC Med Res Methodol*. 2005;5:28.

48. Hernán MA, Robins JM. *Causal Inference: What If*. 1st edition. Boca Raton: Chapman & Hall/CRC; 2020.
49. Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. *Eur J Epidemiol*. 2020;35(7):615–617.
50. Hammerton G, Munafò MR. Causal inference with observational data: the need for triangulation of evidence. *Psychological Medicine*. 2021;51(4):563–578.
51. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov*. 2009;6:4.
52. Steckler A, McLeroy KR. The Importance of External Validity. *Am J Public Health*. 2008;98:9–10.
53. Fernainy P, Cohen AA, Murray E, Losina E, Lamontagne F, Sourial N. Rethinking the pros and cons of randomized controlled trials and observational studies in the era of big data and advanced methods: a panel discussion. *BMC Proc*. 2024;18:1.
54. Grimes DA, Schulz KF. Bias and causal associations in observational research. *The Lancet*. 2002;359(9302):248–252.
55. Jager KJ, Zoccali C, MacLeod A, Dekker FW. Confounding: What it is and how to deal with it. *Kidney International*. 2008;73(3):256–260.
56. Smith LH. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. *Curr Epidemiol Rep*. 2020;7(4):179–189.
57. Fang Y. Two basic statistical strategies of conducting causal inference in real-world studies. *Contemp Clin Trials*. 2020;99:106193.
58. Bettega F, Mendelson M, Leyrat C, Bailly S. Use and reporting of inverse-probability-of-treatment weighting for multicategory treatments in medical research: a systematic review. *J Clin Epidemiol*. 2024;170:111338.
59. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-value. *Ann Intern Med*. 2017;167:268–275.
60. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. *Epidemiology*. 2004;15(5):615.
61. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection Bias and Information Bias in Clinical Research. *Nephron Clinical Practice*. 2010;115(2):c94–99.
62. Smith LH. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. *Curr Epidemiol Rep*. 2020;7:179–189.
63. Holmberg MJ, Andersen LW. Collider Bias. *JAMA*. 2022;327(13):1282–1283.
64. Hernán MA, Monge S. Selection bias due to conditioning on a collider. *BMJ*. 2023;381:p1135.
65. Digitale JC, Martin JN, Glidden DV, Glymour MM. Key concepts in clinical epidemiology: collider-conditioning bias. *J Clin Epidemiol*. 2023;161:152–156.
66. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*. 2016;9:211–217.

67. Hernán MA, Cole SR. Invited Commentary: Causal Diagrams and Measurement Bias. *Am J Epidemiol*. 2009;170(8):959–62.
68. Boe LA, Shaw PA, Midthune D, Gustafson P, Kipnis V, Park E, et al. Issues in Implementing Regression Calibration Analyses. *American Journal of Epidemiology*. 2023;192(8):1406–1414.
69. Suzuki E, Shinozaki T, Yamamoto E. Causal Diagrams: Pitfalls and Tips. *J Epidemiol*. 2020;30(4):153–162.
70. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. *Epidemiology*. 1999;10:37.
71. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*. 2022;142:264–267.
72. Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. *JAMA*. 2022;327(11):1083–1084.
73. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620–632.
74. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*. 2022;328:2446–2447.
75. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*. 2016;183 (8):758–764.
76. Fu EL. Target Trial Emulation to Improve Causal Inference from Observational Data: What, Why, and How? *Journal of the American Society of Nephrology*. 2023;34(8):1305.
77. Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. VANTAGE6: an open source priVAcY preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. *AMIA Annu Symp Proc*. 2021;2020:870–877.
78. Mihaylov I, Kańduła M, Krachunov M, Vassilev D. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biol Direct*. 2019;14:22.
79. Braveman P, Egerter S, Williams DR. The Social Determinants of Health: Coming of Age. *Annual Review of Public Health*. 2011;32:381–398.
80. Salgado M, Madureira J, Mendes AS, Torres A, Teixeira JP, Oliveira MD. Environmental determinants of population health in urban settings. A systematic review. *BMC Public Health*. 2020;20:853.
81. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*. 2016;113(27):7345–7352.
82. Heacock ML, Lopez AR, Amolegbe SM, Carlin DJ, Henry HF, Trottier BA, et al. Enhancing Data Integration, Interoperability, and Reuse to Address Complex and Emerging Environmental Health Problems. *Environ Sci Technol*. 2022;56(12):7544–7552.
83. Martin LT, Nelson C, Yeung D, Acosta JD, Qureshi N, Blagg T, et al. The Issues of Interoperability and Data Connectedness for Public Health. *Big Data*. 2022;10:S19–24.

84. Listl S, Jürges H, Watt RG. Causal inference from observational data. *Community Dentistry and Oral Epidemiology*. 2016;44(5):409–415.
85. González-García J, Estupiñán-Romero F, Tellería-Orriols C, González-Galindo J, Palmieri L, Faragalli A, et al. Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct. *Archives of Public Health*. 2021;79:221.
86. Bernal-Delgado E, Estupiñán-Romero F, Zeynep O, Lyons R, Ivanković D. Real world evidence for European health policy-making: making the case for interoperability. *European Journal of Public Health*. 2019;29:ckz185.819.
87. Rujano MA, Boiten J-W, Ohmann C, Canham S, Contrino S, David R, et al. Sharing sensitive data in life sciences: an overview of centralized and federated approaches. *Briefings in Bioinformatics*. 2024;25(4):bbae262.
88. Kalogirou V, Charalabidis Y. The European Union Landscape on Interoperability Standardisation: Status of European and National Interoperability Frameworks. In: Popplewell K, Thoben K-D, Knothe T, Poler R, editors. *Enterprise Interoperability VIII*. Cham: Springer International Publishing; 2019. p. 359–68.
89. Bates DW, Samal L. Interoperability: What Is It, How Can We Make It Work for Clinicians, and How Should We Measure It in the Future? *Health Serv Res*. 2018;53:3270-3277.
90. Reraei R, Chiew TK, Lee SP, Shams Aliee Z. Interoperability evaluation models: A systematic review. *Comput Ind*. 2014;65:1-23.
91. Kouroubali A, Katehakis DG. The new European interoperability framework as a facilitator of digital transformation for citizen empowerment. *Journal of Biomedical Informatics*. 2019;94:103166.
92. NIFO - National Interoperability Framework Observatory. The European Interoperability Framework in detail. [cited 6 Sep 2024]. Available: <https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/european-interoperability-framework-detail>
93. European Commission. Directorate General for Informatics. New European interoperability framework: promoting seamless services and data flows for European public administrations. Publications Office; 2017 [cited 8 Sep 2024]. Available: <https://data.europa.eu/doi/10.2799/78681>
94. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
95. Krans NA, Ammar A, Nymark P, Willighagen EL, Bakker MI, Quik JTK. FAIR assessment tools: evaluating use and performance. *NanoImpact*. 2022;272:100402.
96. Devaraju A, Huber R. An automated solution for measuring the progress toward FAIR research data. *Patterns*. 2021;2(11):100370.
97. Lia M, Colella DD, Longo A, Zappatore M. euFAIR: A Digital Tool for Assessing the FAIR Principles. In: Sales TP, Araújo J, Borbinha J, Guizzardi G, editors. *Advances in Conceptual modeling*. Cham: Springer Nature Switzerland. 2023. p. 49-58.
98. FAIR-Aware: Assess Your Knowledge of FAIR. FAIRsFAIR. 2020 [cited 12 Dec 2024]. Available: <https://www.fairsfair.eu/fair-aware>.
99. Devaraju A, Huber R. F-UJI – An Automated FAIR Data Assessment Tool. 2024.

100. Gaignard A, Rosnet T, De Lamotte F, Lefort V, Devignes M-D. FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. *J Biomed Semant.* 2023;14:7.
101. The FAIR Maturity Evaluation Service. <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#/>. Accessed 12 Dec 2024.
102. Mills A. Rethinking Jurisdiction in International Law. *British Yearbook of International Law.* 2014;84:187–239.
103. Molnár-Gábor F, Sellner J, Pagil S, Slokenberga S, Tzortzatou-Nanopoulou O, Nyström K. Harmonization after the GDPR? Divergences in the rules for genetic and health data sharing in four member states and ways to overcome them by EU measures: Insights from Germany, Greece, Latvia and Sweden. *Seminars in Cancer Biology.* 2022;84:271–283.
104. Srinivasan S. Compromises in Healthcare Privacy due to Data Breaches. 2016 [cited 2024 Sep 27]; Available from: https://core.ac.uk/outputs/236412348/?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1
105. Pool J, Akhlaghpour S, Fatehi F, Burton-Jones A. A systematic analysis of failures in protecting personal health data: A scoping review. *International Journal of Information Management.* 2024;74:102719.
106. Vukovic J, Ivankovic D, Habl C, Dimnjakovic J. Enablers and barriers to the secondary use of health data in Europe: general data protection regulation perspective. *Archives of Public Health.* 2022;80:115.
107. Petrelli F, Luciani A, Perego G, Dognini G, Colombelli PL, Ghidini A. Therapeutic and prognostic role of vitamin D for COVID-19 infection: A systematic review and meta-analysis of 43 observational studies. *The Journal of Steroid Biochemistry and Molecular Biology.* 2021;211:105883.
108. Dobbins M, Decorby K, Choi BCK. The Association between Obesity and Cancer Risk: A Meta-Analysis of Observational Studies from 1985 to 2011. *International Scholarly Research Notices.* 2013;2013(1):680536.
109. Roi-Teeuw HM la, Smeden M van, Bos M, Wilde SM de, Yang B, Rutten FH, et al. Estimated causal effects of common respiratory infections on cardiovascular risk: a meta-analysis. *Open Heart.* 2023;10(2):e002501.
110. Paul M, Leeflang MM. Reporting of systematic reviews and meta-analysis of observational studies. *Clin Microbiol Infect.* 2021;27(3):311–314.
111. Fontán-Vela M, Kissling E, Nicolay N, Braeye T, Evercooren IV, Hansen CH, et al. Relative vaccine effectiveness against COVID-19 hospitalisation in persons aged ≥ 65 years: results from a VEBIS network, Europe, October 2021 to July 2023. *Eurosurveillance.* 2024;29:2300670.
112. Casaletto J, Bernier A, McDougall R, Cline MS. Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer. *Annu Rev Genomics Hum Genet.* 2023;24:347–368.
113. Hallock H, Marshall SE, 't Hoen PAC, Nygård JF, Hoorne B, Fox C, et al. Federated Networks for Distributed Analysis of Health Data. *Front Public Health.* 2021;9: 712569.
114. Rafi TH, Noor FA, Hussain T, Chae D-K. Fairness and privacy preserving in federated learning: A survey. *Information Fusion.* 2024;105: 102198.

115. Casaletto J, Bernier A, McDougall R, Cline MS. Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer. *Annual Review of Genomics and Human Genetics*. 2023;24: 347–368.
116. Wang D, Shi S, Zhu Y, Han Z. Federated Analytics: Opportunities and Challenges. *IEEE Network*. 2022;36: 151–158.
117. Rafi TH, Noor FA, Hussain T, Chae D-K. Fairness and privacy preserving in federated learning: A survey. *Inf Fusion*. 2024;105:102198.
118. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*. 2014;43(6):1929–1944.
119. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. *Data Intelligence*. 2020;2:96–107.
120. Abboud LA, Bogaert P, Fehr A, Urbanski D, Tolonen H, Noguer-Zambran I, et al. The new Joint Action on Health Information: information for action (InfAct)! *European Journal of Public Health*. 2018;28:cky213.651.
121. Bogaert P, Schutte N. Towards a Population Health Information Research Infrastructure. *European Journal of Public Health*. 2021;31:ckab164.572.
122. BY-COVID. [cited 30 Mar 2023]. Available: <https://by-covid.org/>
123. Abboud L, Cosgrove S, Kesisoglou I, Bogaert P. Results from the TEHDAS country visits: State-of-play of 12 EU member states in relation to the EHDS2. *Eur J Public Health*. 2022;32 Supplement_3:ckac129.366.
124. Kessissoglou IA, Cosgrove SM, Abboud LA, Bogaert P, Peolsson M, Calleja N. Are EU member states ready for the European Health Data Space? Lessons learnt on the secondary use of health data from the TEHDAS Joint Action. *Eur J Public Health*. 2024;34(6):1102–1108.
125. Joint Action Towards the European Health Data Space – TEHDAS1. Tehdas. [cited 1 Jan 2025]. Available: <https://tehdas.eu/tehdas1/>.
126. Second Joint Action Towards the European Health Data Space – TEHDAS2. Tehdas. [cited 1 Jan 2025]. Available: <https://tehdas.eu/>.
127. Kotsareli K, Tsachouridis P. European Health Data Space: A New Era in EU Health. *HAPSc Policy Briefs Ser*. 2023;4:91–6.
128. Home - EHDS2 Pilot - Official website. [cited 1 Jan 2025]. <https://ehds2pilot.eu/>.

SECTION II

Integration of real-world data for public health research in Belgium

CHAPTER 1

Evaluating methodological approaches to assess the severity of infection with SARS-CoV-2 variants

Meurisse M, Van Oyen H, Blot K, Catteau L, Serrien B, Klamer S, Cauët E, Robert A*, Van Goethem N*. Evaluating methodological approaches to assess the severity of infection with SARS-CoV-2 variants: scoping review and applications on Belgian COVID-19 data. BMC Infectious Diseases. 2022;22:839. *Shared last author

ABSTRACT

Background: Differences in the genetic material of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants may result in altered virulence characteristics. Assessing the disease severity caused by newly emerging variants is essential to estimate their impact on public health. However, causally inferring the intrinsic severity of infection with variants using observational data is a challenging process on which guidance is still limited. We describe potential limitations and biases that researchers are confronted with and evaluate different methodological approaches to study the severity of infection with SARS-CoV-2 variants.

Methods: We reviewed the literature to identify limitations and potential biases in methods used to study the severity of infection with a particular variant. The impact of different methodological choices is illustrated by using real-world data of Belgian hospitalized COVID-19 patients.

Results: We observed different ways of defining coronavirus disease 2019 (COVID-19) disease severity (e.g., admission to the hospital or intensive care unit *versus* the occurrence of severe complications or death) and exposure to a variant (e.g., linkage of the sequencing or genotyping result with the patient data through a unique identifier *versus* categorization of patients based on time periods). Different potential selection biases (e.g., overcontrol bias, endogenous selection bias, sample truncation bias) and factors fluctuating over time (e.g., medical expertise and therapeutic strategies, vaccination coverage and natural immunity, pressure on the healthcare system, affected population groups) according to the successive waves of COVID-19, dominated by different variants, were identified. Using data of Belgian hospitalized COVID-19 patients, we were able to document (i) the robustness of the analyses when using different variant exposure ascertainment methods, (ii) indications of the presence of selection bias and (iii) how important confounding variables are fluctuating over time.

Conclusions: When estimating the unbiased marginal effect of SARS-CoV-2 variants on the severity of infection, different strategies can be used and different assumptions can be made, potentially leading to different conclusions. We propose four best practices to identify and reduce potential bias introduced by the study design, the data analysis approach, and the features of the underlying surveillance strategies and data infrastructure.

1. INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), causing coronavirus disease 2019 (COVID-19) and responsible for a worldwide public health crisis, evolves continuously via genetic changes (e.g., mutations), resulting in an expanding phylogenetic diversity [1]. It has been shown that differences in the genetic material of the virus may result in altered virulence characteristics [2–4]. These genetic changes can be detected with Next Generation Sequencing (NGS) techniques, used to sequence the entire genome (i.e., Whole Genome Sequencing, WGS) or targeted regions of the genome of the virus. Virus variants can be differentiated based on one or more genetic mutations. When a virus variant is already identified, it can also be detected using so-called presumptive genotyping methods, such as S-gene sequencing or PCR screening targeting specific single nucleotide polymorphisms (SNPs), insertions or deletions (i.e., probe detection). Continuous monitoring of circulating variants and assessment of emerging variants is the aim of a genomic surveillance system. The World Health Organization (WHO), in collaboration with the European Centre for Disease Prevention and Control (ECDC) and other partners, defines Variants Of Concern (VOCs) based on (1) the presence of genetic alterations that are expected to alter viral characteristics like transmissibility, immune escape, disease severity or effectiveness of diagnostic or therapeutic tools, and (2) the transmission that influences the prevalence in the population [5,6]. Since the beginning of the pandemic, different VOCs (Alpha, Beta, Gamma, Delta and Omicron) have been circulating [5,7]. These VOCs present an increased risk for public health and should be prioritized in public health research. A timely assessment of the disease severity caused by newly emerging variants is essential in order to mitigate the impact on the healthcare system through appropriate public health measures and to provide timely recommendations for policy making and healthcare (e.g., hospital surge capacity preparedness strategies).

Randomization of exposure to a certain SARS-COV-2 variant, as one would do in a randomized controlled trial (RCT), is not conceivable, making the assessment of severity of infection with variants reliant on observational data. These ‘real-world’ data sources often result from the secondary use of routine care data, such as registries, surveys, medical records, insurance claims data, or government databases. As opposed to data of randomized studies conducted under controlled conditions, observational data are generally full of interactions and biases [8]. However, to identify a causal relationship between SARS-CoV-2 variants (exposure) and COVID-19 disease severity (outcome), we need exposure groups that are similar on both known and unknown factors that may differ between subjects and affect the relationship between the exposure and outcome. This can be referred to as exchangeability of exposure groups (i.e., when the unexposed group is a good approximation for the disease experience of the exposed group had they not been exposed) [9]. When exchangeability of

exposure groups can't be obtained by design, it can still be pursued through statistical methods (e.g., adjustment, matching, inverse probability weighting), thereby mimicking randomization. However, assessing the severity of SARS-CoV-2 variants using observational data in a causal research framework is a challenging endeavor, on which guidance is still limited. Reliability of real-world evidence depends strongly on the quality of the data, assumptions on potential confounding and statistical procedures used.

In this manuscript, we describe potential limitations and biases that researchers can be confronted with when studying the severity of infection with SARS-CoV-2 variants. Further, we evaluate the impact of different methodological choices by analyzing data of Belgian hospitalized COVID-19 patients. Finally, best practices to identify and reduce potential bias introduced by the design of the study and the data analysis approach, or related to the underlying surveillance strategies and data infrastructure, are proposed.

2. METHODS

A scoping review was conducted to (1) summarize methodological approaches used to study the severity of SARS-CoV-2 variants within different countries using observational data and (2) identify limitations and potential biases resulting from the study design, data analysis approach, underlying surveillance strategies, or data infrastructure. An electronic search was conducted in the PubMed database for the period of the 1st of March 2020 until the 22nd of June 2022, using the following search string: ("SARS-CoV-2 variants" [Supplementary Concept] OR "SARS-CoV-2 variant*" [TIAB] OR "Severe Acute Respiratory Syndrome Coronavirus 2 Variant*" [TIAB] OR "SARS-CoV-2 lineage*" [TIAB] OR "alpha variant"[TIAB] OR "delta variant"[TIAB] OR "beta variant"[TIAB] OR "gamma variant" [TIAB] OR "omicron variant" [TIAB]) AND ("Virulence" [Mesh] OR "virulence" [TIAB] OR "sever*" [TIAB] OR "pathogenic*" [TIAB] OR "death" [TIAB] OR "hospital*" [TIAB] OR "mortality" [TIAB] OR "fatal*" [TIAB] OR "complication*" [TIAB] OR "natural course" [TIAB] OR "Virulence" [Mesh] OR "Severity of Illness Index" [Mesh]) AND ("Observational Study" [Mesh] OR "observational stud*" [TW] OR "Cohort Studies"[Mesh] OR "cohort stud*" [TW] OR "Case-Control Studies" [Mesh] OR "case-control stud*" [TW] OR "Cross-Sectional Studies" [Mesh] OR "cross-sectional stud*" [TW] OR "data-linkage stud*" [TIAB]). We limited the search to English articles focusing on human populations. A first screening phase based on titles and abstracts was conducted [MM], and out-of-topic articles were excluded. A second screening stage based on the full texts was conducted in duplicate by two independent reviewers [MM, NVG]. Articles without a full text available were excluded, in addition to reviews, meta-analyses, non-research evidence (e.g., guidelines, websites, reports, policy documents letters), case reports and preprints. Studies not using observational data for their analyses, only focusing on vaccine

effectiveness, not doing statistical inference (only descriptive, no exposure groups), not considering COVID-19 disease severity as an outcome or SARS-CoV-2 variants as a risk factor were also excluded. Discrepancies between the two reviewers were solved by discussion. Additional articles of interest were identified by hand-searching and researching reference lists of selected articles. Information on the following aspects was extracted from included articles in tabular format: (1) first author, (2) year and (3) journal of publication, (4) DOI, (5) country in which the study was conducted, (6) SARS-CoV-2 variants under study (exposure groups), (7) exposure ascertainment (method for variant determination), (8) classification level of viral genomic variation, (9) outcome ascertainment (definition of disease severity), (10) study population, (11) study period, (12) reported potential selection biases following selection of samples for variant determination (13) other potential selection biases reported, (14) confounding factors taken into account, (15) other reported biases or challenges (not related to selection bias), and (16) main conclusion of the study.

The impact of different methodological choices or indications of the presence of biases as identified through the literature search were subsequently illustrated using example data originating from the COVID-19 surveillance in Belgium. A conceptual causal framework to evaluate the effect of SARS-CoV-2 variants on disease severity in a population of hospitalized patients, the corresponding Directed Acyclic Graph (DAG) explicitly stating causal assumptions, and data requirements were described elsewhere [10]. This previous work includes a description of the data infrastructure, allowing individual-level data linkage of selected variables from existing Belgian COVID-19 registries and compliance with the identified data requirements. The linkage of data on hospitalized COVID-19 patients [11], COVID-19 test results (including sequencing information) [12], administered COVID-19 vaccines, and socio-economic indicators was executed through the national registry number.

3. RESULTS

We identified 281 articles using the indicated search criteria. A total of 65 articles were selected through the first screening phase (title and abstract screening), 50 articles were finally included based on full text screening, and 6 additional articles were identified through hand-searching and searching of reference lists of included articles (Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram in Appendix A – Figure A1). The completed extraction form describing the characteristics of all included studies is available in Appendix B.

The results have been categorized into four sections highlighting different aspect of the methodological considerations when studying severity of infection with a SARS-CoV-2 variant. In ‘section I’, we describe various definitions of COVID-19 disease severity. In ‘section II’, we focus on the different SARS-CoV-2 variant exposure ascertainment methods. In ‘section III’, potential selection

biases are described. In ‘section IV’, we present a non-exhaustive list of factors fluctuating over time that need to be taken into account when studying successive waves of COVID-19 dominated by different variants.

Section I: defining COVID-19 disease severity

The severity of a COVID-19 infection can be defined based on several criteria and within different study populations. Within the general population, severity can, among others, be approached as a visit to a general practitioner or the emergency department [13,14], hospitalization [4,14–18], intensive care unit (ICU) admission [17–19] or death [18,20–22]. Other studies may consider the hospitalized population as the population group of interest and may define severity based on the need for ICU admission [2] or the occurrence of death [2,23], but also based on characteristics of the hospital care, e.g., respiratory or organ support [2,23]. A distinction can be made between, on the one hand, objective measures of disease severity, such as the monitoring of laboratory biomarkers and “hard” outcomes such as death, and, on the other hand, classifying disease severity based on the patient’s healthcare trajectory, which depends on contextual factors (e.g., clinical interpretations and decisions based on the patient’s condition or pre-existing comorbidities, therapeutic guidelines, or available medical care). The WHO proposed the WHO Clinical progression Scale, ranging from 0 to 10 and based on a patients progress through the healthcare system, as a measure for COVID-19 disease severity that can be used for a broad range of studies and enables comparability between studies [24].

The disease severity outcome definition may determine both the size and direction of the effect measure. Furthermore, risk factors may differ between alternative disease severity outcomes. For example, risk factors for severe complications, such as acute respiratory distress syndrome (ARDS) or death, may be different from those that determine whether a patient is admitted to the ICU. Nursing home residents for instance have characteristics increasing their risk for severe complications (e.g., old age, comorbidities), while clinicians might not be in favor of admitting these patients to the ICU as to prevent disproportionate care [25]. As such, it is important to specify in advance the disease severity outcome(s) of interest and to maintain awareness that conclusions for one outcome might not be transferable to another.

Section II: exposure ascertainment

Three approaches to define the exposure to a SARS-CoV-2 variant within observational studies were observed in literature: individual-level linkage through a unique identified of (1) WGS or (2) presumptive genotyping results, or (3) categorization of patients based on time periods.

Exposure ascertainment based on whole-genome sequencing

The exposure to a variant is ideally confirmed through WGS of the viral isolates obtained from the clinical sample and the subsequent linkage of the obtained sequencing results (e.g., the Pangolin lineage) with the patient data through a unique identifier. However, most countries do not exhaustively perform WGS on all COVID-19 positive samples, but only on a proportion of these samples (see section III), resulting in limited sample sizes when studying severity of infection with a variant. Moreover, a secured data infrastructure needs to be in place to enable the subsequent linkage of the sequencing results with clinical and epidemiological data on the individual patient-level. An example of this study set-up is the observational cohort study in Denmark conducted by Bager *et al.* [26], for which they linked SARS-CoV-2 genomic data with administrative Danish Health Registers in order to estimate the risk of hospital admission in individuals with the Alpha variant compared with those with other SARS-CoV-2 lineages. As such, the study population was restricted to the proportion of cases to which viral genome data could be linked. Given the high WGS capacity in Denmark, this corresponded to 60% of all individuals diagnosed with SARS-CoV-2 during the study period.

Exposure ascertainment based on presumptive genotyping methods

A second approach to define exposure to a variant is to use presumptive genotyping methods. For example, Wolter *et al.* [15] assessed the clinical severity of infections with the omicron variant in South-Africa using S gene target failure (SGTF) on the Thermo Fisher Scientific TaqPath COVID-19 PCR test as a proxy. Patients were classified into SGTF or non-SGTF exposure groups based on the identification of the His69_Val70del in the spike protein. However, this can only be accomplished when the TaqPath COVID-19 PCR test was used. Moreover, using SGTF as a proxy for Omicron infections should be limited to well-defined time periods as otherwise other variants also harboring the His69_Val70del, such as Alpha, may be misclassified as Omicron. Also, the Omicron BA.2 sublineage does not contain the His69_Val70del and is therefore not identifiable by SGTF.

Exposure ascertainment based on time periods

A third approach to define exposure to a variant is the categorization of patients based on time periods with known variant circulation. This approach does not require the linkage of viral genome sequencing results. Abdullah *et al.* [2] were able to rapidly report the decreased severity of COVID-19 disease in the Omicron-driven fourth wave by comparing the clinical profile of patients admitted at a large hospital in South-Africa. As an advantage, sample sizes are not limited by the linkage to non-exhaustive sequencing or genotyping results and the study population does not suffer from selection bias resulting from a non-random selection of samples (see section III) to determine the viral genomic profile.

However, accurately defining the exposure groups highly depends on a representative genomic surveillance system to accurately monitor circulating variants in place and time. This approach may lead to misclassification of variants, especially when the studied time periods are relatively close to each other or in periods with co-circulating variants, possibly diluting the observed effects [13].

Illustration of the impact of different exposure ascertainment methods

Belgian data of hospitalized COVID-19 patients were used to illustrate the impact of the different exposure ascertainment approaches. The clinical severity of infection with the Omicron variant as compared to the Delta variant has previously been assessed among hospitalized COVID-19 patients in Belgium [27], where severity was defined as being hospitalized for COVID-19 and either experiencing an ARDS event, and/or being admitted to ICU and/or in-hospital mortality. The main analysis was restricted to hospitalized patients registered in the Clinical Hospital Survey (CHS) and with a confirmed (based on WGS) or compatible (based on presumptive genotyping) Omicron or Delta infection as obtained through the linkage with the COVID-19 TestResult database. In addition, a sensitivity analysis was performed, only including patients with a WGS-confirmed Omicron or Delta infection [27]. Here, as an illustration, we apply the same analysis to a study population where the exposure groups were defined based on the date of hospital admission during restricted time periods with an estimated 100% circulation of the Omicron or Delta variant, respectively. These time periods were defined based on the representative baseline genomic surveillance in the Belgian general population [28]. The Delta exposure group was defined as those patients being admitted to the hospital between the 30th of August 2021 and the 14th of November 2021, while the Omicron exposure group was defined as those patients being admitted to the hospital between the 31st of January 2022 and the 28th of March 2022 (see Figure 1).

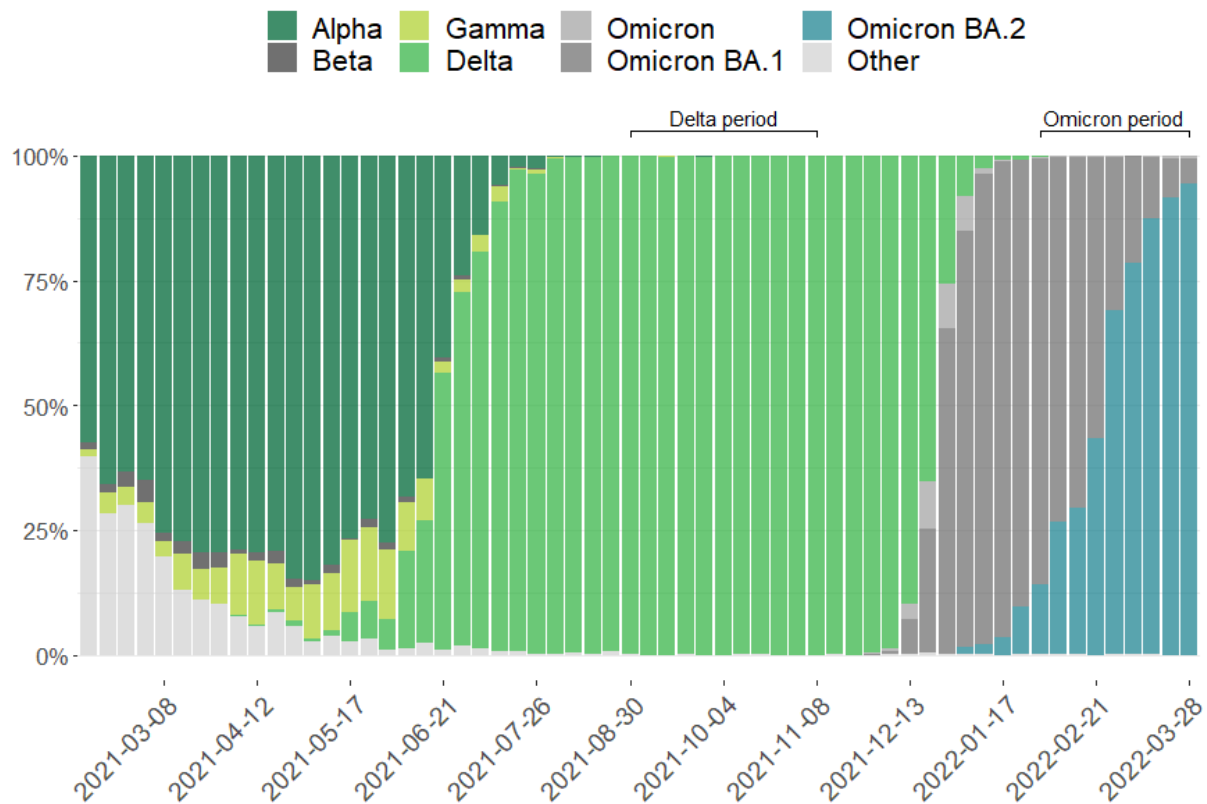


Figure 1. Share of SARS-CoV-2 variants of concern (VOC: Alpha, Beta, Gamma, Delta, Omicron (sub-lineage not specified), Omicron BA.1, Omicron BA.2, and other) per week (x-axis labels indicating the first day of the week) in Belgium as identified through the baseline genomic surveillance. An indication of the exposure groups (Omicron versus Delta variant) by time period is given: Delta period 30th of August – 14th of November 2021 and Omicron period 31st of January – 28th of March 2022.

All data analyses were conducted on the 11th of April 2022, which is almost five months after reporting the first confirmed Omicron case in Belgium. The sample sizes, distribution of cases over time and the obtained causal inference estimates for severe COVID-19 are presented in Figure 2 for each variant exposure ascertainment method.

The variant of infection was determined by WGS or genotyping for around 12% of the hospitalized COVID-19 patients registered in the CHS. Hence, the size of the study population is substantially reduced when defining exposure to a variant based on the linkage with viral genomic data. The exposure groups as defined based on the time period were compared with the exposure groups as defined based on the variant information obtained through linkage with the WGS or genotyping information from the COVID-19 TestResult database to identify exposure misclassification (see Table 1). For a limited number of patients categorized based on the time period, the obtained variant information through WGS or presumptive genotyping contradicted the exposure group, leading to potential differential exposure misclassification bias. For example, 0.4% of patients in the Delta exposure group as defined based on the time period were in reality infected with an Omicron variant.

Table 1. Contrasting, on the one hand, the SARS-CoV-2 variant exposure status as defined based on time periods with known variant circulation with, on the other hand, the SARS-CoV-2 variant exposure status as defined through linkage with Whole Genome Sequencing (WGS) and presumptive genotyping results in the COVID-19 TestResult database.

	Delta based on time period ^a	Omicron based on time period ^b
Conf. or comp. Delta^c	272 (12.0%)	1 (0.0%)
Conf. or comp. Omicron^c	8 (0.4%)	249 (11.5%)
Other variant	1 (0.0%)	1 (0.0%)
Unknown	1,983 (87.6%)	1,919 (88.4%)
Total	2,264	2,170

^a 30th of August – 14th of November 2021

^b 31st of January – 28th of March 2022

^c Conf. = confirmed through WGS. Comp. = compatible, based on presumptive genotyping.

When looking at the inference estimates as obtained from the analyses executed with the three different approaches for exposure status ascertainment, we observed the same direction of the effect, i.e., the risk for severe COVID-19 (experiencing an ARDS event, or being admitted to the ICU or in-hospital mortality) was found to be lower for hospitalized COVID-19 patients when infected with the Omicron variant, compared to hospitalized COVID-19 patients when infected with the Delta variant (see Figure 2). However, the effect size obtained in the study population selected based on the time periods (RR = 0.48) differs from the effect sizes obtained within the study population for which the exposure status was defined based on WGS (RR = 0.66) and/or presumptive genotyping (RR = 0.63). The confidence interval (CI) of the effect size obtained in the study population selected based on the time periods (CI = [0.38; 0.58]) is also narrower than those of the effect sizes obtained through the other two approaches (CI = [0.29; 1.03], CI = [0.30; 0.97]), indicating a higher precision of the former estimate. Furthermore, the estimated standardized risk of severe COVID-19 differs between the different approaches (see Figure 2). It is important to consider the selected study population (depending on the exposure ascertainment method) across which the marginalization was done when interpreting the standardized risk estimates.



Figure 2. Illustration of different approaches for SARS-CoV-2 variant exposure status ascertainment: (1) exposure defined based on WGS, (2) exposure defined based on WGS or presumptive genotyping and (3) exposure defined based on time periods with known variant circulation. The different approaches are compared in terms of the obtained sample sizes, the inference estimates and their 95% confidence intervals (expressed as standardized risk and risk ratio) for severe COVID-19 (experiencing an acute respiratory distress syndrome (ARDS) event, being admitted to the intensive care unit (ICU), or in-hospital mortality), and the distribution over time of the number of hospitalized COVID-19 patients included in the analysis by week of diagnosis (7-day rolling average). Please note the scale differences in the y-axis in the distribution plots.

Section III: potential selection biases

Selection bias is any bias resulting from factors associated with the exposure and/or outcome affecting the selection of the study participants [29]. In the current context, selection bias can be introduced when the subset of samples selected for WGS or presumptive genotyping is not representative of the underlying population. To reduce this bias, an appropriate sampling strategy for sequencing or presumptive genotyping (e.g., ensuring representativeness based on the geographical location, patient demographics, and disease severity) is recommended [30,31]. However, reaching representative sample selection for sequencing or presumptive genotyping is often challenging in current surveillance programs.

Selection bias introduced by conditioning on viral load

Technical considerations related to the viral RNA abundance in samples exist when selecting samples for NGS. In many settings, the only samples routinely available for variant identification will be residual diagnostic samples [32]. The viral load changes dynamically over the course of the infection (highest in the first week following disease onset [33,34], with a more rapid decline among vaccinated individuals [35]) and between different tissues (highest in specimens from the lower respiratory tract [36]). As such, viral load measurements depend on the clinical specimen and on the timing of sampling in the disease trajectory [37,38]. Sequencing samples with a low viral load (high cycle threshold (Ct) value) is often not recommended due to a rapid drop in the success rate of obtaining a complete or nearly complete viral genome [39], especially when sequencing capacities are limited and cost-efficiency of genomic surveillance has to be taken into account. For example, in the UK, samples are considered positive when minimal two target genes are amplified with a Ct value of less than 37 when using the TaqPath™ COVID-19 RT-PCR, whereas they are selected for WGS when N gene or ORF1ab target Ct value is less than 30 [40]. Likewise, Coolen *et al.* [41] describe a cut-off Ct value of 30 for samples to be sequenced with a high SARS-CoV-2 genome coverage using a reverse complement PCR technique, while generally a RT-PCR test result is considered positive when multiple target genes have a Ct value less than 35-40 [37]. For S-gene targeted Sanger sequencing, similar viral load cut-offs have been reported [42]. It has been demonstrated that the viral load is associated with the infectiousness [38,43] and severity of the infection [37,44–48]. Thus, by selecting samples with a sufficiently high viral load for sequencing or PCR probe detection, we might be selecting samples of COVID-19 cases with a more severe SARS-CoV-2 infection, no longer representative of the target population. In addition, viral loads can differ relatively between different SARS-CoV-2 variants [49–53]. Teyssou *et al.* [49] for example observed higher viral loads for the Delta variant compared to the Beta and historical variants.

The assumptions on the data-generating process and the biological pathways are graphically represented using a DAG in Figure 3.A. Briefly, the causal assumptions encoded in DAGs can identify the variables (minimally sufficient adjustment sets) that we need to control for in order to eliminate confounding, as well as helping to recognize variables that, if controlled for, bias the analysis [54]. Conditioning (also referred to as ‘controlling’ or ‘adjusting’) on a variable can be achieved through either sample restriction, stratification, regression adjustment or matching to examine the association of exposure and outcome within levels of the conditioned variable [55]. More details on DAGs can be found elsewhere [54,56,57].

We assume that viral load is a mediator on the causal path between the SARS-CoV-2 variant and COVID-19 disease severity. By selecting COVID-19 cases with an identified SARS-CoV-2 variant of the infection, there is conditioning on a descendant of the viral load (i.e., the mediator), which means there is also conditioning on viral load itself. Conditioning on viral load blocks the indirect path between the exposure and outcome, taking away part of the causal effect of interest. The selection bias which is introduced in this way can be referred to as overcontrol bias [10,58]. Furthermore, there might be a common unmeasured cause (U) of viral load and COVID-19 disease severity (such as immunosuppression [59,60]). Colliders do not generate an unconditional association between the variables that determine it (in a DAG). As viral load is a collider (i.e., a common effect of the treatment – or variable associated with the treatment – and the outcome – or variable associated with the outcome) on the path ‘SARS-CoV-2 variant $\rightarrow$ Viral load $\leftarrow$ U $\rightarrow$ COVID-19 disease severity’, conditioning on viral load would open a backdoor path and introduce a spurious association between SARS-CoV-2 variant and COVID-19 disease severity. This is a form of endogenous selection bias, introduced by conditioning on an intermediate collider [58].

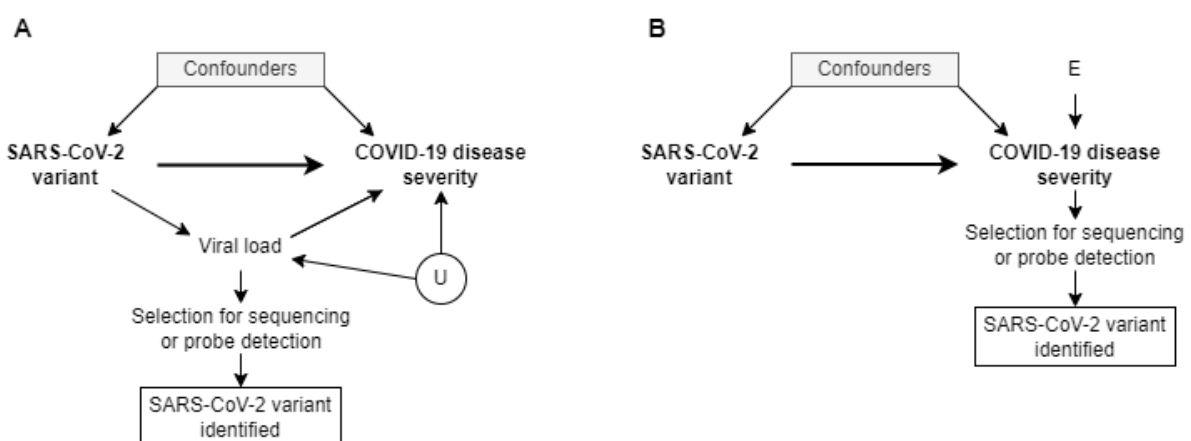


Figure 3. Directed Acyclic Graphs (DAGs) representing the potential selection biases introduced by the selection of COVID-19 cases by the availability of an identified SARS-CoV-2 variant of infection and (A) the selection of samples for variant identification based on the viral load in the sample, or (B) the selection of samples for variant identification based on the disease severity of a patient. Square nodes represent the conditioning on a variable,

while circular nodes represent unobserved or unknown variables. U: unmeasured confounders. E: error term. Based on a figure from Van Goethem et al. [10].

To limit this potential selection bias, in Denmark they aim to sequence the viral genome of all positive samples, with no limitation on the Ct value [61]. However, the chance of successfully obtaining a viral genome remains smaller for samples with a higher Ct value, and as such completely eliminating a selection of samples based on Ct values is not feasible with the current technical limitations. Furthermore, in periods with increased numbers of cases, Denmark's ambition has been shown not to be reachable and some restriction of samples based on the Ct value were still required [4,62,63].

Given the complex relationships that exist between the viral load, SARS-CoV-2 variants and severity of illness (e.g., a modifying effect by age [64]), it is difficult to assess the implications and consequences of selecting samples for sequencing based on the viral load for the subsequent epidemiological analyses.

Selection bias introduced by conditioning on COVID-19 disease severity

In most countries, exhaustive sequencing of eligible positive diagnostic samples remains challenging. In Europe, the percentage of sequences generated and shared of reported COVID-19 cases is 2.7% since the 10th of January 2020 [65] (as accessed on the 7th of April 2022), however large differences in sequence coverage are observed between countries [65,66]. When the sequencing of eligible samples is not exhaustive, the selection might be biased towards those samples of COVID-19 cases with specific characteristics. For example, samples of patients with a severe disease outcome might be preferentially selected for variant identification in a hospitalized cohort. When assuming that there are some factors (which can be taken together in an error term E) affecting COVID-19 disease severity (see figure 3.B), the outcome (COVID-19 disease severity) is a collider on the path between the exposure (SARS-CoV-2 variant) and the error term of the outcome (E). As such, selecting samples of COVID-19 patients with a severe disease outcome (i.e., selection on the outcome), opens a non-causal path 'SARS-CoV-2 variant → COVID-19 disease severity ← E'. As a result, the association between COVID-19 disease severity and SARS-CoV-2 variant does not represent the causal effect of interest, and the study's internal validity is affected. The bias introduced in this way is referred to as sample truncation bias, a form of endogenous selection bias introduced by conditioning on the outcome (or a descendent of the outcome). [58]. Furthermore, selecting patients based on certain patient characteristics (e.g., based on travel history or vaccination status) hampers the external validity or representativeness of the study, by inducing differences between the study population and the target population [67].

Illustration of study population selection and assessment of potential selection bias

The data of Belgian hospitalized COVID-19 patients with variant information results from different selection steps, as visualized in Figure 4. Firstly, diagnostic COVID-19 testing is based on national testing strategies, designed to set rules for prioritization of testing, and thus not at random (Figure 4.A). The testing strategy in Belgium has changed over time. For instance, testing of asymptomatic high-risk contacts was put on hold at the end of 2021, to decrease the pressure on laboratories [68]. This results in differential selection of patients based on their disease severity, which might change over time. Secondly, the study population is restricted to COVID-19 patients admitted to the hospital (Figure 4.B) and registered in the CHS (Figure 4.C). Again, a selection on the severity of disease in patients is performed, potentially resulting in collider bias and hampering both the internal and external validity of the study [69]. However, the advantage of working within this sub-cohort of hospitalized patients is the availability of detailed clinical information, allowing to adjust for important confounders in the study. Thirdly, for 9.8% of the Belgian hospitalized COVID-19 patients admitted after the 1st of March 2021 and registered in the CHS, the variant of infection was confirmed through WGS (Figure 4.D). However, this percentage fluctuates with time. For example, during the third and fourth epidemiological wave (15th of February – 27th of June 2021 and 4th of October – 26th of December retrospectively) the absolute number of hospitalized patients with variant information confirmed through WGS increases and the percentage coverage decreases, while during the interwave period in between the absolute number decreases and the percentage coverage increases (see Appendix C – Figure A2). The Belgian genomic surveillance consists of (1) a baseline surveillance, in which 5-10% of positive COVID-19 samples from sentinel laboratories are requested to be selected at random for WGS and (2) an active surveillance, where samples are requested to be selected for the detection of variants for patients with specific characteristics (i.e., patients with a travel history, re-infection, chronic infection or vaccination breakthrough infection) [70]. A random selection of samples to be sequenced, as in the baseline surveillance, limits potential selection biases. Hence, in this setting it is recommended to select patients with variant information available through baseline surveillance for the study population to assess severity of SARS-CoV-2 variants, which requires a correct reporting of the indication for selection for sequencing.

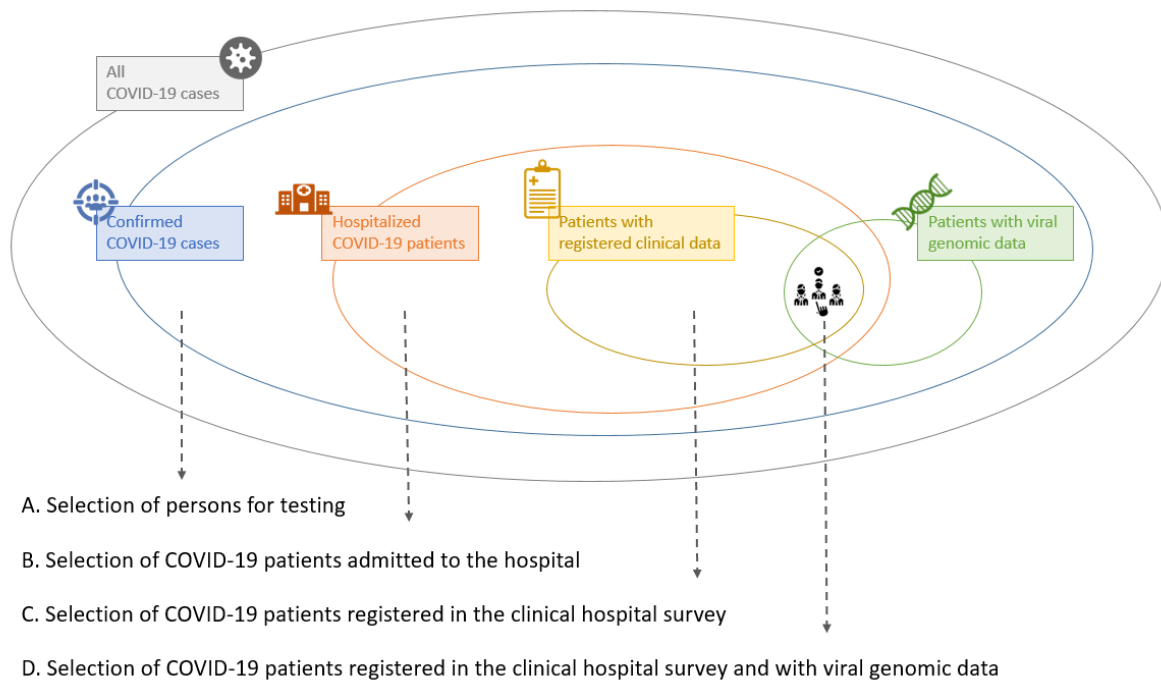


Figure 4. Overview of the different selection steps to obtain the study population of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Surveillance (CHS). This study population has been used to study the severity of SARS-CoV-2 variants. Selecting study participants can introduce selection bias.

There are multiple options to assess selection bias when studying the severity of SARS-CoV-2 variants. First, selection of study participants only introduces bias when associated with exposure and/or outcome. Bager *et al.* [4] studied the risk of hospitalization associated with infection with the SARS-CoV-2 Alpha variant compared to other co-circulating variants in Denmark, and they assessed the presence of selection bias by looking at, on the one hand, the relative risk of hospitalization among individuals without variant information relative to a reference group of non-Alpha strains (adjusted RR = 0.96 [0.90; 1.04]) and, on the other hand, the relative risk of hospitalization of individuals with variant info for the Alpha strain (adjusted RR = 1.23 [1.10; 1.38]) relative to the same reference group. They did not observe a strong association between the availability of variant information and hospitalization. Second, selection bias can be identified by comparing the baseline characteristics of COVID-19 cases with variant info to those without, assessing whether patients selected into the study are representative of the target population (i.e., checking the external validity). This was illustrated using the data of Belgian hospitalized COVID-19 patients, by comparing baseline characteristics between the patients admitted between 1st of March 2021 – 28th of March 2022 with confirmed (by WGS) variant information available through baseline surveillance and without any variant information available (neither confirmed by WGS, nor compatible, based on presumptive genotyping) (Table 2). Patients with confirmed variant information available through baseline surveillance are slightly older, more frequently have a non-European ethnicity and certain comorbidities compared to patients without variant information available. Furthermore, they were more frequently fully vaccinated with

a primary vaccination schedule, admitted to a university hospital and had a severe COVID-19 infection. As such, the study population, selected based on the availability of confirmed variant info through baseline surveillance, differs from and cannot be considered truly representative of the general hospitalized population (i.e., the population of interest). The higher frequency of patients with a severe disease outcome in the study population indicates selection on the outcome variable, and as such the presence of sample truncation bias. Furthermore, selection of samples of COVID-19 patients based on certain patient characteristics (e.g., age, vaccination status, comorbidities) might hamper the generalizability of results.

Table 2. Baseline characteristics of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Surveillance (CHS).

	Hospitalized patients (admitted between 1/3/2021 – 28/3/2022) with available variant information (confirmed by WGS ^a) from baseline surveillance (n = 1767)			Hospitalized patients (admitted between 1/3/2021 – 28/3/2022) without any available variant information (n = 28728)		
		%	n		%	n
Demographics						
Age (years), median (IQR)	69 (50-87)		1767	66 (45-80)		28728
Male gender, n (%)	945	53.5	1765	14776	51.5	28714
Nursing home resident, n (%)	109	6.2	1754	1698	6.0	28347
Ethnicity, n (%)						
European	559	78.2	715	10631	84.8	12535
North-African	86	12.0	715	1173	9.4	12535
Sub-Saharan African	35	4.9	715	322	2.6	12535
Asian	29	4.1	715	301	2.4	12535
Hispanic	6	0.8	715	108	0.9	12535
Comorbidities						
Cardiovascular Disease, n (%)	622	35.3	1763	8208	28.6	28663
History of Arterial Hypertension, n (%)	597	33.9	1763	9529	33.2	28663
Diabetes mellitus, n (%)	394	22.3	1763	5530	19.3	28663

Obesity, n (%)	211	12.0	1763	3478	12.1	28663
Chronic Pulmonary Disease, n (%)	310	17.6	1763	3916	13.7	28663
Chronic Neurological Disease, n (%)	207	11.7	1763	2349	8.2	28663
Chronic Cognitive Deficit, n (%)	175	9.9	1763	2400	8.4	28663
Chronic Renal Disease, n (%)	332	18.8	1763	3363	13.4	28663
Chronic Liver Disease, n (%)	54	3.1	1763	774	2.7	28663
Solid Cancer, n (%)	269	15.3	1763	2884	10.1	28663
Hematological Cancer, n (%)	77	4.4	1763	529	2.1	28663
Chronic Immunosuppression, n (%)	147	8.3	1763	622	2.2	28663
Socio-economic status						
Education level, n (%)						
Lower	249	25.8	965	3904	25.5	15299
Lower secondary	290	30.1	965	4644	30.4	15299
Higher secondary	270	28.0	965	4046	26.4	15299
Post-secondary higher education	156	16.2	965	2705	17.7	15299
Population density ^b , median (IQR)	1105 (419-2590)		1738	740 (332-1778)		25916
Median taxable income per capita ^c , median (IQR)	26733 (23807-28756)		1738	26787 (23807-28352)		25916
Exposure						
Place of infection, n (%)						
Community-acquired	1430	81.6	1752	24491	86.5	28302
Hospital-acquired ^d	231	13.2	1752	2302	8.1	28302
Nursing home-acquired	91	5.2	1752	1509	5.3	28302
Vaccination status						
Vaccination category, n (%)						

Not vaccinated	747	42.3	1767	15010	52.2	28728
Partially vaccinated	62	3.5	1767	1080	3.8	28728
Fully vaccinated	497	28.1	1767	6779	23.6	28728
Fully vaccinated + booster	461	26.1	1767	5859	20.4	28728
Disease characteristics						
Fever at admission, n (%)	376	50.9	738	6273	48.3	12983
Viral syndrome at admission, n (%)	322	43.3	744	5727	44.8	12789
Lower respiratory symptoms at admission, n (%)	520	68.6	758	8908	67.7	13162
Upper respiratory symptoms at admission, n (%)	69	9.6	716	1449	11.6	12528
Anosmia at admission, n (%)	56	8.0	703	1040	8.4	12341
CRP (mg/l) on admission, median (IQR)	95 (41-186)		136	102 (47-168)		1387
Hospital characteristics						
Type of hospital, n (%)						
General hospital	584	33.1	1767	21417	74.6	28728
General hospital with university character	700	39.6	1767	4982	17.3	28728
University hospital	483	27.3	1767	2329	8.1	28728
Mean ICU occupancy during hospital stay, median (IQR)	19 (12-32)		1767	27 (15-40)		28728
Outcomes						
Severe COVID-19 ^e , n (%)	440	25.2	1748	5840	20.7	28147
ICU admission, n (%)	298	16.9	1766	3325	11.6	28722
ICU transfer, n (%)						
General hospital	81	27.2	298	2228	67.0	3325
General hospital with university character	126	42.3	298	677	20.4	3325
University hospital	91	30.5	298	420	12.6	3325

In-hospital mortality, n (%)	230	13.2	1747	3309	11.8	28131
Invasive ventilation, n (%)	89	34.0	262	1089	42.5	2564
ECLS, n (%)	7	2.7	262	85	3.4	2535
Hospital length of stay (days), median (IQR)	9 (4-21)		1767	7 (3-15)		28727

^a Whole Genome Sequencing

^b Population density at the place of residence of the patient (postal code level).

^c Median net taxable income per capita at the place of residence of the patient (postal code level).

^d Diagnosis or onset of symptoms more than 8 days after hospital admission of the patient.

^e Experiencing an acute respiratory distress syndrome (ARDS) event, being admitted to the intensive care unit (ICU), or in-hospital mortality

Section IV: Factors fluctuating over time needed to be taken into account when studying successive waves of COVID-19 dominated by different variants

Different factors affecting the relationship between SARS-CoV-2 variants and disease severity can fluctuate over time, which is important to consider when studying variants not co-circulating in time. We discuss in more detail a non-exhaustive list of such factors fluctuating over time: (i) medical expertise, therapeutic strategies and guidelines, (ii) vaccination coverage and natural immunity, (iii) pressure on the healthcare system and (iv) affected population groups.

Level of medical expertise, therapeutic strategies and guidelines

First, the level of medical expertise and availability of clinical tools has increased throughout the epidemic. As an illustration, COVID-19 mortality scores developed at the pandemic's beginning are no longer accurate and need to be evaluated regularly according to the changing clinical context [71]. With the emergence of the first VOCs (Beta and Alpha) in the autumn of 2020 [6], clinical knowledge had already increased compared to the first months of the epidemic, with wild-type and other SARS-CoV-2 strains circulating [72–77]. Experience, practical guidelines and recommendations (e.g., for specific therapies) were quickly shared [77]. In the beginning of the epidemic, therapies existed mainly of repurposed therapeutics, but currently novel therapeutic options have been developed, among which antiviral medication (e.g., remdesivir, molnupiravir, paxlovid), anti-inflammatory drugs (e.g., dexamethasone) and monoclonal antibodies [78]. However, effectivity of therapeutics and recommended treatments might be altered with newly emerging variants [79,80]. Hence, when studying the severity of variants separated in time, taking into account evolving medical expertise, guidelines for treatment and availability of clinical therapies can be important to eliminate confounding bias. Data to adjust for this confounding bias are often lacking. Restricting the analysis to

a limited time period can increase the homogeneity in medical expertise and available therapeutics between the exposed and unexposed population.

Vaccination coverage and natural immunity

An unprecedented vaccine response was seen during the course of the epidemic. Large-scale vaccination campaigns have been undertaken across Europe, gradually increasing the vaccination coverage with a primary course to 72.6% [81] (as accessed on the 11th of April) in the total population of reporting countries within the EU and European Economic Area. Vaccination has been shown to be effective in preventing, to various degrees, a symptomatic infection [82–84], hospitalization [82,85–91], ICU admission [87] and death [82,89–91]. However, the effectiveness of vaccines can change when new variants emerge [92–94]. When comparing the severity of variants circulating in time periods with a different level of vaccine-induced immunity in the population, adjusting for the vaccination status of persons in the study population reduces confounding bias. In addition, different vaccine types, brands and regimens, introduced at different time points, may have an impact on vaccine effectiveness [95,96].

Next to vaccination, immunity can also be induced by a previous infection with a SARS-CoV-2 virus, referred to as natural immunity [97]. Similarly, the effectiveness of a previous infection to reduce the risk of a reinfection and severe disease outcomes can also differ between variants of reinfection [98,99]. Altarawneh *et al.* [98] observed that natural immunity was less effective in preventing a reinfection with the Omicron variant, characterized by specific mutations that might enable the virus to evade the immune system of an individual, compared to a reinfection with the Alpha, Beta and Delta variant. To remove confounding bias introduced by variable natural immunity, adjustment for an individual's previous infection can also be done. It is, however, only possible to adjust for documented previous infections, and depending on the testing strategy, which changes over time, only a fraction of the SARS-CoV-2 infections will truly be captured by routine surveillance (mainly symptomatic infections will be captured).

Due to limited vaccination coverage and natural immunity in the general population during the first year of the epidemic, confounding bias due to these factors might be limited when comparing variants circulating in this time period. However, with large-scale vaccination programs and increased virus circulation due to relaxation of COVID-19 measures after the first year of the epidemic, assessing severity of variants circulating during later time period requires adjustment for vaccination status and previous infection of individuals to obtain unbiased estimates. It is also important to consider the time since the last dose of vaccination and for reinfection, as vaccine effectiveness and natural immunity wane [99–102].

Pressure on the healthcare system

Factors affecting the pressure on the healthcare system, e.g., medical equipment stocks, surge of cases, hospital load and staff availability, vary over time. Different epidemiological waves, during which pressure on the healthcare system is high, follow each other and are separated by interwave periods with reduced pressure on the healthcare system. In periods of high pressure on the healthcare system, the delay between the onset of symptoms and diagnosis or hospital admission generally increases [103]. This delay has been shown to be positively associated with the severity of COVID-19 [104–106]. Additionally, hospital and ICU loads increase in epidemiological wave periods, with increased loads associated with increased disease severity and higher in-hospital mortality [107–110] of COVID-19 patients. Ideally, analyses should be adjusted for hospital load [111], ICU load [10,112] and COVID-19 incidence [111,113] to eliminate spurious association through these factors.

Affected population groups

Patient characteristics of the affected population, like age, gender, ethnicity, socio-economic status and nursing home residence, might also fluctuate over time depending on the outbreak dynamics. Within the hospital setting, as wards or ICUs overcrowd during epidemic waves, a clinical triage and admission of more severely ill patients also occurs, which can lead to a hospitalized patient population with different demographics and clinical outcomes depending on the epidemiological curve.

Illustration of the fluctuations of important risk factors over time

The fluctuation over time of important risk factors for severe COVID-19 was illustrated using the data from Belgian hospitalized COVID-19 patients. By plotting the vaccination status and the presence of a documented previous infection of these patients over time (see Appendix D-E – Figure A3-4), we could observe that the immunity of patients (vaccine-induced and natural) has increased over the course of the epidemic. However, the acquired immunity can wane with time or might be more efficiently evaded by certain variants. Furthermore, the percentage of nursing home residents and median age of hospitalized COVID-19 patients changes over time (see Appendix F-G – Figure A5-6), with in certain periods more vulnerable individuals (nursing home residents, individuals with a higher age) entering the population of hospitalized COVID-19 patients. These individuals are more likely to develop severe outcomes. Lastly, we can see that the ICU occupancy goes up and down with the different epidemiological waves (see Appendix G – Figure A6), with the height of the peaks also changing with time. When assessing severity of variants in the population of Belgian hospitalized COVID-19 patients, we took these different factors into account for the estimation of a causal effect in section II (Figure 2).

4. DISCUSSION

Studying the severity of newly emerging SARS-CoV-2 variants is important to assess the potential impact of these variants on patients' health, medical practice and the healthcare system, and to guide public health response. In this manuscript, we reviewed different methodological choices made and biases accounted for or reported in literature studying the causal effect of SARS-CoV-2 variants on COVID-19 disease severity using observational data, which in turn can lead to different conclusions. We described different definitions of COVID-19 disease severity, e.g., admission to the hospital or intensive care unit *versus* the occurrence of severe complications or death. Different ways to define exposure to a SARS-CoV-2 variant were also described, e.g., linkage of the sequencing or genotyping result with patient data through a unique identifier *versus* categorization of patients based on time periods. In a number of the included articles a combination of exposure ascertainment approaches was adopted. One included article reported an additional exposure ascertainment approach, namely, Fillâtre *et al.* [114] classified high-risk contacts of a COVID-19 case based on the identified variant of infection of this case. Furthermore, we presented potential selection biases (e.g., overcontrol bias, endogenous selection bias, sample truncation bias). Selection based on factors associated with the exposure and/or outcome might result in biased relative effect sizes when guidelines and procedures for sequencing change over time. Non-random selection of study participants also affects the external validity of a study, irrespective of a difference in selection procedures in the compared exposure groups. Lastly, a non-exhaustive list of time-dependent factors that might be present when studying the severity of variants was given (e.g., medical expertise and therapeutic strategies, vaccination coverage and natural immunity, pressure on the healthcare system per wave, affected population groups).

The scoping literature review shows the diversity of methodological decisions made, confounding biases taken into account, and selection biases present and reported. Based on the results of this literature review and experiences from the Belgian COVID-19 surveillance data, we present recommendations in the form of four best practices for making methodological decisions and limiting potential bias.

FOUR BEST PRACTICES WHEN STUDYING DISEASE SEVERITY OF SARS-COV-2 VARIANTS

1. Different choices in COVID-19 disease outcome measure might change the size and direction of the causal relationship. It is recommended to define the outcome under study to fit the public health impact that is targeted. For example, when assessing the impact of a new variant on the ICU load in hospitals, admission to ICU can be recommended as disease outcome. While, when

assessing the impact of a new variant on the primary care system, visits to a general practitioner might present the most suitable disease outcome. The size and direction of the effect of a variant on a particular COVID-19 disease outcome, as well as the considered confounding factors, may not be transferrable between different outcome measures.

2. The decision on an approach to define exposure to a SARS-CoV-2 variant should be made on a case-by-case basis depending on the context and setting. However, a representative baseline genomic surveillance, independent of any health outcome, is essential to make this decision and is strongly recommended to facilitate research on this topic.
 - a. When comparing co-circulating SARS-CoV-2 variants, defining exposure to a SARS-CoV-2 variant based on linkage with variant information is recommended, to prevent potential misclassification bias. When limiting the study population to patients with available variant information, a representative sample selection for WGS (or alternatively S-gene sequencing, or PCR probe screening) is necessary in order to limit selection bias (see best practice 3).
 - b. When comparing variants separated in time and with a close to 100% circulation (as obtained from representative genomic surveillance), and when the coverage of COVID-19 patients with available variant information is limited, defining exposure to a SARS-CoV-2 variant based on time of diagnosis is recommended. In this way, larger sample sizes can be obtained and external validity of the study findings can be increased. However, it is important to be aware of the potential misclassification bias. Furthermore, when comparing the severity of variants circulating in strictly separated time periods, it is important to take into account time-dependent factors in the analyses, such as the evolution of immunity status in the population.
 - c. The choice between (a) and (b) affects the timing of the analysis. Using linkage with variant information allows for a quick analysis based on the first patients when a new variant emerges and may be important to predict upcoming loads on the healthcare system. When the time between emergence of a variant and 100% circulation takes several weeks, an analysis may come too late to predict upcoming load on the healthcare system. The speed with which the analysis must be done should be balanced against sample size requirements and confounding/selection biases that can be controlled for.
3. A representative sampling of the study population (e.g., based on the geographical location, patient demographics, and disease severity) is essential for the external validity of the study and to allow inference of the effects for the target population. Guidance on representative genomic surveillance practices can be found elsewhere [115]. When representative sampling is not feasible, it is important to be aware of the potential presence of selection bias and to not make inferences

for target populations for which the study population is not representative. DAGs, graphically displaying the knowledge and assumptions about the causal relationship between exposure and outcome, provide a useful tool to identify potential selection biases.

4. When comparing the severity of variants circulating in different time periods, different important time-dependent risk factors need to be considered as potential confounders. It is recommended to determine, based on expert and continuously updating knowledge, which time-dependent factors might be different between exposure groups, and therefore might distort the association of interest.

5. DECLARATIONS

Availability of data and materials: Individual-level data of the Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Surveillance cannot be shared as open data, as they are too detailed to preserve patients' privacy. Data of the different registries within the LINK-VACC project (Clinical Hospital Survey, COVID-19 TestResult Database, StatBel, and Vaccinnet+) are stored in a secured environment provided by healthdata.be. Linking the individual-level data within this secured environment was possible with the use of a pseudonymized national reference number. Access to this secured healthdata.be environment was granted ad nominatum to scientists involved in Sciensano's surveillance activities. The selected data can be requested by external investigators by filling in the data request form (<https://epistat.wiv-isp.be/datarequest>). When of sensitive nature, the provision of the requested data has to be approved by the Belgian Information Security Committee Social Security & Health, and has to be outlined in a data transfer agreement (DTA) with the data owner (Sciensano).

Competing interests: The authors declare that they have no competing interests.

6. REFERENCES

1. Callaway E. Beyond Omicron: what's next for COVID's viral evolution. *Nature*. 2021;600(7888):204–207.
2. Abdullah F, Myers J, Basu D, Tintinger G, Ueckermann V, Mathebula M, et al. Decreased severity of disease during the first global omicron variant covid-19 outbreak in a large hospital in tshwane, south africa. *International Journal of Infectious Diseases*. 2022;116:38–42.
3. Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, et al. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. *The Lancet*. 2022;399(10332):1303–1312.

4. Bager P, Wohlfahrt J, Fonager J, Rasmussen M, Albertsen M, Michaelsen TY, et al. Risk of hospitalisation associated with infection with SARS-CoV-2 lineage B.1.1.7 in Denmark: an observational cohort study. *Lancet Infect Dis*. 2021;21:1507–1517.
5. WHO. Tracking SARS-CoV-2 variants [Internet]. 2022 [cited 13 Apr 2022]. Available: <https://www.who.int/activities/tracking-SARS-CoV-2-variants>
6. European Centre for Disease Prevention and Control. SARS-CoV-2 variants of concern as of 7 April 2022. In: European Centre for Disease Prevention and Control [Internet]. 7 Apr 2022 [cited 13 Apr 2022]. Available: <https://www.ecdc.europa.eu/en/covid-19/variants-concern>
7. Okpeku M. Possibility of COVID-19 eradication with evolution of a new omicron variant. *Infectious Diseases of Poverty*. 2022;11:30.
8. Straw S, Witte KK. Observational data during the COVID-19 pandemic: opportunity with uncertainty. *Heart*. 2020;106(19):1461–1462.
9. Hernán MA, Robins JM. *Causal Inference: What If*. 1st edition. Boca Raton: Chapman & Hall/CRC; 2020.
10. Van Goethem N, Serrien B, Vandromme M, Wyndham-Thomas C, Catteau L, Brondeel R, et al. Conceptual causal framework to assess the effect of SARS-CoV-2 variants on COVID-19 disease severity among hospitalized patients. *Archives of Public Health*. 2021;79:185.
11. Van Goethem N, Vilain A, Wyndham-Thomas C, Deblonde J, Bossuyt N, Lernout T, et al. Rapid establishment of a national surveillance of COVID-19 hospitalizations in Belgium. *Archives of Public Health*. 2020;78:121.
12. Meurisse M, Lajot A, Dupont Y, Lesenfants M, Klammer S, Rebolledo J, et al. One year of laboratory-based COVID-19 surveillance system in Belgium: main indicators and performance of the laboratories (March 2020–21). *Archives of Public Health*. 2021;79:188.
13. Wang L, Berger NA, Kaelber DC, Davis PB, Volkow ND, Xu R. Comparison of outcomes from COVID infection in pediatric and adult patients before and after the emergence of Omicron. *medRxiv*; 2022. p. 2021.12.30.21268495. Available: <https://www.medrxiv.org/content/10.1101/2021.12.30.21268495v1>
14. Twohig KA, Nyberg T, Zaidi A, Thelwall S, Sinnathamby MA, Aliabadi S, et al. Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. *The Lancet Infectious Diseases*. 2022;22:35–42.
15. Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. *The Lancet*. 2022;399(10323):437–446.
16. Dabrera G, Allen H, Zaidi A, Flannagan J, Twohig K, Thelwall S, et al. Assessment of mortality and hospital admissions associated with confirmed infection with SARS-CoV-2 Alpha variant: a matched cohort and time-to-event analysis, England, October to December 2020. *Eurosurveillance*. 2022;27(20):2100377.
17. Veneti L, Seppälä E, Storm ML, Salamanca BV, Buanes EA, Aasand N, et al. Increased risk of hospitalisation and intensive care admission associated with reported cases of SARS-CoV-2 variants B.1.1.7 and B.1.351 in Norway, December 2020 –May 2021. *PLOS ONE*. 2021;16(10):e0258513.

18. Fisman DN, Tuite AR. Evaluation of the relative virulence of novel SARS-CoV-2 variants: a retrospective cohort study in Ontario, Canada. *CMAJ*. 2021;193(42):E1619–1625.
19. Ulloa AC, Buchan SA, Daneman N, Brown KA. Estimates of SARS-CoV-2 Omicron Variant Severity in Ontario, Canada. *JAMA*. 2022;327(13):1286–1288.
20. Dabrera G, Allen H, Zaidi A, Twohig K, Thelwall S, Marchant E, et al. Assessment of Mortality and Hospital Admissions Associated with Confirmed Infection with SARS-CoV-2 Variant of Concern VOC-202012/01 (B.1.1.7) a Matched Cohort and Time-to-Event Analysis. Rochester, NY: Social Science Research Network; 2021 Mar. Report No.: 3802578. doi:10.2139/ssrn.3802578
21. Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ*. 2021;372:n579.
22. Grint DJ, Wing K, Williamson E, McDonald HI, Bhaskaran K, Evans D, et al. Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England, 16 November to 5 February. *Euro Surveill*. 2021;26(11):2100256.
23. Robinson ML, Morris CP, Betz J, Zhang Y, Bollinger R, Wang N, et al. Impact of SARS-CoV-2 variants on inpatient clinical outcome. *medRxiv*. 2022; 2022.02.02.22270337. doi:10.1101/2022.02.02.22270337
24. Marshall JC, Murthy S, Diaz J, Adhikari NK, Angus DC, Arabi YM, et al. A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases*. 2020;20(8):e192–197.
25. Borasio GD, Gamondi C, Obrist M, Jox R. COVID-19: decision making and palliative care. *Swiss Medical Weekly*. 2020;150:w20233.
26. Bager P, Wohlfahrt J, Fonager J, Albertsen M, Yssing Michaelsen T, Holten Møller C, et al. Increased Risk of Hospitalisation Associated with Infection with SARS-CoV-2 Lineage B.1.1.7 in Denmark. Rochester, NY: Social Science Research Network; 2021 Mar. Report No.: ID 3792894. Available: <https://papers.ssrn.com/abstract=3792894>
27. Van Goethem N, Chung PYJ, Meurisse M, Vandromme M, De Mot L, Brondeel R, et al. Clinical Severity of SARS-CoV-2 Omicron Variant Compared with Delta among Hospitalized COVID-19 Patients in Belgium during Autumn and Winter Season 2021–2022. *Viruses*. 2022;14(6):1297.
28. Maes P, Cuypers L, Baele G, Keyaerts E, Wollants E, Van Ranst M, et al. Genomic surveillance of SARS-CoV-2 in Belgium. National Reference Laboratory (UZ Leuven & KU Leuven); 2021 Jan pp. 1–12. Report No.: 3rd report for 2021. Available: https://assets.uzleuven.be/files/2021-03/genomic_surveillance_update_210121.pdf?utm_medium=email&utm_source=transaction
29. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection Bias and Information Bias in Clinical Research. *NEC*. 2010;115(2): c94–99.
30. European Centre for Disease Prevention and Control. Guidance for representative and targeted genomic SARS-CoV-2 monitoring. In: European Centre for Disease Prevention and Control [Internet]. 3 May 2021 [cited 15 Apr 2022]. Available: <https://www.ecdc.europa.eu/en/publications-data/guidance-representative-and-targeted-genomic-sars-cov-2-monitoring>
31. World Health Organization. SARS-CoV-2 genomic sequencing for public health goals: Interim guidance. 8 Jan 2021 [cited 15 Apr 2022]. Available: https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-genomic_sequencing-2021.1

32. World Health Organization (WHO). Genomic sequencing of SARS-CoV-2: A guide to implementation for maximum impact on public health. 2021 Jan. Available: <https://apps.who.int/iris/rest/bitstreams/1326052/retrieve>
33. Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect Dis*. 2020;20(4):411–412.
34. Zheng S, Fan J, Yu F, Feng B, Lou B, Zou Q, et al. Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January–March 2020: retrospective cohort study. *BMJ*. 2020;369:m1443.
35. Singanayagam A, Hakki S, Dunning J, Madon KJ, Crone MA, Koycheva A, et al. Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *The Lancet Infectious Diseases*. 2022;22(2):183–195.
36. Li C, Zhao C, Bao J, Tang B, Wang Y, Gu B. Laboratory diagnosis of coronavirus disease-2019 (COVID-19). *Clin Chim Acta*. 2020;510:35–46.
37. Shenoy S. SARS-CoV-2 (COVID-19), viral load and clinical outcomes; lessons learned one year into the pandemic: A systematic review. *World J Crit Care Med*. 2021;10(4):132–150.
38. Marc A, Kerioui M, Blanquart F, Bertrand J, Mitjà O, Corbacho-Monné M, et al. Quantifying the relationship between SARS-CoV-2 viral load and infectiousness. Cobey SE, Van der Meer JW, editors. *eLife*. 2021;10:e69302.
39. Chiara M, D’Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Briefings in Bioinformatics*. 2021;22(2):616–630.
40. Williams GH, Llewelyn A, Brandao R, Chowdhary K, Hardisty K-M, Loddio M. SARS-CoV-2 testing and sequencing for international arrivals reveals significant cross border transmission of high risk variants into the United Kingdom. *eClinicalMedicine*. 2021;38:101021.
41. Coolen JPM, Wolters F, Tostmann A, van Groningen LFJ, Bleeker-Rovers CP, Tan ECTH, et al. SARS-CoV-2 whole-genome sequencing using reverse complement PCR: For easy, fast and accurate outbreak and variant analysis. *J Clin Virol*. 2021;144:104993.
42. Jørgensen TS, Blin K, Kuntke F, Salling HK, Michaelsen TY, Albertsen M, et al. A rapid, cost efficient and simple method to identify current SARS-CoV-2 variants of concern by Sanger sequencing part of the spike protein gene. medRxiv; 2021 Mar p. 2021.03.27.21252266. doi:10.1101/2021.03.27.21252266
43. Marks M, Millat-Martinez P, Ouchi D, Roberts C h, Alemany A, Corbacho-Monné M, et al. Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. *The Lancet Infectious Diseases*. 2021;21(5):629–636.
44. Pujadas E, Chaudhry F, McBride R, Richter F, Zhao S, Wajnberg A, et al. SARS-CoV-2 viral load predicts COVID-19 mortality. *The Lancet Respiratory Medicine*. 2020;8(9):e70.
45. Fajnzylber J, Regan J, Coxen K, Corry H, Wong C, Rosenthal A, et al. SARS-CoV-2 viral load is associated with increased disease severity and mortality. *Nat Commun*. 2020;11:5493.

46. Knudtzen FC, Jensen TG, Lindvig SO, Rasmussen LD, Madsen LW, Hoegh SV, et al. SARS-CoV-2 viral load as a predictor for disease severity in outpatients and hospitalised patients with COVID-19: A prospective cohort study. *PLOS ONE*. 2021;16(10):e0258421.
47. Liu Y, Yan L-M, Wan L, Xiang T-X, Le A, Liu J-M, et al. Viral dynamics in mild and severe cases of COVID-19. *Lancet Infect Dis*. 2020;20(6):656–657.
48. Kwon J-S, Kim JY, Kim M-C, Park SY, Kim B-N, Bae S, et al. Factors of Severity in Patients with COVID-19: Cytokine/Chemokine Concentrations, Viral Load, and Antibody Responses. *Am J Trop Med Hyg*. 2020;103(6):2412–2418.
49. Teyssou E, Delagrèverie H, Visseaux B, Lambert-Niclot S, Brichler S, Ferre V, et al. The Delta SARS-CoV-2 variant has a higher viral load than the Beta and the historical variants in nasopharyngeal samples from newly diagnosed COVID-19 patients. *J Infect*. 2021;83(4):e1–3.
50. Kim J-M, Rhee JE, Yoo M, Kim HM, Lee N-J, Woo SH, et al. Increase in Viral Load in Patients With SARS-CoV-2 Delta Variant Infection in the Republic of Korea. *Frontiers in Microbiology*. 2022;13.
51. King KL, Wilson S, Napolitano JM, Sell KJ, Rennert L, Parkinson CL, et al. SARS-CoV-2 variants of concern Alpha and Delta show increased viral load in saliva. medRxiv. 2022; 2022.02.10.22270797. doi:10.1101/2022.02.10.22270797
52. Hay JA, Kissler SM, Fauver JR, Mack C, Tai CG, Samant RM, et al. Viral dynamics and duration of PCR positivity of the SARS-CoV-2 Omicron variant. medRxiv; 2022 Jan p. 2022.01.13.22269257. doi:10.1101/2022.01.13.22269257
53. Puhach O, Adea K, Hulo N, Sattonnet P, Genecand C, Iten A, et al. Infectious viral load in unvaccinated and vaccinated individuals infected with ancestral, Delta or Omicron SARS-CoV-2. *Nat Med*. 2022;28:1491–1500.
54. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*. 2022;142:264–267.
55. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*. 2008;8:70.
56. Glymour MM. Using Causal Diagrams to Understand Common Problems in Social Epidemiology. *Methods in social epidemiology*. Jossey-Bass/Wiley; 2006. pp. 393–428.
57. Suzuki E, Shinozaki T, Yamamoto E. Causal Diagrams: Pitfalls and Tips. *J Epidemiol*. 2020;30(4):153–162.
58. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annu Rev Sociol*. 2014;40:31–53.
59. Lahmer T, Erber J, Schmid RM, Schneider J, Spinner CD, Lupp P, et al. SARS-CoV-2 viral load dynamics in immunocompromised critically ill patients on remdesivir treatment. *Multidisciplinary Respiratory Medicine*. 2022;17:825.
60. Meiring S, Tempia S, Bhiman JN, Buys A, Kleynhans J, Makhasi M, et al. Prolonged Shedding of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) at High Viral Loads Among Hospitalized Immunocompromised Persons Living With Human Immunodeficiency Virus (HIV), South Africa. *Clinical Infectious Diseases*. 2022;75(1):e144–156.

61. de Villiers CB, Blackburn L, Cook S, Janus J, Johnson E, Kroese M. Next generation sequencing for SARS-CoV-2/COVID-19. PHG Foundation; 2021 Mar pp. 1–116. Available: https://www.finddx.org/wp-content/uploads/2021/05/2021_04_21_NGS-for-sars-cov-2-compr.pdf
62. Lyngse FP, Mølbak K, Skov RL, Christiansen LE, Mortensen LH, Albertsen M, et al. Increased transmissibility of SARS-CoV-2 lineage B.1.1.7 by age and viral load. *Nat Commun*. 2021;12:7251.
63. Danish Covid-19 Genome Consortium. Genomic overview of SARS-CoV-2 in Denmark. 4 Apr 2022 [cited 8 Apr 2022]. Available: <https://www.covid19genomics.dk/statistics>
64. Kim Y, Cheon S, Jeong H, Park U, Ha N-Y, Lee J, et al. Differential Association of Viral Dynamics With Disease Severity Depending on Patients' Age Group in COVID-19. *Frontiers in Microbiology*. 2021;12:712260.
65. Freunde von GISAID e.V. GISAID - Submission Tracker Global. [cited 7 Apr 2022]. Available: <https://www.gisaid.org/index.php?id=208>
66. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. *medRxiv*. 2021; 2021.08.21.21262393. doi:10.1101/2021.08.21.21262393
67. Findley MG, Kikuta K, Denly M. External Validity. *Annual Review of Political Science*. 2021;24:365–393.
68. European Centre for Disease Prevention and Control. Evaluation of the SARS-CoV-2 testing policy in Belgium from June to December 2021. In: European Centre for Disease Prevention and Control [Internet]. 14 Feb 2022 [cited 17 Apr 2022]. Available: <https://www.ecdc.europa.eu/en/publications-data/covid-19-evaluation-testing-policy-belgium-june-december-2021>
69. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun*. 2020;11:5749.
70. RAG subgroep testing. AANBEVELINGEN VOOR DE SELECTIE VAN STALEN VOOR DE SEQUENTIEBEPALING VAN HET VOLLEDIGE GENOOM IN HET KADER VAN SURVEILLANCE - UPDATE AUGUST 2021. Sciensano; 2021 Aug. Available: https://covid-19.sciensano.be/sites/default/files/Covid19/20210830_Advice_RAG_Update%20indications%20for%20sequencing_NL.pdf
71. Soto-Mota A, Marfil-Garza BA, Obeso SC, Rodriguez EJM, Vazquez DAC, Tadeo-Espinoza H, et al. Prospective predictive performance comparison between clinical gestalt and validated COVID-19 mortality scores. *Journal of Investigative Medicine*. 2022;70(2):415–420.
72. Shang Y, Pan C, Yang X, Zhong M, Shang X, Wu Z, et al. Management of critically ill patients with COVID-19 in ICU: statement from front-line intensive care experts in Wuhan, China. *Ann Intensive Care*. 2020;10:73.
73. Dhala A, Sasangohar F, Kash B, Ahmadi N, Masud F. Rapid Implementation and Innovative Applications of a Virtual Intensive Care Unit During the COVID-19 Pandemic: Case Study. *Journal of Medical Internet Research*. 2020;22(9):e20143.
74. Li T, Lu H, Zhang W. Clinical observation and management of COVID-19 patients. *Emerging Microbes & Infections*. 2020;9(1):687–690.

75. Griffin KM, Karas MG, Ivascu NS, Lief L. Hospital Preparedness for COVID-19: A Practical Guide from a Critical Care Perspective. *Am J Respir Crit Care Med*. 2020;201(11):1337–1344.
76. Li H, Zheng S, Liu F, Liu W, Zhao R. Fighting against COVID-19: Innovative strategies for clinical pharmacists. *Research in Social and Administrative Pharmacy*. 2021;17(1):1813–1818.
77. Aziz S, Arabi YM, Alhazzani W, Evans L, Citerio G, Fischkoff K, et al. Managing ICU surge during the COVID-19 crisis: rapid guidelines. *Intensive Care Med*. 2020;46(7):1303–1325.
78. Aleem A, Akbar Samad AB, Slenker AK. *Emerging Variants of SARS-CoV-2 And Novel Therapeutics Against Coronavirus (COVID-19)*. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024. Available: <http://www.ncbi.nlm.nih.gov/books/NBK570580/>
79. Yamasoba D, Kosugi Y, Kimura I, Fujita S, Uriu K, Ito J, et al. Neutralisation sensitivity of SARS-CoV-2 omicron subvariants to therapeutic monoclonal antibodies. Sensitivity of novel SARS-CoV-2 Omicron subvariants, BA.2.11, BA.2.12.1, BA.4 and BA.5 to therapeutic monoclonal antibodies. *The Lancet Infectious Diseases*. 2022;22(7):942–943.
80. Tada T, Zhou H, Dcosta BM, Samanovic MI, Mulligan MJ, Landau NR. Partial resistance of SARS-CoV-2 Delta variants to vaccine-elicited antibodies and convalescent sera. *iScience*. 2021;24(11):103341.
81. European Centre for Disease Prevention and Control (ECDC). COVID-19 Vaccine Tracker. 19 May 2022 [cited 11 Apr 2022]. Available: <https://vaccinetracker.ecdc.europa.eu/public/extensions/covid-19/vaccine-tracker.html#uptake-tab>
82. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *New England Journal of Medicine*. 2021;384(15):1412–1423.
83. Bernal JL, Andrews N, Gower C, Robertson C, Stowe J, Tessier E, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. *BMJ*. 2021;373: n1088.
84. Hall VJ, Foulkes S, Saei A, Andrews N, Oguti B, Charlett A, et al. COVID-19 vaccine coverage in health-care workers in England and effectiveness of BNT162b2 mRNA vaccine against infection (SIREN): a prospective, multicentre, cohort study. *The Lancet*. 2021;397(10286):1725–1735.
85. Bajema KL, Dahl RM, Prill MM, Meites E, Rodriguez-Barradas MC, Marconi VC, et al. Effectiveness of COVID-19 mRNA Vaccines Against COVID-19–Associated Hospitalization — Five Veterans Affairs Medical Centers, United States, February 1–August 6, 2021. *MMWR Morb Mortal Wkly Rep*. 2021;70(37):1294–1299.
86. Rosenberg ES, Dorabawila V, Easton D, Bauer UE, Kumar J, Hoen R, et al. Covid-19 Vaccine Effectiveness in New York State. *New England Journal of Medicine*. 2022;386:116–127.
87. Olson SM, Newhams MM, Halasa NB, Price AM, Boom JA, Sahni LC, et al. Effectiveness of BNT162b2 Vaccine against Critical Covid-19 in Adolescents. *New England Journal of Medicine*. 2022;386:713–723.
88. Hyams C, Marlow R, Maseko Z, King J, Ward L, Fox K, et al. Effectiveness of BNT162b2 and ChAdOx1 nCoV-19 COVID-19 vaccination at preventing hospitalisations in people aged at least 80 years: a test-negative, case-control study. *The Lancet Infectious Diseases*. 2021;21(11):1539–1548.
89. Andrews N, Stowe J, Kirsebom F, Toffa S, Sachdeva R, Gower C, et al. Effectiveness of COVID-19 booster vaccines against COVID-19-related symptoms, hospitalization and death in England. *Nat Med*. 2022;28:831–837.

90. Nunes B, Rodrigues AP, Kislaya I, Cruz C, Peralta-Santos A, Lima J, et al. mRNA vaccine effectiveness against COVID-19-related hospitalisations and deaths in older adults: a cohort study based on data linkage of national health registries in Portugal, February to August 2021. *Eurosurveillance*. 2021;26(38): 2100833.
91. Arbel R, Sergienko R, Friger M, Peretz A, Beckenstein T, Yaron S, et al. Effectiveness of a second BNT162b2 booster vaccine against hospitalization and death from COVID-19 in adults aged over 60 years. *Nat Med*. 2022;28:1486-1490.
92. Luring AS, Tenforde MW, Chappell JD, Gaglani M, Ginde AA, McNeal T, et al. Clinical severity of, and effectiveness of mRNA vaccines against, covid-19 from omicron, delta, and alpha SARS-CoV-2 variants in the United States: prospective observational study. *BMJ*. 2022;376:e069761.
93. Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New England Journal of Medicine*. 2021;385(7):585–594.
94. Hayawi K, Shahriar S, Serhani MA, Alashwal H, Masud MM. Vaccine versus Variants (3Vs): Are the COVID-19 Vaccines Effective against the Variants? A Systematic Review. *Vaccines*. 2021;9(11):1305.
95. Au WY, Cheung PP-H. Effectiveness of heterologous and homologous covid-19 vaccine regimens: living systematic review with network meta-analysis. *BMJ*. 2022;377:e069989.
96. Self WH. Comparative Effectiveness of Moderna, Pfizer-BioNTech, and Janssen (Johnson & Johnson) Vaccines in Preventing COVID-19 Hospitalizations Among Adults Without Immunocompromising Conditions — United States, March–August 2021. *MMWR Morb Mortal Wkly Rep*. 2021;70(38):1337-1343.
97. Nordström P, Ballin M, Nordström A. Risk of SARS-CoV-2 reinfection and COVID-19 hospitalisation in individuals with natural and hybrid immunity: a retrospective, total population cohort study in Sweden. *The Lancet Infectious Diseases*. 2022;22(6):781-790.
98. Altarawneh HN, Chemaitelly H, Hasan MR, Ayoub HH, Qassim S, AlMukdad S, et al. Protection against the Omicron Variant from Previous SARS-CoV-2 Infection. *New England Journal of Medicine*. 2022;386(13):1288–1290.
99. Braeye T, van Loenhout J, Brondeel R, Stouten V, Hubin P, Billuart M, et al. COVID-19 Vaccine effectiveness against symptomatic infection and hospitalization in Belgium, July 2021-APRIL 2022. medRxiv. 2022.
100. Menni C, May A, Polidori L, Louca P, Wolf J, Capdevila J, et al. COVID-19 vaccine waning and effectiveness and side-effects of boosters: a prospective community study from the ZOE COVID Study. *Lancet Infect Dis*. 2022;22(7):1002-1010.
101. Chemaitelly H, Tang P, Hasan MR, AlMukdad S, Yassine HM, Benslimane FM, et al. Waning of BNT162b2 Vaccine Protection against SARS-CoV-2 Infection in Qatar. *New England Journal of Medicine*. 2021;385(24):e83.
102. Goldberg Y, Mandel M, Bar-On YM, Bodenheimer O, Freedman LS, Ash N, et al. Protection and Waning of Natural and Hybrid Immunity to SARS-CoV-2. *New England Journal of Medicine*. 2022;386(23):2201-2212.
103. Ward T, Johnsen A. Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLOS ONE*. 2021;16(10):e0257978.

104. Li L, Sun W, Han M, Ying Y, Wang Q. A Study on the Predictors of Disease Severity of COVID-19. *Med Sci Monit*. 2020;26:e927167-1-e927167-8.
105. Dananché C, Elias C, Hénaff L, Amour S, Kuczewski E, Gustin M-P, et al. Baseline clinical features of COVID-19 patients, delay of hospital admission and clinical outcome: A complex relationship. *PLOS ONE*. 2022;17(1):e0261428.
106. Alaa A, Qian Z, Rashbass J, Bengier J, van der Schaar M. Retrospective cohort study of admission timing and mortality following COVID-19 infection in England. *BMJ Open*. 2020;10(11):e042712.
107. Rossman H, Meir T, Somer J, Shilo S, Gutman R, Ben Arie A, et al. Hospital load and increased COVID-19 related mortality in Israel. *Nat Commun*. 2021;12:1904.
108. Bravata DM, Perkins AJ, Myers LJ, Arling G, Zhang Y, Zillich AJ, et al. Association of Intensive Care Unit Patient Load and Demand With Mortality Rates in US Department of Veterans Affairs Hospitals During the COVID-19 Pandemic. *JAMA Network Open*. 2021;4(1):e2034266.
109. Strålin K, Wahlström E, Walther S, Bennet-Bark AM, Heurgren M, Lindén T, et al. Mortality in hospitalized COVID-19 patients was associated with the COVID-19 admission rate during the first year of the pandemic in Sweden. *Infectious Diseases*. 2022;54(2):145–151.
110. Lytras T, Tsiodras S. Total patient load, regional disparities and in-hospital mortality of intubated COVID-19 patients in Greece, from September 2020 to May 2021. *Scand J Public Health*. 2022;50(6):671-675.
111. Patone M, Thomas K, Hatch R, Tan PS, Coupland C, Liao W, et al. Mortality and critical care unit admission associated with the SARS-CoV-2 lineage B.1.1.7 in England: an observational cohort study. *The Lancet Infectious Diseases*. 2021;21(11):1518–1528.
112. Goethem NV, Vandromme M, Oyen HV, Haarhuis F, Brondeel R, Catteau L, et al. Severity of infection with the SARS-CoV-2 B.1.1.7 lineage among hospitalized COVID-19 patients in Belgium. *PLOS ONE*. 2022;17(6):e0269138.
113. Edward PR, Lorenzo-Redondo R, Reyna ME, Simons LM, Hultquist JF, Patel AB, et al. Severity of Illness Caused by Severe Acute Respiratory Syndrome Coronavirus 2 Variants of Concern in Children: A Single-Center Retrospective Cohort Study. *Journal of the Pediatric Infectious Diseases Society*. 2022;11(10):440-447.
114. Fillâtre P, Dufour M-J, Behillil S, Vatan R, Reusse P, Gabellec A, et al. A new SARS-CoV-2 variant with high lethality poorly detected by RT-PCR on nasopharyngeal samples: an observational study. *Clin Microbiol Infect*. 2022;28(2):298.e9-298.e15.
115. European Centre for Disease Prevention and Control. Guidance for representative and targeted genomic SARS-CoV-2 monitoring – 3 May 2021. ECDC: Stockholm. 2021 [cited 17 Dec 2024]. Available: <https://www.ecdc.europa.eu/en/publications-data/guidance-representative-and-targeted-genomic-sars-cov-2-monitoring>.

7. APPENDICES

Appendix A – PRISMA flow diagram

Appendix B – Data extraction file

Appendix C – Coverage of Belgian hospitalized COVID-19 patients with confirmed SARS-CoV-2 variant information

Appendix D – Vaccination status of Belgian hospitalized COVID-19 patients

Appendix E – Documented previous SARS-CoV-2 infections of Belgian hospitalized COVID-19 patients

Appendix F – Nursing home residents within the population of Belgian hospitalized COVID-19 patients

Appendix G – Age and mean ICU occupancy of Belgian hospitalized COVID-19 patients

Appendix A – PRISMA flow diagram

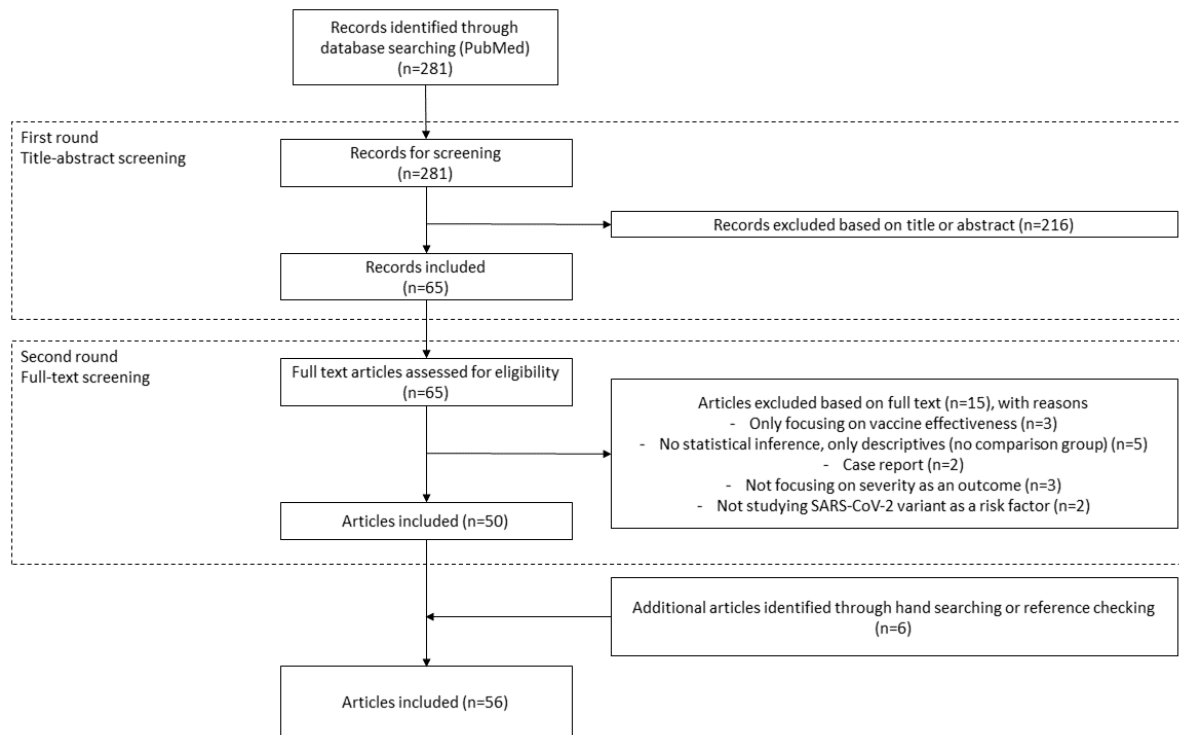


Figure A1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram, presenting the selection process of articles included in the scoping literature review, conducted to summarize methodological approaches to study the severity of SARS-CoV-2 variants using observational data and identify limitations and potential biases resulting from the study design, data analysis approach, underlying surveillance strategies, or data infrastructure.

Appendix B – Data extraction file

An Excel file was generated, containing extracted data from 56 articles included in the scoping review.

You can download the Excel file from the following publication:

Meurisse M, Van Oyen H, Blot K, Catteau L, Serrien B, Klamer S, et al. Evaluating methodological approaches to assess the severity of infection with SARS-CoV-2 variants: scoping review and applications on Belgian COVID-19 data. *BMC Infect Dis.* 2022;22:839. <https://doi.org/10.1186/s12879-022-07777-6>

Appendix C – Coverage of Belgian hospitalized COVID-19 patients with confirmed SARS-CoV-2 variant information

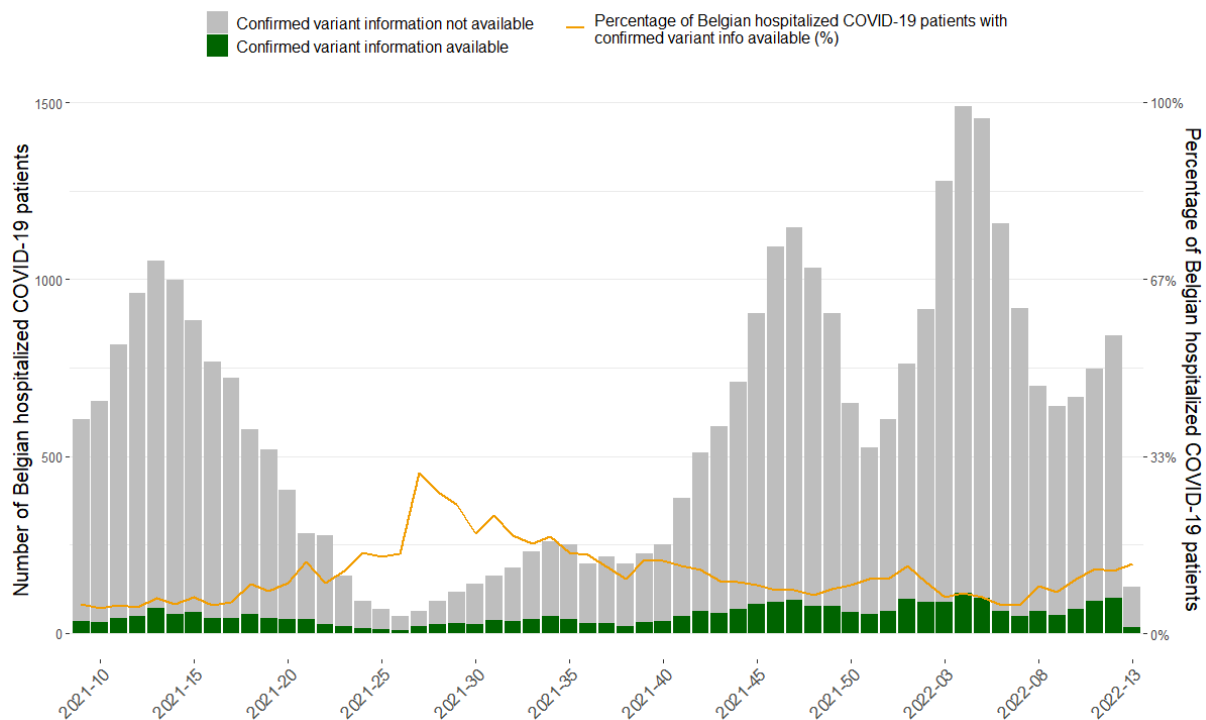


Figure A2. The number of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Surveillance (CHS) and admitted between the 1st of March 2021 and the 28th of March 2022 by the availability of SARS-CoV-2 variant information (variant information confirmed by Whole Genome Sequencing (WGS) available versus no confirmed variant information available) (left y-axis) and the coverage of Belgian hospitalized COVID-19 patients with confirmed variant information, i.e. percentage of patients with available confirmed (by WGS) viral information, among the of Belgian hospitalized COVID-19 patients admitted between the 1st of March 2021 and the 28th of March 2022 (right y-axis).

Appendix D – Vaccination status of Belgian hospitalized COVID-19 patients

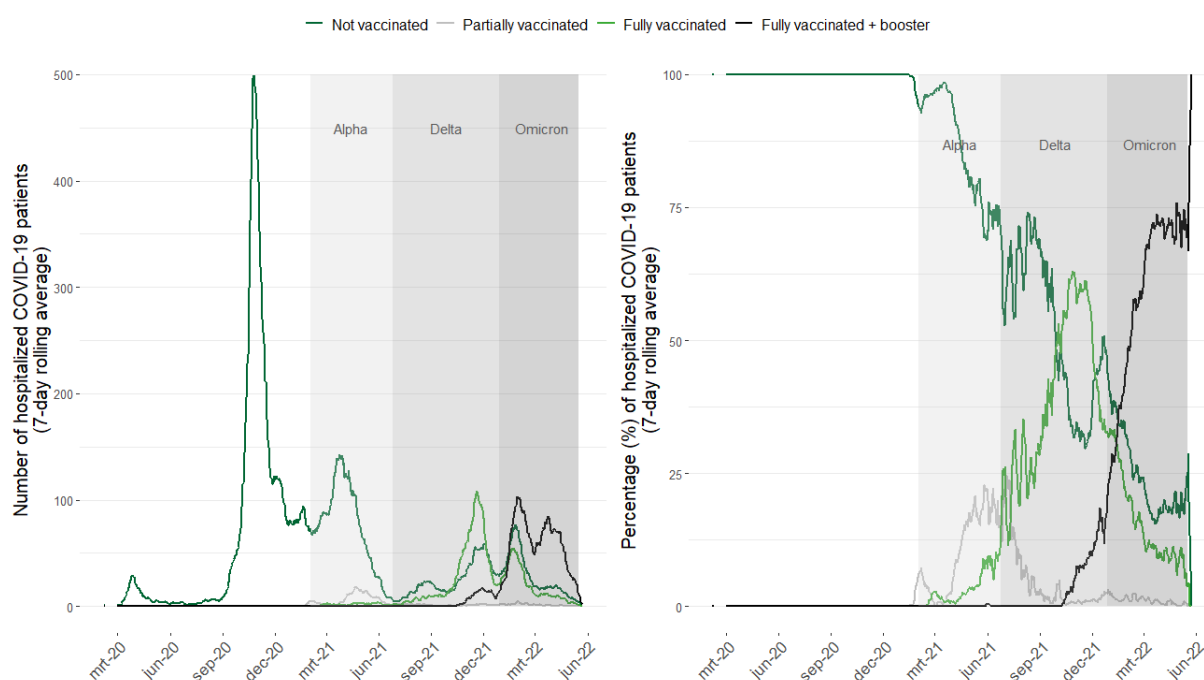


Figure A3. Vaccination status of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Survey (CHS). (Left) Number of Belgian hospitalized COVID-19 patients with a certain vaccination status over time, 7-day rolling average. (Right) Percentage of Belgian hospitalized COVID-19 patients with a certain vaccination status over time, 7-day rolling average. Periods of dominance of SARS-CoV-2 variants (more than 50% presence in baseline surveillance) are indicated as areas on the plot. Not vaccinated = no dose of a vaccine. Partially vaccinated = one dose of the BNT162b2, mRNA-1273 or NVX-CoV2373 vaccine. Fully vaccinated = one dose of the Ad26.COV2.S vaccine, two doses of the BNT162b2, mRNA-1273 or NVX-CoV2373 vaccine, or a mixture of two doses of the latter three vaccines (= primary vaccination schedule). Fully vaccinated + booster = primary vaccination schedule plus an additional dose of the BNT162b2 or mRNA-1273 vaccine.

Appendix E – Documented previous SARS-CoV-2 infections of Belgian hospitalized COVID-19 patients

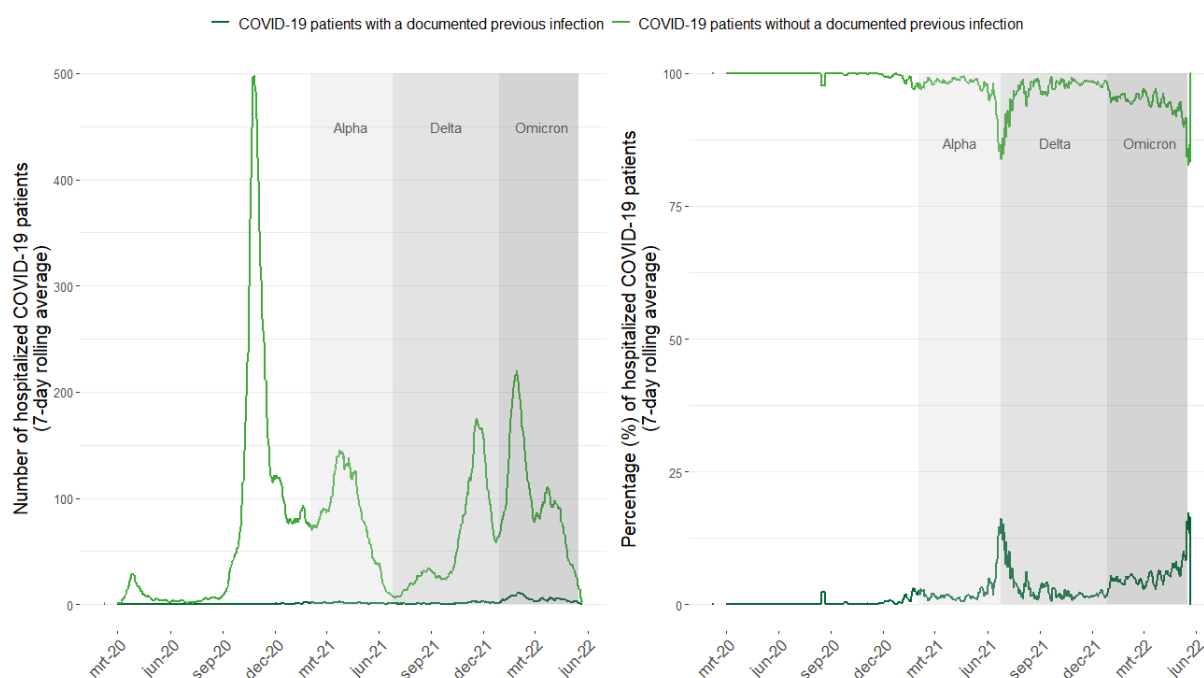


Figure A4. Documented previous SARS-CoV-2 infections of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Survey (CHS). (Left) Number of Belgian hospitalized COVID-19 patients with and without a documented previous infection, 7-day rolling average. (Right) Percentage of Belgian hospitalized COVID-19 patients with and without a documented previous infection, 7-day rolling average. Periods of dominance of SARS-CoV-2 variants (more than 50% presence in baseline surveillance) are indicated as areas on the plot.

Appendix F – Nursing home residents within the population of Belgian hospitalized COVID-19 patients

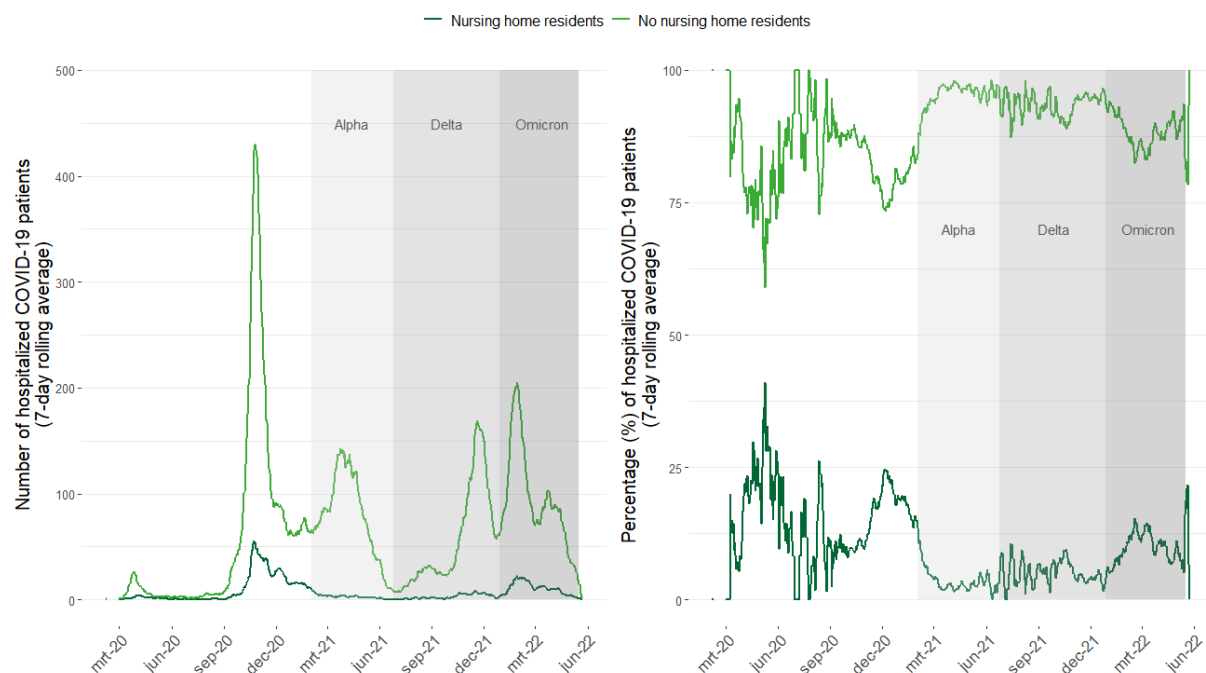


Figure A5: Nursing home residents within the population of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Survey (CHS). (Left) Number of Belgian hospitalized COVID-19 patients that are nursing home residents, 7-day rolling average. (Right) Percentage of Belgian hospitalized COVID-19 nursing home residents, 7-day rolling average. Periods of dominance of SARS-CoV-2 variants (more than 50% presence in baseline surveillance) are indicated as areas on the plot.

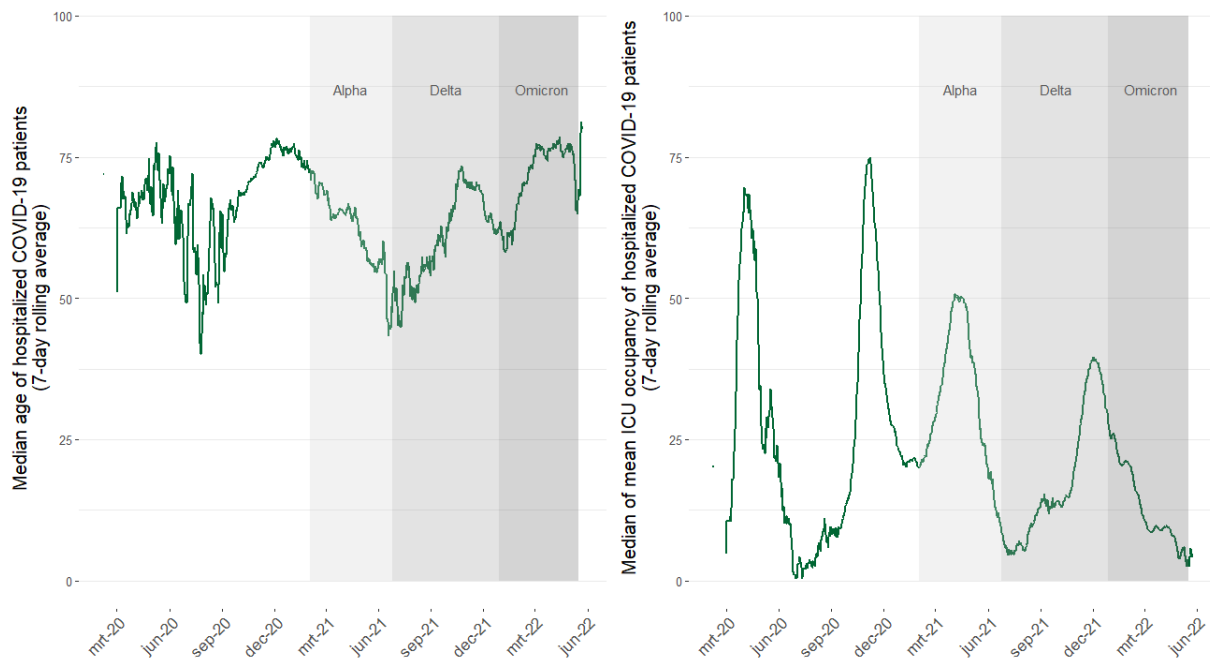
Appendix G – Age and mean ICU occupancy of Belgian hospitalized COVID-19 patients

Figure A6. (Left) Median age of Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Survey (CHS), 7-day rolling average. (Right) Median of the mean ICU occupancy of Belgian hospitalized COVID-19 patients registered in the CHS, 7-day rolling average. Periods of dominance of SARS-CoV-2 variants (more than 50% presence in baseline surveillance) are indicated as areas on the plot.

CHAPTER 2

Impact of homologous and heterologous prime-boost vaccination on the clinical severity of SARS-CoV-2 Omicron infections

Meurisse M, Catteau L, van Loenhout J A F, Braeye T, De Mot L, Serrien B, Blot K, Cauët E, Van Oyen H, Cuypers L, Belgian Collaborative Group on COVID-19 Hospital Surveillance, COVID-19 Genomics Belgium Consortium, Robert A*, Van Goethem N*. Homologous and heterologous prime-boost vaccination: impact on clinical severity of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients in Belgium. *Vaccines*. 2023; 11: 378. *Shared last author

ABSTRACT

We investigated the effectiveness of (1) mRNA booster vaccination versus primary vaccination only and (2) heterologous (viral vector–mRNA) versus homologous (mRNA–mRNA) prime-boost vaccination against severe outcomes of BA.1, BA.2, BA.4 or BA.5 Omicron infection (confirmed by whole genome sequencing) among hospitalized COVID-19 patients using observational data from national COVID-19 registries. In addition, it was investigated whether the difference between the heterologous and homologous prime-boost vaccination was homogenous across Omicron sub-lineages. Regression standardization (parametric g-formula) was used to estimate counterfactual risks for severe COVID-19 (combination of severity indicators), intensive care unit (ICU) admission, and in-hospital mortality under exposure to different vaccination schedules. The estimated risk for severe COVID-19 and in-hospital mortality was significantly lower with an mRNA booster vaccination as compared to only a primary vaccination schedule (RR = 0.59 [0.33; 0.85] and RR = 0.47 [0.15; 0.79], respectively). No significant difference was observed in the estimated risk for severe COVID-19, ICU admission and in-hospital mortality with a heterologous compared to a homologous prime-boost vaccination schedule, and this difference was not significantly modified by the Omicron sub-lineage. Our results support evidence that mRNA booster vaccination reduced the risk of severe COVID-19 disease during the Omicron-predominant period.

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to a public health crisis across the world. The clinical severity of COVID-19 ranges from asymptomatic, mild or moderate (e.g., experiencing fever, headache, diarrhea, and myalgia) to severe disease (e.g., experiencing dyspnea, hypoxemia, acute respiratory distress syndrome (ARDS)) [1]. SARS-CoV-2 vaccination has shown to provide effective protection against severe SARS-CoV-2 infections requiring admission to the hospital [2–6] or intensive care unit (ICU) [6,7], and COVID-19-related death [2,3,5]. In Belgium, a large-scale vaccination campaign was initiated on 5 January 2021 [8,9] followed by the roll-out of a booster vaccination campaign from September 2021 onwards, resulting in 79% of the total population with a primary vaccination schedule, 62% with a booster dose, and 33% with two booster doses by the beginning of December 2022 [10]. The BNT162b2 vaccine (Comirnaty®, Pfizer–BioNTech), mRNA-1273 vaccine (Spikevax®, Moderna), Ad26.COVS.2 vaccine (COVID-19 vaccine Janssen®, Johnson & Johnson) and ChAdOx1-S vaccine (Vaxzevria®, Oxford–AstraZeneca) were initially approved by the European Medicines Agency (EMA), and have widely been administered for primary vaccination throughout the Belgian population [10]. Different types of vaccines exist, each training the human immune system differently [11]. Pfizer and Moderna vaccines are mRNA vaccines (i.e., artificially created viral mRNA), while the latter two are viral vector vaccines (i.e., genetic material of the virus placed in a viral vector) [12–15]. mRNA vaccines BNT162b2 and mRNA-1273 were primarily used for booster vaccination in Belgium [10], potentially matching the vaccine type (mRNA or viral vector) of the primary vaccination (i.e., a homologous booster) or not (i.e., a heterologous booster). SARS-CoV-2 vaccines act through humoral immune responses (e.g., B-cell proliferation into neutralizing antibody-producing plasma cells, neutralizing antibodies target the viral spike protein and block the virus' ability to enter human cells), cellular immune responses (e.g., T cell proliferation into cytotoxic T cells) and memory responses of B- and T-cells [11,16]. It has been previously indicated for non-SARS-CoV-2 vaccines that heterologous prime-boost vaccination might result in more durable and broader vaccine-induced immune response and a higher number of neutralizing antibodies [17,18]. The same has been suggested for SARS-CoV-2 vaccines [19–21]; however, this might differ according to circulating SARS-CoV-2 variants, and the vaccine types and order of vaccine types used for primary and booster vaccination [17,20]. To our knowledge, the literature comparing the effectiveness of SARS-CoV-2 'viral vector–mRNA' and 'mRNA–mRNA' prime-boost vaccination during a period of Omicron circulation is still limited.

At the same time, emerging SARS-CoV-2 variants accumulate mutations over time, potentially resulting in viral characteristics that allow them to more efficiently evade vaccine-induced immunity. A

mismatch between the variant and the vaccine may subsequently reduce the efficacy of the vaccine [22,23]. For example, due to its large number of mutations in the viral spike gene (i.e., the target of neutralizing antibodies), the Omicron variant (lineage B.1.1.529) was found to be more efficient in evading neutralizing antibodies than previously circulating variants [24–27] and initial evidence shows a lower effectiveness of booster vaccination against severe clinical outcomes of a Omicron infection compared to previous variants. Hence, after its emergence in late 2021, it has caused significant concerns about the enduring effectiveness of vaccines. Moreover, between sub-lineages of the Omicron variant, characterized by important differences in the S protein of the virus, differences in resistance to neutralization and vaccine effectiveness can be observed [28–30]. In Belgium, the initially circulating Omicron BA.1 lineage was first detected in November 2021 [31] and the BA.1 and BA.1.1 lineages jointly represented more than 80% of the positive samples by the first week of January 2022 [32]. Subsequently, BA.2 emerged and gradually replaced BA.1 during early 2022, reaching 80% of circulating strains by the beginning of March 2022 [33]. Next, BA.4 and BA.5 lineages emerged and BA.5 became more prevalent from May 2022 onwards [10,34]. By the beginning of December 2022, BA.5 and its sub-lineages (e.g., BQ.1*) were the predominantly circulating strains [35]. In addition to a potential change in the capacity to evade neutralizing antibodies, the regularly occurring mutations may also alter the viral pathogenicity [36,37]. Therefore, continuously evaluating the impact of newly emerging variants is essential.

The current study aimed to compare the effectiveness of (1) mRNA booster vaccination versus primary vaccination only and (2) a heterologous (viral vector–mRNA) versus homologous (mRNA–mRNA) prime-boost vaccination scheme in preventing severe clinical outcomes of SARS-CoV-2 Omicron infection among hospitalized patients in Belgium, and whether the effect of heterologous versus homologous prime-boost vaccination scheme was homogenous across different Omicron sub-lineages (i.e., is the difference in risk for severe clinical outcomes between heterologous and homologous prime-boost vaccination modified by the Omicron sub-lineage of infection). This study focused on hospitalized COVID-19 patients in Belgium and aims to assess this effect using observational data. By conducting this study, we aim to inform vaccine policy and contribute to knowledge used for the development of strategies for booster campaigns during the continuously evolving COVID-19 epidemic.

2. METHODS

2.1 Target Trial Specification

This observational study was developed to emulate two conceptual target randomized control trials comparing the effectiveness of (1) a primary vaccination (viral vector or mRNA) plus mRNA booster vaccination (i.e., homologous or heterologous prime-boost vaccination scheme) versus a primary vaccination (viral vector or mRNA) without booster vaccination and (2) a heterologous (viral vector prime–mRNA boost) versus homologous (mRNA prime–mRNA boost) prime-boost vaccination scheme in preventing severe clinical outcomes of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients in Belgium. Key elements of the protocol of the two target trials have been specified in Table A1 (Appendix A).

In both hypothetical target trials, adults (age ≥ 18 , with age defined at the start of COVID-19 related hospital care) are randomly assigned to one of the intervention arms. In the first target trial, the intervention arms consist of (a) an mRNA (two doses of BNT162b2 or mRNA-1273) or viral vector (one dose of Ad26.COVS.2 or two doses of ChAdOx1-S) primary vaccination plus an mRNA (BNT162b2 or mRNA-1273) booster vaccination and (b) an mRNA or viral vector primary vaccination without booster vaccination (see Figure 1). In the second target trial the intervention arms consist of (a) heterologous prime-boost vaccination, defined as a viral vector (one dose of Ad26.COVS.2 or two doses of ChAdOx1-S) primary vaccination plus mRNA (BNT162b2 or mRNA-1273) booster vaccination, and (b) homologous prime-boost vaccination, defined as an mRNA (two doses of BNT162b2 or mRNA-1273) primary vaccination plus mRNA (BNT162b2 or mRNA-1273) booster vaccination (see Figure 1). Individuals assigned to one of the intervention arms are only eligible for inclusion in the study if they subsequently acquire a SARS-CoV-2 infection (either community- or hospital-acquired) that is confirmed by a reverse transcription polymerase chain reaction (RT-PCR) or antigen (AG) laboratory test and, in addition, if the COVID-19 disease progression of this infection requires COVID-19 related hospital care. Individuals are not eligible when the detected SARS-CoV-2 infection occurs less than 14 days after the last vaccination dose of their vaccination schedule and when the lineage assigned based on a whole genome sequencing (WGS) test result indicates an infection with a variant other than Omicron. Patient follow up for eligible patients starts at the beginning of COVID-19 related hospital care (time zero, T_0) and ends at hospital discharge. Baseline patient characteristics are collected at the start of the COVID-19 related hospital care, and clinical severity (severe COVID-19, ICU admission, and in-hospital mortality) is assessed at hospital discharge.

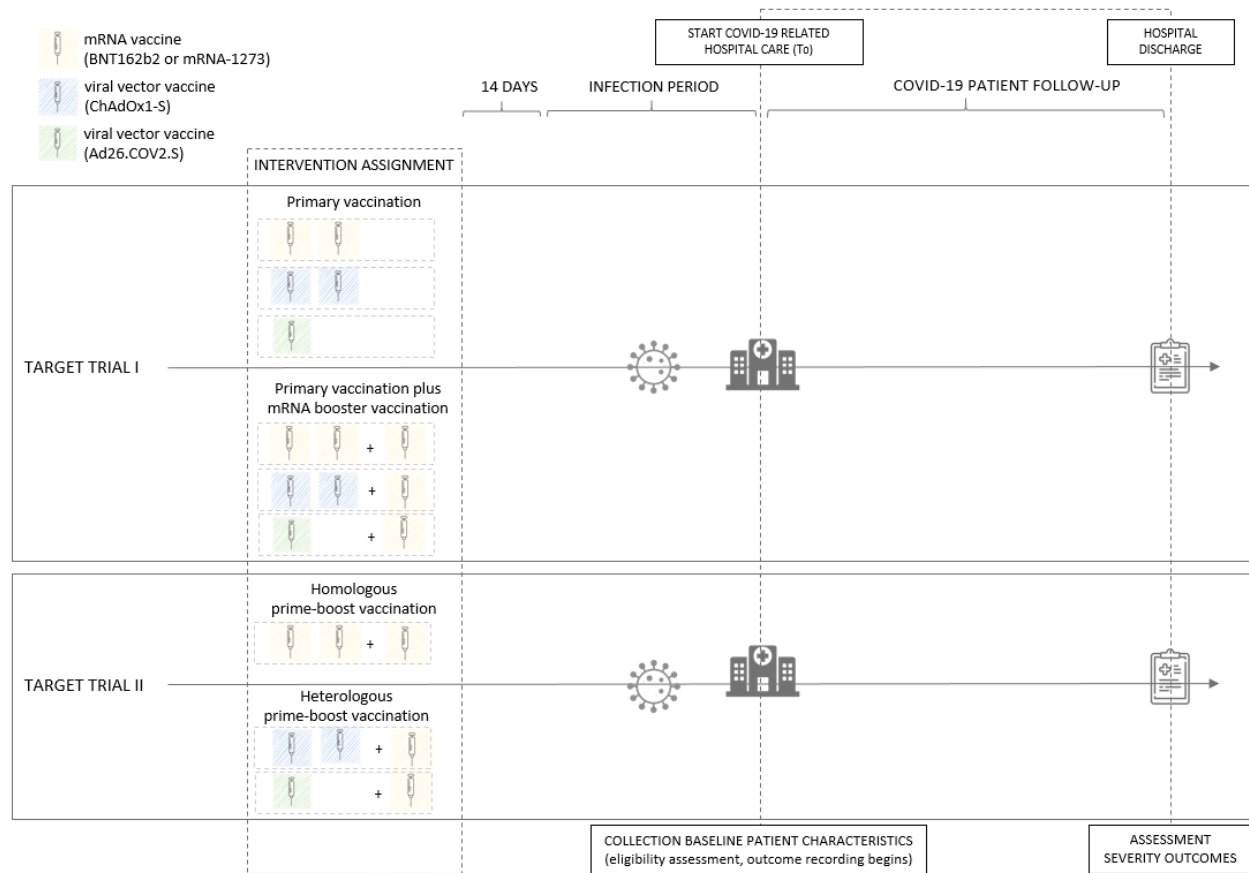


Figure 1. Graphical representation of the conceptual target randomized control trials comparing the effectiveness of (1) primary vaccination plus mRNA booster vaccination versus primary vaccination without booster vaccination (target trial I) and (2) heterologous versus homologous prime-boost vaccination (target trial II), in preventing severe clinical outcomes of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients in Belgium. Individuals assigned to one of the intervention arms, with a SARS-CoV-2 infection (community- or hospital-acquired, confirmed by an RT-PCR or AG test) acquired at least 14 days after the last administered vaccination dose, requiring COVID-19 related hospital care and fitting with the eligibility criteria, are followed up from the beginning of COVID-19 related hospital care until hospital discharge. Baseline patient characteristics are collected at the start of the COVID-19 related hospital care (T₀) and clinical severity (severe COVID-19, ICU admission, and in-hospital mortality) is assessed at hospital discharge.

2.2 Target Trial Emulation

Observational data from existing national surveillance systems and administrative sources were leveraged to emulate the specified hypothetical target trials.

Data Sources

Secure individual-level linkage of selected variables from different existing data sources was conducted within the framework of the LINK-VACC project, initiated by Sciensano, the Belgian national institute for health. Linkage was executed using a pseudonymized national registry number within a secured environment (HealthData.be platform). Data on Belgian hospitalized COVID-19 patients registered in the Clinical Hospital Survey (CHS, including information on comorbidities and severity indicators) [38] were linked to data from the COVID-19 TestResult database [39–41] (to obtain information on reinfections and the variant of infection), data from the national vaccine registry (Vaccinnet+) [42], data on socio-economic factors obtained from the Belgian Statistical Office (StatBel), data from the surge capacity survey (to obtain information on the hospital bed occupancy) [38] and data from the Common Base Registry for Healthcare Actor (CoBRHA, to obtain information on the recognition and activity of people as healthcare actors) [43]. The data infrastructure has been described elsewhere [44].

Eligibility Criteria

The study population consisted of symptomatic laboratory-confirmed COVID-19 patients admitted to a Belgian hospital between 1 September 2021 (i.e., the start of the Belgian booster vaccination campaign) and 23 November 2022 (i.e., the date of data extraction minus two weeks to allow for sufficient follow-up time), with an admission form reported in the CHS. To comply with the eligibility criteria of the target trial, hospitalized patients who were tested for COVID-19 in a screening context were excluded, as well as patients diagnosed with COVID-19 more than 20 days before hospital admission or after hospital discharge, with the aim of only considering patients who received hospital care for their COVID-19 related pathology and excluding patients who had reported severe clinical outcomes unrelated to COVID-19. Furthermore, underaged patients (<18 years) were excluded, as well as transferred or readmitted patients, and patients admitted to a psychiatric hospital or a hospital without an ICU. Patients with a missing pseudonymized national registry number were additionally excluded, as this information was required to link to other registries. Furthermore, patients with an unknown date of diagnosis, starting date of COVID-19-related hospital care or date of discharge were not included. As we assessed the risk of severe clinical outcomes of a SARS-CoV-2 Omicron infection under exposure to different vaccination schedules, only patients infected with Omicron were

considered for the analysis. Four sub-lineages of Omicron could be identified in the eligible patients by linking WGS-confirmed lineage data from COVID-19 TestResult database: BA.1, BA.2, BA.4 and BA.5. Due to the low prevalence of BA.4 infections among the eligible patients with an WGS-confirmed Omicron infection (13 out of 1,119 patients) and co-circulation of BA.4 and BA.5 in time, BA.4 and BA.5 were combined and further referred to as BA.4/5. For the main analysis, infection with an Omicron variant was ascertained through the linkage with the WGS-confirmed lineage data from the COVID-19 TestResult database. Patients without a linked WGS-confirmed lineage or with a linked WGS-confirmed lineage different from Omicron BA.1, BA.2, BA.4 or BA.5 were excluded from the study population in the main analysis. As linkage with the non-exhaustive WGS test results limited the sample size significantly (only for 11% of the eligible patients a WGS-confirmed lineage could be linked) and might have resulted in selection bias [45], a sensitivity analysis was performed where patients were categorized based on their COVID-19 diagnosis date within time periods with known circulation of the Omicron sub-lineages [35]. Patients diagnosed with COVID-19 during a restricted time period with an approximated circulation of BA.1, BA.2 or BA.4/5 of more than 80%, were assigned to the BA.1, BA.2 or BA.4/5 Omicron sub-lineage, respectively. These restricted time periods were set by considering the distribution of variants in the representative Belgian baseline genomic surveillance [35]. The time periods to assign the BA.1, BA.2 or BA.4/5 Omicron sub-lineage were set to 3 January 2022–13 February 2022, 14 March 2022–22 May 2022 and later than 22 June 2022, respectively. Patients with a COVID-19 diagnosis date outside these predefined periods were excluded from the study population in the sensitivity analysis.

Intervention Strategies and Assignment

Intervention groups could be determined in the population eligible for inclusion through individual-level linkage with data from the national vaccine registry (Vaccinnet+). Vaccination status of eligible patients was assigned on the date of the COVID-19 diagnosis resulting in the hospitalization.

To emulate the target trials, patients who had completed one of the vaccination schedules as defined in the intervention arms and who had received the last dose of their vaccination schedule at least 14 days before the date of diagnosis of the confirmed COVID-19 infection were assigned to the respective intervention group. Patients with a vaccination schedule not corresponding to the definition of one of the intervention arms, were excluded from the study population (e.g., not or partially vaccinated patients, patients who received a heterologous primary vaccination scheme, patients who had received more than one booster dose, patients who received a non-mRNA booster vaccination).

Randomization was mimicked by conditioning on a set of factors that were assumed to confound the exposure-outcome pathway. We assumed that individuals were exchangeable within levels of the

confounders. A directed acyclic graph (DAG) was constructed based on a priori assumptions about the causal relationship between COVID-19 vaccination scheme (i.e., the exposure) and COVID-19 disease severity (see Figure 2). The DAG includes both observed and unobserved variables. The code to generate the DAG with the 'DAGitty' software [46] is provided in the Appendix B. A sufficient adjustment set to close the non-causal backdoor paths captured in the DAG was identified: patients' comorbidities, age, gender, pregnancy status, host genetics, socio-economic status (SES), previous infection(s), whether the patient is a nursing home resident, whether the patient is a healthcare worker, the hospital of admission and ICU load during COVID-19-related hospital stay, whether the patient had a hospital-acquired infection, treatment strategies used during the time of admission, and Omicron sub-lineage of infection.

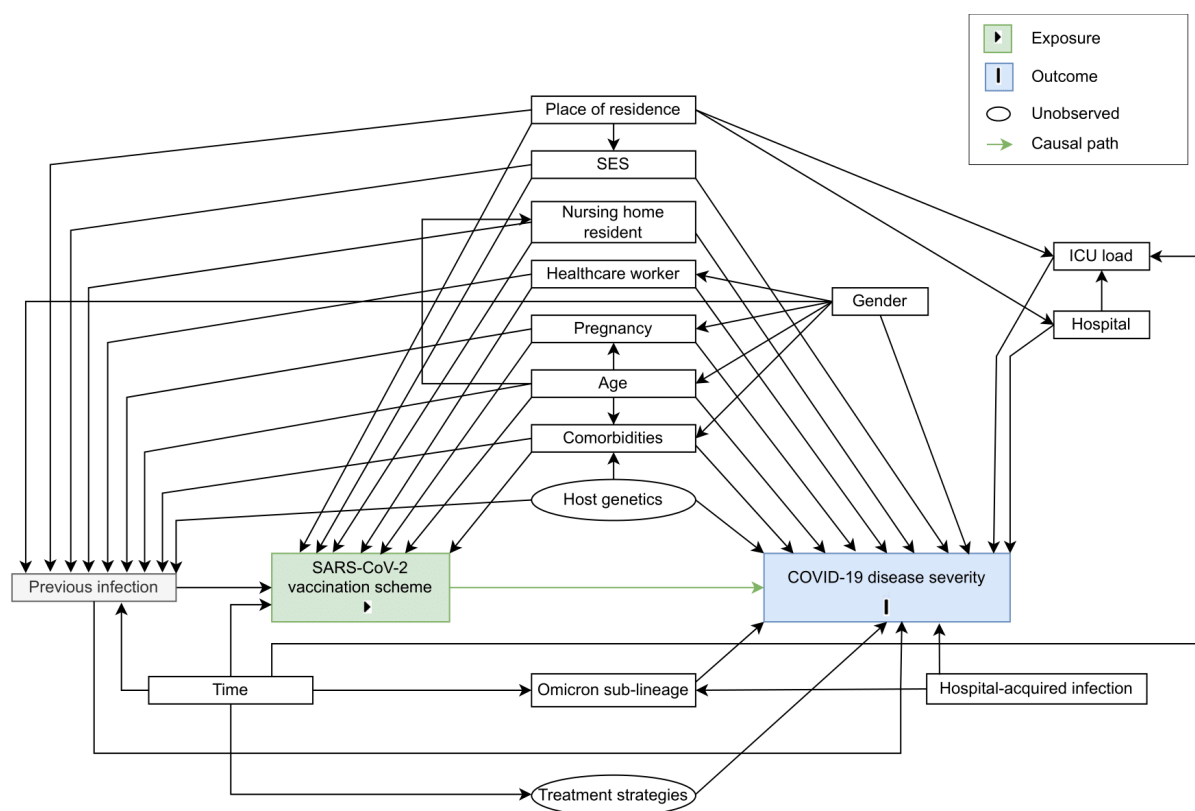


Figure 2. Directed Acyclic Graph (DAG) representing the causal assumptions that were made when assessing the effect of SARS-CoV-2 vaccination scheme (i.e., the exposure, indicated in green) on COVID-19 disease severity (i.e., the outcome, indicated in blue). Causal paths are shown as green arrows.

The sufficient adjustment set, determining which covariates we should ideally consider to close the non-causal backdoor paths as captured in the DAG, was approximated by the information available in the different registries within the LINK-VACC project.

Individual comorbidities, registered in the CHS, were grouped into three categories based on their strength as a risk factor for severe COVID-19: no underlying comorbidities, medium-risk underlying comorbidities (i.e., cardiovascular disease, chronic liver disease, obesity, immunosuppressive therapy,

cognitive disorder, diabetes, or chronic lung disease) and high-risk comorbidities (i.e., solid cancer, hematological cancer, chronic kidney disease, immunodeficiency, chronic neurological disease, or people with an organ transplant). SES was approximated by education level and income categories at the individual level, and by the median net taxable income per capita and population density of the postal code of residence. Education level categories were obtained by categorizing the assigned International Standard Classification of Education (ISCED) levels into ‘low’ (ISCED0–ISCED2), ‘medium’ (ISCED3–ISCED4) and ‘high’ (ISCED5–ISCED8). Income categories were determined by breaking up the net income of the patient’s household into deciles and categorizing these deciles into ‘low’ (decile 1–4), ‘medium’ (decile 5–7) and ‘high’ (decile 8–10). As not only vaccination- but also infection-induced immunity contributes to a patient’s preexisting immunity, previous infection was considered as an important confounder in the analysis. We defined that a patient had a documented previous infection when the sample, for diagnosis of the SARS-CoV-2 infection for which the patient is hospitalized, was collected more than 60 days after the date of sample collection from a previous documented positive COVID-19 test (RT-PCR or AG) [47]. Whether or not a patient was a nursing home resident and whether the infection was nosocomial (i.e., acquired in the hospital) was captured in the CHS. Information on a patient’s activity as a healthcare worker was obtained by combining information from the CoBRHA registry, which contains records on healthcare professionals by degree, and the patient’s admission form in the CHS. The hospital of admission was categorized into three hospital types (general hospitals, university hospitals or general hospitals with a university character). We assumed that treatment strategies were constant over the study period. Information on host genetics was missing.

Follow-up Period and Outcomes of Interest

As in the target trial specification, the outcomes of the causal inference study were (1) the development of severe COVID-19, as defined by the occurrence of either ICU admission, ARDS, or in-hospital mortality (all causes), (2) ICU admission and (3) in-hospital mortality separately. Patients were followed up from hospital admission until hospital discharge, with details on their COVID-19 disease outcome reported in the discharge form.

2.3 Statistical Analyses

Characteristics of patients in the selected study population were presented by their vaccination scheme: primary vaccinated patients (viral vector or mRNA) without booster vaccination, patients who have received a homologous (mRNA–mRNA) prime-boost vaccination scheme, and patients who have received a heterologous (viral vector–mRNA) prime-boost vaccination scheme. Continuous variables

were presented by their median value and interquartile range (IQR), while categorical variables were described as a counts (n) and proportions (%).

Missing values in variables included in the multivariable model (see further) were imputed with the R package ‘MICE’ (multivariate imputation via chained equations) [48]. We assumed that the missing values were missing at random (MAR), i.e., missingness can be explained by other covariates and is not related to the missing value itself. Twenty-fold multiple imputation was executed, with predictive mean matching, logistic regression and polytomous logistic regression for numerical, binary and categorical variables, respectively. The variables for severe COVID-19 and nosocomial infection were passively imputed. Twenty iterations were performed to obtain convergence of the Markov chain Monte Carlo (MCMC) algorithm.

A logistic regression model was built, adjusted for the following covariates: comorbidity group, age, gender, nursing home residence, median net taxable income per capita and population density of the postal code of residence, education level category, income category, previous infection, type of hospital and mean ICU occupancy rate in the hospital of admission during the stay, whether it was a hospital-acquired infection, and a two-way interaction between the Omicron sub-lineage of infection and the vaccination schedule. Covariates for pregnancy status and whether the patient is a healthcare worker were not included in the model because of their low prevalence in the study population, resulting in separation in the regression models. Numeric variables were added both as linear and quadratic term in the model.

Standardization [49] using the fitted logistic regression model (parametric G-formula) was performed in each imputed dataset, where we predict counterfactual risks of potential outcomes under intervention (i.e., under primary vaccination without booster vaccination (R_0) and primary vaccination plus mRNA booster vaccination (R_1) in the first emulated target trial, and under homologous (R_0) and heterologous (R_1) prime-boost vaccination in the second emulated target trial) using the fitted adjusted logistic regression model. Marginal effect estimates (risk difference, $RD = R_1 - R_0$, and relative risk, $RR = R_1/R_0$) were obtained by contrasting the estimated counterfactual risks under each intervention arm. Bootstrapping (with 1000 bootstrap replicates) was performed, using the R package ‘boot’ [50], to obtain the standard error of effect estimates. Rubin’s rules were used to pool the effect estimates and corresponding standard errors from each imputed dataset, to calculate 95% Confidence Intervals (CIs).

Furthermore, we assessed whether the estimated effect of a homologous versus heterologous prime-boost scheme on clinical severity is modified by the Omicron sub-lineage of infection. By fitting a model with the two-way interaction between the Omicron sub-lineage of infection and the vaccination schedule, counterfactual risks of the binary outcomes under exposure of the control or intervention

vaccination schedule (primary exposure) and a specific Omicron sub-lineage of infection (secondary exposure) could be estimated (e.g., $R_{1,BA.1}$ indicating the counterfactual risk for a severity outcome under the intervention vaccination schedule—primary vaccination plus mRNA booster vaccination in target trial I and heterologous prime-boost vaccination in target trial II—and when being exposed to the BA.1 lineage, $R_{0,BA.1}$ indicating the counterfactual risk for a severity outcome under the control vaccination schedule—primary vaccination only in target trial I and homologous prime-boost vaccination in target trial II—and when being exposed to the BA.1 lineage). Counterfactual risks under both intervention arms were contrasted under exposure to BA.1 ($RD_{BA.1} = R_{1,BA.1} - R_{0,BA.1}$; $RR_{BA.1} = R_{1,BA.1}/R_{0,BA.1}$), BA.2 ($RD_{BA.2} = R_{1,BA.2} - R_{0,BA.2}$; $RR_{BA.2} = R_{1,BA.2}/R_{0,BA.2}$) and BA.4/5 ($RD_{BA.4/5} = R_{1,BA.4/5} - R_{0,BA.4/5}$; $RR_{BA.4/5} = R_{1,BA.4/5}/R_{0,BA.4/5}$) lineages of infection (for an explanation on the counterfactual definition of effect modification, see the manuscript of Bours [51]). According to the counterfactual definition of effect modification, effect modification can be observed on the additive scale if at least two of the estimated risk differences ($RD_{BA.1}$, $RD_{BA.2}$ and $RD_{BA.4/5}$) are different and on the multiplicative scale if at least two of the estimated relative risks ($RR_{BA.1}$, $RR_{BA.2}$, $RR_{BA.4/5}$) are different.

3. RESULTS

3.1 Characteristics of the Study Population

On 7 December 2022 (i.e., the date of data extraction), 133,002 admission forms were registered in the CHS. From these, 117,051 records of patients not meeting the eligibility criteria or duplicate records of the same patient were excluded. This resulted in 15,951 unique patient records with available admission information and meeting the eligibility criteria. The study population selection procedures for the main and sensitivity analysis of both emulated trials are graphically represented in Figure 3.

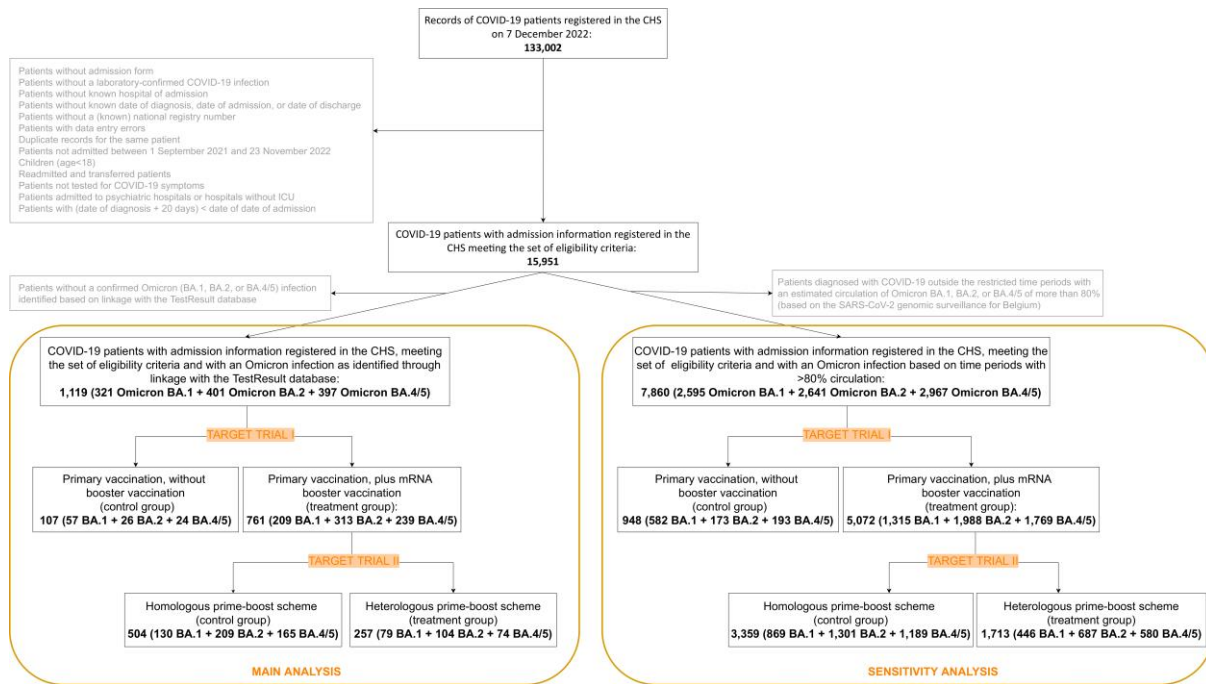


Figure 3. Flow chart representing the study selection procedures.

Main Analysis

For 1,119 out of the 15,951 patients, the Omicron variant of infection was assigned based on linkage to WGS-confirmed lineage data. From those, 868 had a vaccination schedule corresponding to one of the intervention arms of target trial I or II and were subsequently included in the analysis. They were admitted to the hospital for COVID-19 between 16 December 2021 and 10 November 2022. In the study population, 59% (513/868) were male, 11% (99/868) were a nursing home resident, and median age was 79 years (IQR: 70-86). 3% (29/868) of the patients in the study population had a documented previous infection and 56.9% (494/868) had at least one underlying comorbidity characterized by a high risk for severe COVID-19.

Patient characteristics by vaccination scheme (primary vaccinated patients versus patients with a homologous prime-boost vaccination scheme versus patients with a heterologous prime-boost vaccination scheme) are presented in Table 1. Patients who had received a booster vaccine dose in our study population tended to be older, more frequently were male and nursing home resident and more frequently had certain comorbidities (e.g., cardiovascular disease, arterial hypertension, chronic renal disease), less frequently had an identified previous infection, and more frequently were characterized by severe clinical COVID-19 outcomes (e.g., ICU admission, in-hospital mortality), compared to patients with only a primary vaccination schedule. Patient characteristics between patients who received a homologous and heterologous prime-boost vaccination scheme were more similar.

Sensitivity Analysis

For 7,860 out of the 15,951 patients, the Omicron sub-lineage of infection was assigned based on their date of COVID-19 diagnosis, falling within pre-defined time periods with known circulation of the Omicron BA.1, BA.2 and BA.4/5 sub-lineages. From those, 6020 had a vaccination schedule corresponding to one of the intervention arms of target trial I or II and were subsequently included in the analysis. They were admitted to the hospital for COVID-19 between 18 October 2021 and 23 November 2022. In the study population, 52% (3,133/6,020) were male, 14% (814/6,020) were a nursing home resident, and median age was 78 years (IQR: 67–86). 6.2% (373/6,020) of the patients in the study population had a documented previous infection and 47.4% (2,852/6,020) had at least one underlying comorbidity characterized by a high risk for severe COVID-19.

Table 1. Characteristics by vaccination scheme (primary vaccinated patients versus patients with a homologous prime-boost vaccination scheme versus patients with a heterologous prime-boost vaccination scheme) within a multi-center study to assess the effect of SARS-CoV-2 vaccination scheme on clinical outcomes among hospitalized COVID-19 patients.

	Primary Vaccination (No Booster) (n = 107)	Primary Vaccination, Plus mRNA Booster Vaccination	
		Homologous Prime- Boost Vaccination (n = 504)	Heterologous Prime- Boost Vaccination (n = 257)
	Number/Total Number (Percent)		
Demographics			
Age (years), median (IQR)	68 (51–80)	79 (72–85)	84 (74–89)
Male gender, n/N (%)	60/107 (56.1)	300/504 (59.5)	153/257 (59.5)
Nursing home resident, n/N (%)	5/106 (4.7)	68/497 (13.7)	26/256 (10.2)
Healthcare worker, n/N (%)	1/105 (1.0)	4/500 (0.8)	3/256 (1.2)
Comorbidities			
Cardiovascular Disease, n/N (%)	36/107 (33.6)	277/504 (55.0)	151/257 (58.8)
History of Arterial Hypertension, n/N (%)	39/107 (36.4)	218/504 (43.3)	104/257 (40.5)
Diabetes mellitus, n/N (%)	27/107 (25.2)	142/504 (28.2)	70/257 (27.2)
Obesity, n/N (%)	8/107 (7.5)	39/504 (7.7)	23/257 (8.9)

Chronic Pulmonary Disease, n/N (%)	28/107 (26.2)	119/504 (23.6)	74/257 (28.8)
Chronic Neurological Disease, n/N (%)	21/107 (19.6)	111/504 (22.0)	40/257 (15.6)
Chronic Cognitive Deficit, n/N (%)	15/107 (14.0)	65/504 (12.9)	28/257 (10.9)
Chronic Renal Disease, n/N (%)	22/107 (20.6)	136/504 (27.0)	85/257 (33.1)
Chronic Liver Disease, n/N (%)	4/107 (3.7)	18/504 (3.6)	5/257 (1.9)
Solid Cancer, n/N (%)	20/107 (18.7)	111/504 (22.0)	48/257 (18.7)
Hematological Cancer, n/N (%)	6/107 (5.6)	18/504 (3.6)	10/257 (3.9)
Solid organ transplantation, n/N (%)	2/107 (1.9)	7/504 (1.4)	6/257 (2.3)
Chronic Immunosuppression, n/N (%)	8/107 (7.5)	18/504 (3.6)	14/257 (5.4)
Pregnancy, n/N (%)	5/107 (4.7)	0/504 (0.0)	1/257 (0.4)
Socio-economic status			
Education level, n/N (%)			
Low	41/72 (56.9)	300/445 (67.4)	155/235 (66.0)
Middle	24/72 (33.3)	87/445 (19.6)	58/235 (24.7)
High	7/72 (9.7)	58/445 (13.0)	22/235 (9.4)
Income category, n/N (%)			
Low income	55/94 (58.5)	269/479 (56.2)	146/251 (58.2)
Middle income	32/94 (34.0)	147/479 (30.7)	75/251 (29.9)
High income	7/94 (7.4)	63/479 (13.2)	30/251 (12.0)
Population density ^a , median (IQR)	2,591 (484-2,591)	798 (390-2,591)	941 (357-2,591)
Median taxable income per capita ^b , median (IQR)	23,807 (23,807–27,514)	27,075 (23,807–28,985)	26,892 (23,807–28,855)
Exposure			
Hospital-acquired infection, n/N (%)	7/106 (6.6)	33/498 (6.6)	15/256 (5.9)

Documented previous infections ^c , n/N (%)	8/107 (7.5)	12/504 (2.4)	9/257 (3.5)
Hospital characteristics			
Mean ICU occupancy during hospital stay, median (IQR)	13 (7–20)	8 (4–14)	9 (5–16)
Clinical outcomes			
Severe ^d COVID-19, n/N (%)	29/106 (27.4)	98/501 (19.6)	40/256 (15.6)
ICU admission, n/N (%)	18/107 (16.8)	54/504 (10.7)	19/257 (7.4)
In-hospital mortality, n/N (%)	17/105 (16.2)	57/500 (11.4)	30/256 (11.7)
ARDS, n/N (%)	4/107 (3.7)	18/504 (3.6)	6/257 (2.3)
Hospital length of stay (days), median (IQR)	9 (4–14)	9 (5–16)	9 (5–17)

^a Population density at the patient's place of residence (postal code level)

^b median net taxable income per capita at the patient's place of residence (postal code level)

^c number of patients with one or multiple documented previous infections, i.e., with the date of diagnosis of the SARS-CoV-2 infection causing hospitalization more than 60 days after the date of sample collection from a previous documented positive COVID-19 test (RT-PCR or AG); ^d admission to an intensive care unit (ICU), occurrence of acute respiratory distress syndrome (ARDS), or in-hospital mortality.

3.2 Effect Estimates

3.2.1 Target Trial I: Primary vaccination plus mRNA booster vaccination versus primary vaccination without booster vaccination

Main Analysis

The effect estimates (RDs and RRs) for the three considered clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality) in the study population, are presented in Table 2. The estimated counterfactual risk for severe COVID-19 (defined based on a combination of severity indicators, including ARDS, ICU admission, and in-hospital mortality) among hospitalized COVID-19 patients in Belgium infected with the BA.1, BA.2, or BA.4/5 Omicron sub-lineage (as identified through linkage with WGS-confirmed lineage data) was significantly lower when having received an mRNA booster vaccination compared to when only having received a primary vaccination schedule (RD = -0.13 [-0.24; -0.01], RR = 0.59 [0.33; 0.85]). In addition, when only looking at in-hospital mortality as severity indicator, the estimated risk was significantly lower when having received the booster dose (RD = -0.13

$[-0.25; 0.00]$, $RR = 0.47 [0.15; 0.79]$). There was no significant difference in counterfactual risk for ICU admission ($RD = -0.04 [-0.11; 0.04]$, $RR = 0.74 [0.22; 1.26]$).

Table 2. Counterfactual risk of potential outcomes under different vaccination schedules (R , in %), risk difference (RD) and relative risk (RR) estimates, and 95% confidence interval (CI) within a multi-center study estimating the effect of the SARS-CoV-2 vaccination schedule (primary vaccination plus mRNA booster vaccination versus primary vaccination without booster vaccination) on severe clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality) of BA.1, BA.2 or BA.4/5 Omicron infection (identified through linkage with WGS-confirmed lineage data) among hospitalized COVID-19 patients in Belgium.

	Primary Vaccination, without Booster Vaccination (Control Group)	Primary Vaccination, Plus mRNA Booster Vaccination (Intervention Group)	Intervention Effect	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Severe COVID-19	0.31 [0.19; 0.42]	0.18 [0.15; 0.21]	-0.13 [-0.24; -0.01]	0.59 [0.33; 0.85]
ICU admission	0.14 [0.06; 0.21]	0.10 [0.08; 0.12]	-0.04 [-0.11; 0.04]	0.74 [0.22; 1.26]
In-hospital mortality	0.24 [0.12; 0.35]	0.11 [0.09; 0.13]	-0.13 [-0.25; 0.00]	0.47 [0.15; 0.79]

Sensitivity Analysis

The effect estimates (RD s and RR s) for the three considered clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality) in the study population, are presented in Table A2 (Appendix C). As in the main analysis, in the sensitivity analysis the estimated counterfactual risk for severe COVID-19 (defined based on a combination of severity indicators, including ARDS, ICU admission, and in-hospital mortality) and in-hospital mortality among hospitalized COVID-19 patients in Belgium infected with the BA.1, BA.2, or BA.4/5 Omicron sub-lineage (as defined by the time period of COVID-19 diagnosis) was significantly lower when having received an mRNA booster vaccination as compared to when only having received a primary vaccination schedule (severe COVID-19: $RD = -0.05 [-0.08; -0.02]$, $RR = 0.75 [0.61; 0.88]$; in-hospital mortality: $RD = -0.05 [-0.08; -0.02]$, $RR = 0.67 [0.51; 0.82]$). No significant difference in standardized risk for ICU admission was observed ($RD = -0.01 [-0.03; 0.01]$, $RR = 0.87 [0.63; 1.11]$).

3.2.2 Target Trial II: Heterologous versus homologous prime-boost vaccination

Main Analysis

The effect estimates (RDs and RRs) for the three considered clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality), overall and under exposure to Omicron sub-lineage BA.1, BA.2 and BA.4/5, separately, are presented in Table 3. No significance difference was observed in estimated counterfactual risk for severe COVID-19 (defined based on a combination of severity indicators, including ARDS, ICU admission, and in-hospital mortality), ICU admission and in-hospital mortality among hospitalized COVID-19 patients in Belgium infected with the BA.1, BA.2 or BA.4/5 Omicron sub-lineage (as identified through linkage with WGS-confirmed lineage data) when having received a heterologous prime-boost vaccination as compared to when having received a homologous schedule (severe COVID-19: RD = -0.04 [-0.10; 0.02], RR = 0.80 [0.53; 1.07]; ICU admission: RD = -0.03 [-0.07; 0.02], RR = 0.73 [0.34; 1.12]; in-hospital mortality: RD = 0.00 [-0.05; 0.04], RR = 0.96 [0.55; 1.38]). Furthermore, we assessed whether this estimated effect was homogenous across different Omicron sub-lineages. We observed no significant additive and multiplicative modification by the Omicron sub-lineage of the effect of the vaccination schedule (heterologous versus homologous prime-boost vaccination) on severe COVID-19, ICU admission, and in-hospital mortality, since none of the estimated risk differences and relative risks were significantly different from each other (see Table A3, Appendix D).

Table 3. Counterfactual risk of potential outcomes under intervention (*R*, in %), risk difference (RD) and relative risk (RR) estimates, and 95% confidence interval (CI) predicted with the Omicron sub-lineage as observed (overall, identified through linkage with WGS test results) and under exposure to Omicron sub-lineage BA.1, BA.2 and BA.4/5, separately, within a multi-center cohort study estimating the effect of the SARS-CoV-2 vaccination schedule (heterologous versus homologous prime-boost vaccination) on severe clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality; different outcomes indicated in bold) of BA.1, BA.2 or BA.4/5 Omicron infection (identified through linkage with WGS-confirmed lineage data) among hospitalized COVID-19 patients in Belgium.

Outcome: Severe COVID-19				
	Homologous Prime-Boost Scheme (Control Group)	Heterologous Prime-Boost Scheme (Intervention Group)	Intervention Effect	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Overall	0.20 [0.16; 0.23]	0.16 [0.11; 0.20]	-0.04 [-0.10; 0.02]	0.80 [0.53; 1.07]

Omicron BA.1	0.19 [0.12; 0.27]	0.13 [0.06; 0.20]	-0.07 [-0.16; 0.03]	0.65 [0.22; 1.07]
Omicron BA.2	0.19 [0.13; 0.24]	0.14 [0.08; 0.21]	-0.04 [-0.13; 0.04]	0.77 [0.32; 1.23]
Omicron BA.4/5	0.21 [0.13; 0.29]	0.21 [0.10; 0.33]	0.00 [-0.12; 0.12]	1.01 [0.41; 1.61]
Outcome: ICU admission				
	Homologous prime-boost scheme (Control group)	Heterologous prime-boost scheme (Intervention group)	Intervention effect	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Overall	0.11 [0.08; 0.13]	0.08 [0.04; 0.11]	-0.03 [-0.07; 0.02]	0.73 [0.34; 1.12]
Omicron BA.1	0.08 [0.03; 0.13]	0.04 [0.00; 0.08]	-0.04 [-0.10; 0.02]	0.51 [0.00; 1.15]
Omicron BA.2	0.10 [0.06; 0.15]	0.06 [0.01; 0.11]	-0.04 [-0.11; 0.02]	0.59 [0.00; 1.19]
Omicron BA.4/5	0.15 [0.08; 0.23]	0.17 [0.06; 0.28]	0.02 [-0.10; 0.13]	1.12 [0.27; 1.97]
Outcome: In-hospital mortality				
	Homologous prime-boost scheme (Control group)	Heterologous prime-boost scheme (Intervention group)	Intervention effect	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Overall	0.12 [0.09; 0.15]	0.11 [0.07; 0.15]	0.00 [-0.05; 0.04]	0.96 [0.55; 1.38]
Omicron BA.1	0.14 [0.07; 0.20]	0.11 [0.04; 0.18]	-0.03 [-0.11; 0.06]	0.81 [0.18; 1.45]
Omicron BA.2	0.12 [0.07; 0.16]	0.11 [0.05; 0.16]	-0.01 [-0.09; 0.06]	0.89 [0.23; 1.56]
Omicron BA.4/5	0.09 [0.03; 0.15]	0.13 [0.04; 0.23]	0.04 [-0.06; 0.14]	1.42 [0.00; 2.53 × 10 ⁵]

Sensitivity Analysis

The effect estimates (RDs and RRs) for the three considered clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality), overall and under exposure to Omicron sub-lineage BA.1, BA.2 and BA.4/5 separately, are presented in Table A4 (Appendix E). No significance difference was observed in estimated counterfactual risk for severe COVID-19 (defined based on a combination of severity indicators, including ARDS, ICU admission, and in-hospital mortality), ICU admission and in-hospital mortality among hospitalized COVID-19 patients in Belgium infected with the BA.1, BA.2 or BA.4/5 Omicron sub-lineages (as defined by the time period of COVID-19 diagnosis) when having received a heterologous prime-boost vaccination as compared to having received a homologous schedule (severe

COVID-19: RD = 0.01 [-0.01 ; 0.03], RR = 1.06 [0.91; 1.21]; ICU admission: RD = 0.00 [-0.01; 0.02], RR = 1.02 [0.78; 1.26]; in-hospital mortality: RD = 0.00 [-0.02; 0.02], RR = 1.01 [0.83; 1.19]). Furthermore, we observed no significant additive and multiplicative modification by the Omicron sub-lineage of the effect of the vaccination schedule (heterologous versus homologous prime-boost vaccination) on severe COVID-19, ICU admission and in-hospital mortality, since none of the estimated risk differences and relative risks were significantly different from each other (see Table A5, Appendix F).

4. DISCUSSION

In this study, we compared the effectiveness of (1) an mRNA booster vaccine versus primary vaccination only and (2) a heterologous (viral vector–mRNA) versus homologous (mRNA–mRNA) prime-boost vaccination in preventing severe clinical outcomes of SARS-CoV-2 Omicron infection (BA.1, BA.2 or BA.4/5 Omicron sub-lineage) among hospitalized COVID-19 patients in Belgium, and whether this effect was homogenous across the different Omicron sub-lineages. This study was conducted using observational data from national COVID-19 registries.

We observed that administering a booster dose (mRNA vaccine) in addition to a homogenous primary vaccination (any type) schedule significantly reduced the estimated risk of progression of an BA.1, BA.2 or BA.4/5 Omicron infection to severe COVID-19 (experiencing an ARDS event, or ICU transfer, or in-hospital mortality) and in-hospital mortality in the population of patients treated for COVID-19 in a Belgian hospital. This is in line with results from previous studies, in which the effectiveness of an mRNA booster vaccination against severe Omicron infections was found to increase, compared to patients with only a primary vaccination [26,52,53]. Therefore, based on previous and the current results, in a period of predominant Omicron circulation, the administration of a booster vaccine dose can be recommended. We observed a similar, though weaker, and non-significant, trend in effectiveness of mRNA booster vaccination against ICU admission of patients treated for COVID-19 in a Belgian hospital. While this could potentially be attributed partly to limited statistical power due to the small sample size, other factors may also have influenced these findings. Further investigation with larger datasets is warranted to better understand this relationship. Furthermore, in contrast to severity outcomes such as mortality or ARDS, ICU admission reflects clinical decision making which partly depends on contextual factors (such as therapeutic guidelines in place and varying healthcare capacity). As the median age was 79 and 84 years in the cohort with a homologous and heterologous prime-boost vaccination schedule respectively, (versus 68 years among those that received a primary vaccination schedule only), it could be that this older patient population was less likely to be admitted

into ICU due to perceived clinical futility. Further investigations in a younger population (e.g., excluding patients older than 85 years old) could be carried out in the future.

No difference in estimated risk of severe BA.1, BA.2 or BA.4/5 Omicron infection was observed between a homologous and heterologous prime-booster scheme with an mRNA booster vaccine in the study population. Early results suggested that mixing of vaccine types for primary and booster vaccination might result in a stronger immune response [19–21] and therefore might offer higher protection against severe COVID-19. However, this was not observed in our study population of hospitalized COVID-19 patients in Belgium who were infected with Omicron BA.1, BA.2 or BA.4/5. This might be related to the SARS-CoV-2 lineages we consider in this study, namely Omicron BA.1, BA.2, BA.4 or BA.5, characterized by their ability to more efficiently evade vaccine-induced neutralizing antibodies compared to previously circulating variants [24–27]. Furthermore, in this study we compare the effectiveness of SARS-CoV-2 ‘viral vector–mRNA’ and ‘mRNA–mRNA’ prime-boost vaccination and results might depend on the vaccine types and order of the vaccine types used for primary and booster vaccination. Previously, there might have been a concern for less protection against severe COVID-19 disease in people vaccinated with the less effective Ad26.COVS.S and ChAdOx1-S vaccines [54]. However, our results indicate that after booster vaccination with the BNT162b2 or mRNA-1273 vaccine, the type of vaccine used for primary vaccination might no longer be of concern and ‘viral vector–mRNA’ and ‘mRNA–mRNA’ prime-boost vaccination schemes might provide similar protection against severe Omicron infection for hospitalized COVID-19 patients in Belgium. We found, during the circulation of BA.1, BA.2 and BA.4/5 Omicron sub-lineages, no evidence of a difference in protection against severe clinical outcomes between homologous and heterologous prime boost between Omicron sub-lineages.

This study has several important strengths. First, detailed patient information was available through the CHS, to which data on socio-economic characteristics of the patients, previous documented infections, sequencing results for identification of the SARS-CoV-2 Omicron sub-lineage of infection, and exhaustive data on administered vaccine doses could be linked. This allowed us to control for many of the identified confounders, making our effect estimates more reliable and limiting confounder bias. Second, even though the CHS is not exhaustive, hospitals reporting to the CHS were geographically distributed over the country, resulting in an increased generalizability of the results for the entire population of hospitalized COVID-19 patients in Belgium. Third, we took into account different clinical severity outcomes of COVID-19, including a composite measure (i.e., severe COVID-19, defined as experiencing an acute respiratory distress syndrome (ARDS) event, or an ICU admission, or in-hospital mortality), a severity measure dependent on clinical decision making (i.e., ICU admission), and a ‘hard’

objective severity measure (i.e., in-hospital mortality). These different outcomes might capture different aspects of vaccine effectiveness and might inform policy in an alternative way.

There are also some limitations to the study. First, in the main analysis, the variant of infection was assigned based on linkage with WGS-confirmed lineage data from the COVID-19 TestResult database. This might result in selection bias and drastically limits the sample size [45]. The small sample sizes reduce the statistical power and the reliability of significant findings [55]. Furthermore, the lineage data obtained from the COVID-19 TestResult database might be slightly biased due to delayed adjustment of the data entry tool when new sub-lineages were defined. A sensitivity analysis was conducted where the variant of infection was assigned based on the timing of diagnosis during restricted time periods with a predominant circulation of BA.1, BA.2 and BA.4/5. Conclusions from this sensitivity analysis remain the same and are not affected by the method of lineage assignment, indicating the robustness of our findings. Second, the DAG was constructed and a sufficient adjustment set was determined based on a priori assumptions about the causal relationship between the exposure and outcome. Hence, there might exist residual bias from unidentified confounders. Furthermore, host genetics was identified as a confounder in the DAG, but was unobserved, and we were not able to adjust for this variable. In addition, the sufficient adjustment set was approximated by available information. For example, the considered previous infections entail only documented previous infections, which is dependent on the testing strategy and care-seeking behavior. Furthermore, alternative definitions for a reinfection can be adopted [47,56]. Our definition of a reinfection was based on the recommendations of the European Center for Disease Prevention and Control (ECDC) to take a minimal period of 60 days to elapse since the previous documented infection to consider an infection as a reinfection [47]. The hospital to which the patient admitted was identified as a confounder and was taken into account by adjusting for the hospital type, not the individual hospital, since (due to the small sample size) including the individual hospital in the models resulted in complete separation. Treatment strategies were assumed to be constant during the study period; however, if this assumption is not correct, adjustment for this confounder is required. Hence, residual confounding bias might be present in the effect estimates. We also have to account for the possibility of model misspecification, while no model misspecification is an assumption for obtaining unbiased estimates through standardization. Third, to investigate if the estimated effect of a homologous versus heterologous prime-booster scheme on clinical severity is modified by the Omicron sub-lineage of infection, the Omicron BA.4 and BA.5 lineages were combined. Viral characteristics (e.g., virulence and transmissibility) of these variants might differ; however, due to the small sample size and low prevalence of BA.4 infections, we were not able to separate the latter sub-lineage in our analysis. However, the BA.4 and BA.5 lineages were co-circulating in time and the Omicron sub-lineage was not

our primary exposure of interest. Fourth, we aimed to only include hospitalized patients registered in the CHS that are treated for COVID-19; however, it is difficult to deduce from the collected data in the CHS for which reason a patient was treated in the hospital. Furthermore, the cause of the documented severe clinical outcomes was difficult to assess from the available data. Fifth, for performing multiple imputation of missing values, the assumption of missingness at random was made. However, the underlying missing data mechanism is unknown. Sixth, different selection steps were performed to obtain data of our study population (i.e., selection of patients tested for COVID-19, treated in the hospital for COVID-19 symptoms, registered in the CHS, and with an Omicron BA.1, BA.2, or BA.4/5 infection identified by linking WGS-confirmed lineage data) [45]. Hence, patients in our study population might have specific characteristics distinctive from the population of all hospitalized patients in Belgium and the general population, thereby limiting generalization. It is important to consider these study selection procedures when interpreting the estimated counterfactual risks and effect estimates (RDs and RRs) tied to the particular population across which the marginalization was carried out.

5. CONCLUSIONS

Our results support evidence that the administration of an mRNA booster vaccine reduces the risk of severe COVID-19 disease among hospitalized patients in a time with predominant Omicron circulation and suggest no difference in protection offered by the contrasted prime-boost vaccination schemes. These results can inform vaccine policy in Belgium, and contribute to knowledge used for the development of strategies for booster campaigns during the continuing COVID-19 epidemic.

6. DECLARATIONS

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. Ethical approval was granted for the gathering of data from hospitalized patients by the Committee for Medical Ethics from the Ghent University Hospital (reference number BC07507) and authorization for possible individual data linkage using the national register number from the Information Security Committee (ISC) Social Security and Health (reference number IVC/KSZG/20/384). Linkage of hospitalized patient data to vaccination, testing, sequencing, healthcare actor, and socio-economic data within the LINK-VACC project was approved by the Medical Ethics Committee UZ Brussels–VUB on 3 February 2021 (reference number 2020/523), and authorization from the ISC Social Security and Health (reference number IVC/KSZG/21/034).

Informed Consent Statement: Informed consent was waived based on art. 6 and 9 of the GDPR. The collection is allowed based on general interest (art. 6 GDPR) and regarding article 9 § 2 of the GDPR: processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health, or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular, professional secrecy.

Data Availability Statement: The individual level datasets generated or analyzed during the current study do not fulfill the requirements for open data access. The data is too dense and comprehensive to preserve patient privacy. The data of the individual data sources (Clinical Hospital Survey, Vaccinnet+, COVID-19 TestResult Database, StatBel, Surge Capacity Survey, and Common Base Registry for HealthCare Actor) within the LINK-VACC project are kept in the pseudonymized environment of healthdata.be and a link between the individual data in each of them takes place thanks to the use of a pseudonymized national reference number managed by healthdata.be under a "project mandate". A "project mandate" consists of a group of individuals, a group of variables and a time period. Access rights to the pseudonymized data in the healthdata.be data warehouse are granted ad nominatum for the scientists involved in the surveillance activities at Sciensano. External investigators with a request for selected data should fill in the data request form (<https://epistat.sciensano.be/datarequest/>, accessed on December 16, 2022). Depending on the type of desired data (anonymous or pseudonymized), the provision of data will have to be assessed by the Belgian Information Security Committee Social Security & Health based on legal and ethical regulations, and is outlined in a data transfer agreement with the data owner (Sciensano).

Acknowledgments: We are very grateful to all clinicians and hospital directions which allowed this systematic and timely data collection and reporting to Sciensano. In particular, we would like to acknowledge the Belgian Collaborative Group on COVID-19 Hospital surveillance: Amir-Samy Aouachria (CHC), Kristof Bafort (MZNL), Leïla Belkhir (Cliniques universitaires St-Luc), Steven Callens (UZ Gent), Vincent Colombie (EpiCURA), Nicolas Dauby (CHU Saint-Pierre), Paul De Munter (UZ Leuven), Pieter Depuydt (UZ Gent), Jessika Deblonde (Sciensano), Didier Delmarcelle (Clinique Saint Jean), Mélanie Delvallee (CHwapi), Rémy Demeester (CHU de Charleroi), Thierry Dugernier (Clinique St-Pierre), Caroline Gheysen (AZ Jan Yperman), Xavier Holemans (GHdC), Frederic Thys (GHdC), Benjamin Kerzmann (CNDG), Sarah Loof (AZ Maria Middelaes), Pierre Yves Machurot (CH de l'Ardenne), Peter Messiaen (Jessa Hospital), Philippe Minette (CHR Jolimont), Jean-Marc Minon (CHR de la Citadelle), Saphia Mokrane (Hôpitaux Iris Sud), Catherine Nachtergal (Cliniques de l'Europe), Séverine Noirhomme (CHR de Namur), Denis Piérard (UZ Brussel), Camelia Rossi (CHU Ambroise Paré),

Carole Schirvel (CHIREC), Erica Sermijn (ASZ Hospital), Frank Staelens (OLV Hospital), Fabio Silvio Taccone (Hôpital Erasme), Filip Triest (AZ Sint Lucas), Eva Van Braeckel (UZ Gent), Jens Van Praet (AZ Sint-Jan), Anke Vanhoenacker (ZNA), Roeland Verstraete (AZ Monica), and Elise Willems (AZ Nikolaas). We would also like to thank all sequencing laboratories for uploading SARS-CoV-2 variant information to the COVID-19 TestResult database. In particular, we would like to thank the members of the COVID-19 Genomics Belgium consortium: Emmanuel André (UZ Leuven), Piet Maes (KU Leuven), Guy Baele (KU Leuven), Simon Dellicour (KU Leuven), Lize Cuypers (UZ Leuven), Marc Van Ranst (UZ Leuven), Barney Potter (KU Leuven), Samuel Hong (KU Leuven), Ruddy Wattiez (UMONS), Carl Vael (AZ Klina), Lynsey Berckmans (AZ Klina), Philippe Selhorst (ITG), Kevin K. Ariën (ITG), Arnaud Marchant (ULB), Coralie Henin (ULB), Benoit Haerlingen (ULB), Ricardo De Mendonca (ULB), Marie-Luce Delforge (ULB), Sonia Van Dooren (UZ Brussel), Bruno Hinckel (UZ Brussel), Hideo Imamura (UZ Brussel), Toon Janssen (UZ Brussel), Ben Caljon (UZ Brussel), Oriane Soetens (UZ Brussel), Denis Piérard (UZ Brussel), Thomas Demuyser (UZ Brussel), Charlotte Michel (LHUB-ULB), Olivier Vandenberg (LHUB-ULB), Sigi van den Wijngaert (LHUB-ULB), Giulia Zorzi (LHUB-ULB), Benoit Kabamba Mukadi (UC Louvain), Jean-Luc Gala (UC Louvain), Bertrand Bearzatto (UC Louvain), Jérôme Ambroise UC Louvain), Philippe Van Lint (GZA), Walter Verstrepen (ZNA), Reinout Naesens (ZNA), Kate Bakelants (Sciensano), Sarah Denayer (Sciensano), Sofieke Klamer (Sciensano), François E. Dufrasne (Sciensano), Pascale Hilbert (IPG), Sylvain Brohée (IPG), Pierre-Emmanuel Léonard (IPG), Deniz Karadurmus (IPG), Jeremie Gras (IPG), Damien Féret (IPG), Barbara Lambert (IPG), Anne Vankeerberghen (OLVZ Aalst), Astrid Holderbeke (OLVZ Aalst), Hans De Beenhouwer (OLVZ Aalst), Lien Cattoir (OLVZ Aalst), Christine Lammens (U Antwerpen), Basil Britto Xavier (U Antwerpen), Marie Le Mercier (UZ Antwerpen), Jasmine Coppens (UZ Antwerpen), Veerle Matheeussen (UZ Antwerpen), Herman Goossens (UZ Antwerpen), Geert A. Martens (AZ Delta), Koen Swaert (AZ Delta), Frederik Van Hoecke (AZ Delta), Dieter Desmet (AZ Delta), Patrick Descheemaeker (AZ Delta), Pierre Bogaerts (UC Louvain), Jonathan Degosserie (UC Louvain), Olivier Denis (UC Louvain), Te-Din Huang (UC Louvain), Dagmar Obbels (Imelda Hospital), Hanne Valgaeren (Imelda Hospital), Johan Frans (Imelda Hospital), Annick Smismans (Imelda Hospital), Paul-Emile Claus (Imelda Hospital), Truus Goegebuer (AZ St-Maarten), Ann Lemmens (AZ St-Maarten), Bea Van den Poel (AZ Jan Portaels), Sonja De Bock (AZ Jan Portaels), Wim Laffut (Heilig Hart Hospital Lier), Ellen Van Even (Heilig Hart Hospital Lier), Jos Van Acker (AZ St-Lucas Gent), Charlotte Verfaillie (AZ St-Lucas Gent), Elke Vanlaere (AZ St-Lucas Gent), Klara De Rauw (AZ St-Lucas Gent), Luc Waumans (Jessa Hospital), Britt Van Meensel (Jessa Hospital), Reinoud Cartuyvels (Jessa Hospital), Severine Berden (Jessa Hospital), Marijke Raymaekers (Jessa Hospital), Bruno Verhasselt (UZ Gent), Cécile Meex (CHU de Liège), Keith Durkin (CHU de Liège), Laurent Gillet (CHU de Liège), Maria Artesi (CHU de Liège), Marie-Pierre Hayette (CHU de Liège), Sébastien Bontems (CHU de Liège), Vincent Bours (CHU de Liège),

Claire Gourzonès (CHU de Liège), Olivier Ek (CHU de Liège), Fabrice Bureau (CHU de Liège), Jorn Hellemans (AZ St-Jan), Merijn Vanhee (AZ St-Jan), and Marijke Reynders (AZ St-Jan). Finally, we acknowledge all persons who played a key role by processing the data technically, especially the colleagues of healthdata.be, eHealth, STATBEL, and IMA.

Conflicts of Interest: The authors declare no conflict of interest.

7. REFERENCES

1. Lamers MM, Haagmans BL. SARS-CoV-2 pathogenesis. *Nat Rev Microbiol.* 2022;20:270–284.
2. Haas EJ, Angulo FJ, McLaughlin JM, Anis E, Singer SR, Khan F, et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet.* 2021;397(10287):1819–1829.
3. Mazagatos C, Monge S, Olmedo C, Vega L, Gallego P, Martín-Merino E, et al. Effectiveness of mRNA COVID-19 vaccines in preventing SARS-CoV-2 infections and COVID-19 hospitalisations and deaths in elderly long-term care facility residents, Spain, weeks 53 2020 to 13 2021. *Eurosurveillance.* 2021;26(24):2100452.
4. Martínez-Baz I, Miqueleiz A, Casado I, Navascués A, Trobajo-Sanmartín C, Burgui C, et al. Effectiveness of COVID-19 vaccines in preventing SARS-CoV-2 infection and hospitalisation, Navarre, Spain, January to April 2021. *Eurosurveillance.* 2021;26(21):2100438.
5. Olson SM, Newhams MM, Halasa NB, Price AM, Boom JA, Sahni LC, et al. Effectiveness of BNT162b2 Vaccine against Critical Covid-19 in Adolescents. *New England Journal of Medicine.* 2022;386(8):713–723.
6. Pawlowski C, Lenehan P, Puranik A, Agarwal V, Venkatakrishnan AJ, Niesen MJM, et al. FDA-authorized mRNA COVID-19 vaccines are effective per real-world evidence synthesized across a multi-state health system. *Med.* 2021;2(8):979–992.
7. Moghadas SM, Vilches TN, Zhang K, Wells CR, Shoukat A, Singer BH, et al. The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States. *Clinical Infectious Diseases.* 2021;73(12):2257–2264.
8. Federal Public Service (FPS) Health, Food Chain Safety and Environment. Vaccination | Coronavirus COVID-19. 2022 [cited 15 Mar 2022]. Available: <https://www.info-coronavirus.be/en/vaccination/>
9. Sciensano. COVID-19 – EPIDEMIOLOGISCH BULLETIN VAN 31 MEI 2022. 2022 May. Available: <https://covid-19.sciensano.be/sites/default/files/Covid19/Meest%20recente%20update.pdf>
10. Sciensano. COVID-19 WEKELIJKS EPIDEMIOLOGISCH BULLETIN (1 DECEMBER 2022). Sciensano; 2022 Jan. Available: https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_Weekly_report_NL.pdf
11. Bellamkonda N, Lambe UP, Sawant S, Nandi SS, Chakraborty C, Shukla D. Immune Response to SARS-CoV-2 Vaccines. *Biomedicines.* 2022;10(7):1464.

12. Kashte S, Gulbake A, El-Amin III SF, Gupta A. COVID-19 vaccines: rapid development, implications, challenges and future prospects. *Human Cell*. 2021;34:711–733.
13. Ndwandwe D, Wiysonge CS. COVID-19 vaccines. *Current Opinion in Immunology*. 2021;71: 111–116.
14. Vanaparthi R, Mohan G, Vasireddy D, Atluri P. Review of COVID-19 viral vector-based vaccines and COVID-19 variants. *Infez Med*. 2021;29(3):328–338.
15. Huang Q, Zeng J, Yan J. COVID-19 mRNA vaccines. *Journal of Genetics and Genomics*. 2021;48(2):107–114.
16. Upreti S, Samant M. A Review on Immunological Responses to SARS-CoV-2 and Various COVID-19 Vaccine Regimens. *Pharm Res*. 2022;39(9):2119–2134.
17. Lu S. Heterologous prime–boost vaccination. *Current Opinion in Immunology*. 2009;21(3):346–351.
18. Levine MZ, Holiday C, Jefferson S, Gross FL, Liu F, Li S, et al. Heterologous prime-boost with A(H5N1) pandemic influenza vaccines induces broader cross-clade antibody responses than homologous prime-boost. *npj Vaccines*. 2019;4(1):1–10.
19. Sapkota B, Saud B, Shrestha R, Al-Fahad D, Sah R, Shrestha S, et al. Heterologous prime–boost strategies for COVID-19 vaccines. *Journal of Travel Medicine*. 2022;29(3):taab191.
20. Cohen G, Jungsomsri P, Sangwongwanich J, Tawinprai K, Siripongboonsitti T, Porntharukchareon T, et al. Immunogenicity and reactogenicity after heterologous prime-boost vaccination with CoronaVac and ChAdox1 nCov-19 (AZD1222) vaccines. *Human Vaccines & Immunotherapeutics*. 2022;18(5): 2052525.
21. Khoo NKH, Lim JME, Gill US, de Alwis R, Tan N, Toh JZN, et al. Differential immunogenicity of homologous versus heterologous boost in Ad26.COV2.S vaccine recipients. *Med (N Y)*. 2022;3(2):104–118.e4.
22. Chemaitelly H, Tang P, Hasan MR, AlMukdad S, Yassine HM, Benslimane FM, et al. Waning of BNT162b2 Vaccine Protection against SARS-CoV-2 Infection in Qatar. *New England Journal of Medicine*. 2021;385:e83.
23. Hirabara SM, Serdan TDA, Gorjao R, Masi LN, Pithon-Curi TC, Covas DT, et al. SARS-COV-2 Variants: Differences and Potential of Immune Evasion. *Frontiers in Cellular and Infection Microbiology*. 2022;11.
24. Li X. Omicron: Call for updated vaccines. *Journal of Medical Virology*. 2022;94(4):1261–1263.
25. Garcia-Beltran WF, St. Denis KJ, Hoelzemer A, Lam EC, Nitido AD, Sheehan ML, et al. mRNA-based COVID-19 vaccine boosters induce neutralizing immunity against SARS-CoV-2 Omicron variant. *Cell*. 2022;185(3):457–466.e4.
26. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. *New England Journal of Medicine*. 2022;386(16):1532–1546.
27. Hoffmann M, Krüger N, Schulz S, Cossmann A, Rocha C, Kempf A, et al. The Omicron variant is highly resistant against antibody-mediated neutralization: Implications for control of the COVID-19 pandemic. *Cell*. 2022;185(3):447–456.e11.

28. Tartof SY, Slezak JM, Puzniak L, Hong V, Frankland TB, Ackerson BK, et al. BNT162b2 vaccine effectiveness against SARS-CoV-2 omicron BA.4 and BA.5. *The Lancet Infectious Diseases*. 2022;22(12):1663-1665.
29. Jacobsen H, Katzmarzyk M, Higdon MM, Jiménez VC, Sitaras I, Bar-Zeev N, et al. Post-Vaccination Neutralization Responses to Omicron Sub-Variants. *Vaccines*. 2022;10(10):1757.
30. Björk J, Bonander C, Moghaddassi M, Rasmussen M, Malmqvist U, Inghammar M, et al. COVID-19 vaccine effectiveness against severe disease from SARS-CoV-2 Omicron BA.1 and BA.2 subvariants – surveillance results from southern Sweden, December 2021 to March 2022. *Eurosurveillance*. 2022;27(18):2200322.
31. Vanmechelen B, Logist A-S, Wawina-Bokalanga T, Verlinden J, Martí-Carreras J, Geenen C, et al. Identification of the First SARS-CoV-2 Lineage B.1.1.529 Virus Detected in Europe. *Microbiol Resour Announc*. 2022;11(1):e0116121.
32. UZ Leuven, KU Leuven. Genomic surveillance of SARS-CoV-2 in Belgium - Report of the National Reference Laboratory. Leuven, Belgium; 2022 Feb. Available: <https://www.uzleuven.be/nl/laboratoriumgeneeskunde/genomic-surveillance-sars-cov-2-belgium/genomic-surveillance-sars-cov-2-belgium-reports>
33. UZ Leuven, KU Leuven. Genomic surveillance of SARS-CoV-2 in Belgium - Report of the National Reference Laboratory. Leuven, Belgium; 2022 Mar. Available: <https://www.uzleuven.be/nl/laboratoriumgeneeskunde/genomic-surveillance-sars-cov-2-belgium/genomic-surveillance-sars-cov-2-belgium-reports>
34. UZ Leuven, KU Leuven. Genomic surveillance of SARS-CoV-2 in Belgium - Report of the National Reference Laboratory. Leuven, Belgium; 2022 Jun. Available: <https://www.uzleuven.be/nl/laboratoriumgeneeskunde/genomic-surveillance-sars-cov-2-belgium/genomic-surveillance-sars-cov-2-belgium-reports>
35. UZ Leuven. Genomic surveillance of SARS-CoV-2 in Belgium. 19 Dec 2022 [cited 19 Dec 2022]. Available: <https://www.uzleuven.be/nl/laboratoriumgeneeskunde/genomic-surveillance-sars-cov-2-belgium>
36. Saito A, Irie T, Suzuki R, Maemura T, Nasser H, Uriu K, et al. Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature*. 2022;602:300–306.
37. Dang S, Ren L, Wang J. Functional mutations of SARS-CoV-2: implications to viral transmission, pathogenicity and immune escape. *Chinese Medical Journal*. 2022;135(10):1213–1222.
38. Van Goethem N, Vilain A, Wyndham-Thomas C, Deblonde J, Bossuyt N, Lernout T, et al. Rapid establishment of a national surveillance of COVID-19 hospitalizations in Belgium. *Archives of Public Health*. 2020;78:121.
39. healthdata.be (Sciensano). DATABASE COVID-19 TESTRESULTS - LaboratoryTestResult. 22 Jun 2022. Available: <https://covid19lab.healthdata.be/data-collection/laboratorytestresult>
40. healthdata.be (Sciensano). DATABASE COVID-19 TESTRESULTS - LaboratoryTestResultVariants. 22 Jun 2022. Available: <https://covid19lab.healthdata.be/laboratorytestresultvariants>

41. Meurisse M, Lajot A, Dupont Y, Lesenfans M, Klamer S, Rebolledo J, et al. One year of laboratory-based COVID-19 surveillance system in Belgium: main indicators and performance of the laboratories (March 2020–21). *Archives of Public Health*. 2021;79:188.
42. Zorg en Gezondheid (Vlaamse overheid). Vaccinatie Systeem (Vaccinnet). 3 Sep 2022 [cited 20 Jun 2022]. Available: <https://www.vaccinnet.be/Vaccinnet/welkom.do>
43. eHealth. CoBRHA - Common Base Registry for HealthCare Actor. 20 Jun 2022 [cited 20 Jun 2022]. Available: <https://www.ehealth.fgov.be/ehealthplatform/nl/service-cobrha-common-base-registry-for-healthcare-actor>
44. Van Goethem N, Serrien B, Vandromme M, Wyndham-Thomas C, Catteau L, Brondeel R, et al. Conceptual causal framework to assess the effect of SARS-CoV-2 variants on COVID-19 disease severity among hospitalized patients. *Archives of Public Health*. 2021;79:185.
45. Meurisse M, Van Oyen H, Blot K, Catteau L, Serrien B, Klamer S, et al. Evaluating methodological approaches to assess the severity of infection with SARS-CoV-2 variants: scoping review and applications on Belgian COVID-19 data. *BMC Infectious Diseases*. 2022;22(1):839.
46. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package “dagitty.” *Int J Epidemiol*. 2016;45(6):1887–1894.
47. European Centre for Disease Prevention and Control (ECDC). Reinfection with SARS-CoV-2: implementation of a surveillance case definition within the EU/EEA. In: European Centre for Disease Prevention and Control [Internet]. 8 Apr 2021 [cited 6 Sep 2022]. Available: <https://www.ecdc.europa.eu/en/publications-data/reinfection-sars-cov-2-implementation-surveillance-case-definition-within-eueea>
48. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(3):1–67.
49. Lee S, Lee W. Application of Standardization for Causal Inference in Observational Studies: A Step-by-step Tutorial for Analysis Using R Software. *J Prev Med Public Health*. 2022;55(2):116–124.
50. Ripley B. Package ‘boot.’ 2022. Available: <https://cran.r-project.org/web/packages/boot/boot.pdf>
51. Bours MJL. Tutorial: A nontechnical explanation of the counterfactual definition of effect modification and interaction. *Journal of Clinical Epidemiology*. 2021;134: 113–124.
52. Buchan SA, Chung H, Brown KA, Austin PC, Fell DB, Gubbay JB, et al. Estimated Effectiveness of COVID-19 Vaccines Against Omicron or Delta Symptomatic Infection and Severe Outcomes. *JAMA Network Open*. 2022;5(9):e2232760.
53. Liu B, Gidding H, Stepien S, Cretikos M, Macartney K. Relative effectiveness of COVID-19 vaccination with 3 compared to 2 doses against SARS-CoV-2 B.1.1.529 (Omicron) among an Australian population with low prior rates of SARS-CoV-2 infection. *Vaccine*. 2022;40(43):6288–6294.
54. Kissling E, Hooiveld M, Martínez-Baz I, Mazagatos C, William N, Vilcu A-M, et al. Effectiveness of complete primary vaccination against COVID-19 at primary care and community level during predominant Delta circulation in Europe: multicentre analysis, I-MOVE-COVID-19 and ECDC networks, July to August 2021. *Eurosurveillance*. 2022;27(21):2101104.

55. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365–376.
56. Yahav D, Yelin D, Eckerle I, Eberhardt CS, Wang J, Cao B, et al. Definitions for coronavirus disease 2019 reinfection, relapse and PCR re-positivity. *Clin Microbiol Infect*. 2021;27(3):315–318.

8. APPENDICES

Appendix A – Target trial emulation protocol

Appendix B – Directed Acyclic Graph specification Dagitty

Appendix C – Sensitivity analysis – Target Trial I: Primary vaccination plus mRNA booster vaccination *versus* primary vaccination without booster vaccination

Appendix D – Main analysis – Target Trial II: Primary vaccination plus mRNA booster vaccination *versus* primary vaccination without booster vaccination

Appendix E – Sensitivity analysis – Target Trial II: Heterologous *versus* homologous prime-boost vaccination

Appendix F – Sensitivity analysis – Target Trial II: Heterologous *versus* homologous prime-boost vaccination

Appendix A – Target trial emulation protocol

Table A1. Protocol of the target randomized control trials and target trial emulation by using observational data

Protocol Component	Target Trial Specification	Target Trial Emulation
Eligibility criteria	- Individuals that acquired a SARS-CoV-2 infection (either community- or hospital-acquired, detected by an RT-PCR or AG laboratory test) more than 14 days after intervention (i.e., more than 14 days after receiving the last dose of their vaccination schedule) and COVID-19 disease progression of this infection requiring COVID-19 related hospital care Enrolment period: From the start of the Belgian booster vaccination campaign until the end of November 2022 Age: adults (age ≥ 18) SARS-CoV-2 variant of infection: Patients infected with the Omicron variant (WGS-confirmed)	- Patients treated for COVID-19 in a Belgian hospital and with an admission form reported in the Clinical Hospital Survey (CHS) were considered. Patients without a laboratory-confirmed (RT-PCR or AG) SARS-CoV-2 infection and who were not tested because they experienced COVID-19 symptoms (i.e., in a screening context) were excluded. Patients with a missing pseudonymized national registry number were excluded. Transferred or readmitted patients, and patients admitted to a psychiatric hospital or a hospital without an intensive care unit (ICU) are excluded. Patients with an unknown date of diagnosis, starting date of COVID-19 related hospital care or date of discharge are not included, as well as patients diagnosed with COVID-19 more than 20 days before hospital admission or after hospital discharge. Enrolment period: Patients hospitalized for COVID-19 in a Belgian hospital between 1 September 2021 (i.e., the start of the Belgian booster vaccination campaign) and 23 November 2022 (i.e., the date of data extraction minus two weeks to allow for sufficient follow-up time). Patients admitted outside this enrolment period are excluded. Age: adults (age ≥ 18) SARS-CoV-2 variant of infection: Patients infected with the Omicron sub-lineages BA.1, BA.2, BA.4 or BA.5 as identified by linkage with WGS-confirmed lineage data from the COVID-19 TestResult database in the main analysis and defined by the time period of diagnosis (10 January 2022 – 27 February 2022, 26 March 2022 – 19 May 2022 and 17 June 2022 – 11 November 2022 for BA.1, BA.2 and BA.4/5 respectively) in the sensitivity analysis. Patients without a linked WGS-confirmed lineage or with a linked WGS-confirmed lineage different from Omicron BA.1, BA.2, BA.4 or BA.5 were excluded from the study population in the main analysis. Patients with a COVID-19 diagnosis date outside these predefined periods were excluded from

		the study population in the sensitivity analysis.
Intervention strategies	Target trial 1: 1) Intervention with mRNA (2 doses of BNT162b2 or mRNA-1273) or viral vector (1 dose of Ad26.COV2.S or 2 doses of ChAdOx1-S) primary vaccination plus an mRNA (BNT162b2 or mRNA-1273) booster vaccination 2) Intervention with mRNA (2 doses of BNT162b2 or mRNA-1273) or viral vector (1 dose of Ad26.COV2.S or 2 doses of ChAdOx1-S) primary vaccination (no booster vaccination)	Same as in the target trial.
	Target trial 2: 1) Intervention with a heterologous prime-boost vaccination, defined as a viral vector (1 dose of Ad26.COV2.S or 2 doses of ChAdOx1-S) primary vaccination and mRNA (BNT162b2 or mRNA-1273) booster vaccination 2) Intervention with a homologous prime-boost vaccination, defined as an mRNA (2 doses of BNT162b2 or mRNA-1273) primary vaccination and mRNA (BNT162b2 or mRNA-1273) booster vaccination	Same as in the target trial.
Assignment procedures	Participants are randomly assigned to an intervention arm. The participants are not assigned blindly, thus will be aware of the intervention strategy to which they have been assigned.	Vaccination status ascertainment based on linkage with the Vaccinnet+ registry. Vaccination status was ascertained on the date of the COVID-19 diagnosis. For a vaccination dose to be considered for vaccination status ascertainment, it has to be administered at least 14 days before the date of COVID-19 diagnosis. Randomization is mimicked by conditioning on a set of factors that are assumed to confound the exposure-outcome pathway (as identified through construction of a Directed Acyclic Graph), resulting in comparable (exchangeable) intervention groups. We assume that individuals are exchangeable within levels of the following patient

		characteristics: patient's comorbidities, age, gender, pregnancy status, host genetics, socio-economic status (SES), previous infection(s), whether the patient is a nursing home resident, whether the patient is a healthcare worker, the hospital of admission and ICU load during COVID-19 related hospital stay, whether the patient had a hospital-acquired infection, treatment strategies used during the time of admission, and Omicron sub-lineage of infection. The adjustment is performed by adjusting for these covariates in the logarithmic regression model.
Follow-up period	- Follow-up starts at the start of the COVID-19 related hospital care - Follow-up ends when the patient is discharged from the hospital	- Follow-up starts at hospitalization due to SARS-CoV-2 infection. - Follow-up ends when the patient is discharged from the hospital
Outcome	Severe COVID-19, ICU transfer, in-hospital mortality	Severe COVID-19, ICU transfer, in-hospital mortality
Causal contrast of interest	Per-protocol effect	Observational analogue of the per-protocol effect

Appendix B – Directed Acyclic Graph specification Dagitty

```

dag {
  bb="0,0,1,1"
  "Clinical severity" [outcome,pos="0.682,0.728"]
  "Healthcare worker" [adjusted,pos="0.514,0.291"]
  "Host genetics" [latent,pos="0.521,0.630"]
  "ICU load" [adjusted,pos="0.840,0.507"]
  "Nursing home resident" [adjusted,pos="0.516,0.200"]
  "Omicron sub-lineage" [adjusted,pos="0.540,0.805"]
  "Place of residence" [pos="0.513,0.036"]
  "Previous infection" [adjusted,pos="0.147,0.738"]
  "SARS-CoV-2 vaccination scheme" [exposure,pos="0.399,0.735"]
  "Treatment strategies" [latent,pos="0.529,0.886"]
  Age [adjusted,pos="0.518,0.434"]
  Comorbidities [adjusted,pos="0.520,0.543"]
  Gender [adjusted,pos="0.626,0.453"]
  HAI [adjusted,pos="0.667,0.800"]
  Hospital [adjusted,pos="0.840,0.614"]
  Pregnancy [adjusted,pos="0.518,0.368"]
  SES [adjusted,pos="0.514,0.124"]
  Time [pos="0.338,0.803"]
  "Healthcare worker" -> "Clinical severity"
  "Healthcare worker" -> "Previous infection" [pos="0.323,0.335"]
  "Healthcare worker" -> "SARS-CoV-2 vaccination scheme"
  "Host genetics" -> "Clinical severity"
  "Host genetics" -> "Previous infection"
  "Host genetics" -> Comorbidities
  "ICU load" -> "Clinical severity"
  "Nursing home resident" -> "Clinical severity"
  "Nursing home resident" -> "Previous infection" [pos="0.265,0.323"]
  "Nursing home resident" -> "SARS-CoV-2 vaccination scheme"
  "Omicron sub-lineage" -> "Clinical severity"
  "Place of residence" -> "ICU load"
  "Place of residence" -> "Previous infection" [pos="0.130,0.186"]
  "Place of residence" -> "SARS-CoV-2 vaccination scheme"
  "Place of residence" -> Hospital
  "Place of residence" <-> SES
  "Previous infection" -> "Clinical severity" [pos="0.438,0.990"]
  "Previous infection" -> "SARS-CoV-2 vaccination scheme"
  "SARS-CoV-2 vaccination scheme" -> "Clinical severity"
  "Treatment strategies" -> "Clinical severity"
  Age -> "Clinical severity"
  Age -> "Nursing home resident" [pos="0.334,0.367"]
  Age -> "Previous infection" [pos="0.321,0.421"]
  Age -> "SARS-CoV-2 vaccination scheme"
  Age -> Comorbidities
  Age -> Pregnancy
  Comorbidities -> "Clinical severity"

```

```

Comorbidities -> "Previous infection" [pos="0.339,0.525"]
Comorbidities -> "SARS-CoV-2 vaccination scheme"
Gender -> "Clinical severity"
Gender -> "Healthcare worker"
Gender -> "Previous infection" [pos="0.322,0.487"]
Gender -> Age
Gender -> Comorbidities [pos="0.583,0.498"]
Gender -> Pregnancy
HAI -> "Omicron sub-lineage"
HAI -> Comorbidities
Hospital -> "Clinical severity"
Hospital -> "ICU load"
Pregnancy -> "Clinical severity"
Pregnancy -> "Previous infection" [pos="0.333,0.379"]
Pregnancy -> "SARS-CoV-2 vaccination scheme"
SES -> "Clinical severity"
SES -> "Previous infection" [pos="0.210,0.168"]
SES -> "SARS-CoV-2 vaccination scheme"
Time -> "ICU load"
Time -> "Omicron sub-lineage"
Time -> "Previous infection"
Time -> "SARS-CoV-2 vaccination scheme"
Time -> "Treatment strategies"
}

```

Appendix C – Sensitivity analysis – Target Trial I: Primary vaccination plus mRNA booster vaccination *versus* primary vaccination without booster vaccination

Table A2. Counterfactual risk of potential outcomes under different vaccination schedules (*R*, in %), risk difference (*RD*) and relative risk (*RR*) estimates, and 95% confidence interval (*CI*) within a multi-center study estimating the effect of the SARS-CoV-2 vaccination schedule (primary vaccination plus mRNA booster vaccination versus primary vaccination without booster vaccination) on severe clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality) of BA.1, BA.2 or BA.4/5 Omicron infection (identified by the time period of COVID-19 diagnosis) among hospitalized COVID-19 patients in Belgium.

	Primary vaccination, without booster vaccination (control group)	Primary vaccination, plus mRNA booster vaccination (intervention group)	Intervention effect	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Severe COVID-19	0.20 [0.17; 0.23]	0.15 [0.14; 0.16]	-0.05 [-0.08; -0.02]	0.75 [0.61; 0.88]
ICU admission	0.08 [0.06; 0.10]	0.07 [0.06; 0.07]	-0.01 [-0.03; 0.01]	0.87 [0.63; 1.11]
In-hospital mortality	0.15 [0.12; 0.18]	0.10 [0.09; 0.11]	-0.05 [-0.08; -0.02]	0.67 [0.51; 0.82]

Appendix D – Main analysis – Target Trial II: Primary vaccination plus mRNA booster vaccination *versus* primary vaccination without booster vaccination

Table A3. Differences between estimated risk differences (RDs) or relative risks (RRs) under exposure of different Omicron sub-lineages, and 95% confidence intervals (CI), for the different severe clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality).

	Additive scale			Multiplicative scale		
	RD _{BA.2} - RD _{BA.1}	RD _{BA.4/5} - RD _{BA.2}	RD _{BA.4/5} - RD _{BA.1}	RR _{BA.2} - RR _{BA.1}	RR _{BA.4/5} - RR _{BA.2}	RR _{BA.4/5} - RR _{BA.1}
Severe COVID-19	0.03 [-0.10; 0.16]	0.04 [-0.11; 0.19]	0.07 [-0.08; 0.22]	0.13 [-0.49; 0.75]	0.24 [-0.53; 1.00]	0.37 [-0.37; 1.10]
ICU admission	0.00 [-0.09; 0.08]	0.06 [-0.07; 0.19]	0.06 [-0.07; 0.18]	0.09 [-0.79; 0.97]	0.52 [-0.52; 1.57]	0.61 [-0.44; 1.66]
In-hospital mortality	0.01 [-0.10; 0.13]	0.05 [-0.07; 0.18]	0.06 [-0.07; 0.19]	0.08 [-0.84; 1.00]	0.52 [-2.53e+05; 2.53e+05]	0.60 [-2.53e+05; 2.53e+05]

Appendix E – Sensitivity analysis – Target Trial II: Heterologous versus homologous prime-boost vaccination

Table A4. Counterfactual risk of potential outcomes under intervention (*R*, in %), risk difference (*RD*) and relative risk (*RR*) estimates, and 95% confidence interval (*CI*) predicted with the Omicron sub-lineage as observed (overall, identified through linkage with WGS test results) and under exposure to Omicron sub-lineage BA.1, BA.2 and BA.4/5 separately, within a multi-center cohort study investigating the effect of the SARS-CoV-2 vaccination schedule (heterologous versus homologous prime-boost vaccination) on severe clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality) of BA.1, BA.2 or BA.4/5 Omicron infection among hospitalized COVID-19 patients in Belgium.

Outcome: Severe COVID-19				
	Homologous prime-boost scheme (control group)	Heterologous prime-boost scheme (intervention group)	Intervention regimen	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Overall	0.15 [0.14; 0.16]	0.16 [0.14; 0.17]	0.01 [-0.01 ; 0.03]	1.06 [0.91; 1.21]
Omicron BA.1	0.16 [0.14; 0.18]	0.18 [0.14; 0.21]	0.02 [-0.02; 0.05]	1.10 [0.85; 1.35]
Omicron BA.2	0.14 [0.12; 0.16]	0.15 [0.12; 0.19]	0.01 [-0.03; 0.05]	1.09 [0.81; 1.36]
Omicron BA.4/5	0.14 [0.12; 0.16]	0.14 [0.11; 0.17]	0.00 [-0.04; 0.03]	0.98 [0.72; 1.24]
Outcome: ICU admission				
	Homologous prime-boost scheme (control group)	Heterologous prime-boost scheme (intervention group)	Intervention regimen	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Overall	0.06 [0.06; 0.07]	0.06 [0.05; 0.08]	0.00 [-0.01; 0.02]	1.02 [0.78; 1.26]
Omicron BA.1	0.06 [0.04; 0.07]	0.06 [0.04; 0.09]	0.01 [-0.02; 0.03]	1.11 [0.69; 1.54]
Omicron BA.2	0.07 [0.05; 0.09]	0.06 [0.04; 0.08]	-0.01 [-0.04; 0.02]	0.88 [0.49; 1.27]
Omicron BA.4/5	0.07 [0.05; 0.08]	0.07 [0.04; 0.09]	0.00 [-0.02; 0.03]	1.05 [0.62; 1.47]
Outcome: In-hospital mortality				
	Homologous prime-boost scheme (control group)	Heterologous prime-boost scheme (intervention group)	Intervention regimen	
	R [95% CI]	R [95% CI]	RD [95% CI]	RR [95% CI]
Overall	0.10 [0.09; 0.11]	0.10 [0.09; 0.12]	0.00 [-0.02; 0.02]	1.01 [0.83; 1.19]

Omicron BA.1	0.12 [0.10; 0.14]	0.12 [0.10; 0.15]	0.01 [-0.03; 0.04]	1.06 [0.77; 1.35]
Omicron BA.2	0.09 [0.08; 0.11]	0.10 [0.07; 0.12]	0.00 [-0.03; 0.03]	1.02 [0.68; 1.35]
Omicron BA.4/5	0.10 [0.07; 0.12]	0.09 [0.06; 0.11]	0.00 [-0.04; 0.02]	0.93 [0.61; 1.25]

Appendix F – Sensitivity analysis – Target Trial II: Heterologous *versus* homologous prime-boost vaccination

Table A5. Differences between estimated risk differences (RDs) or relative risks (RRs) under exposure of different Omicron sub-lineages, and 95% confidence intervals (CI), for the different severe clinical outcomes (severe COVID-19, ICU admission, in-hospital mortality).

	Additive scale			Multiplicative scale		
	$RD_{BA.2} - RD_{BA.1}$	$RD_{BA.4/5} - RD_{BA.2}$	$RD_{BA.4/5} - RD_{BA.1}$	$RR_{BA.2} - RR_{BA.1}$	$RR_{BA.4/5} - RR_{BA.2}$	$RR_{BA.4/5} - RR_{BA.1}$
Severe COVID-19	0.00 [-0.06; 0.05]	-0.01 [-0.07; 0.04]	-0.02 [-0.07; 0.03]	-0.02 [-0.38; 0.35]	-0.10 [-0.48; 0.28]	-0.12 [-0.47; 0.24]
ICU admission	-0.01 [-0.05; 0.02]	0.01 [-0.03; 0.05]	0.00 [-0.04; 0.03]	-0.23 [-0.80; 0.34]	0.17 [-0.41; 0.75]	-0.07 [-0.67; 0.54]
In-hospital mortality	0.00 [-0.05; 0.04]	-0.01 [-0.05; 0.04]	-0.01 [-0.06; 0.03]	-0.02 [-0.46; 0.41]	-0.08 [-0.55; 0.39]	-0.10 [-0.53; 0.33]

SECTION III

**Integration of real-world data for public health
research across jurisdictional borders**

CHAPTER 3

Federated causal inference based on real-world observational data sources

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*, for the BeYond-COVID project. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. BMC Medical Research Methodology. 2023;23:248. [†]Shared first author. *Shared last author.

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. To fill this gap and allow 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. INTRODUCTION

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” [1]. 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 [1,2]. 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 study’s external validity [3]. Performing an RCT 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 [1]. 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) [1,2,4]. 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) [1,5]. 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 et al. (2022) [6] 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 study's internal validity [1,7,8].

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 [9], 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 [10,11]. González-García et al. (2021) [12] 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 [13,14], the Personal Health Train (PHT) [15,16], and VANTAGE6 [17], have been proposed. However, to our knowledge, guidance on the full methodological process to approach causal inference, including the specification of data requirements and guaranteeing interoperability when 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 [18]. 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 [11], Open Science (i.e., transparent and accessible processes and knowledge) [19,20] 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) [12,21] and Population Health Information Research Infrastructure (PHIRI) project [22], 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.

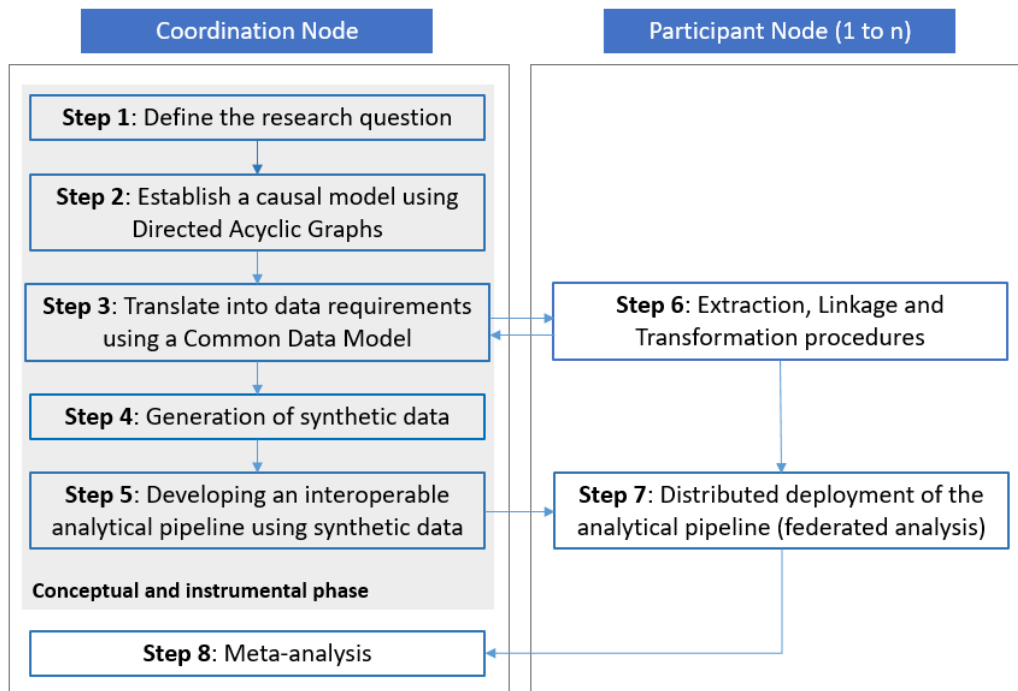


Figure 1. Visual representation of the proposed methodological framework.

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.

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 [23–25]. 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)

Exchangeable exposure groups should be pursued to estimate the causal effect of interest for the defined exposure-outcome relationship, thereby emulating randomization of the exposure as in an RCT [26]. Approaching exchangeability requires the adjustment for all known factors confounding the relationship between exposure and outcome. To identify 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) [1]. 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 [27–29]. 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 [30]. Guidance on the construction of DAGs and the identification of a minimal sufficient adjustment set can be found elsewhere [28,29,31]. 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) [32].

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 labelling must be consistent and follow a pre-specified convention (e.g., Snake Case, Camel Case) [33,34]. 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 within the model description of the CDM in a detailed manner, including (1) the model entity, (2) the variable label, (2) a description of the variable, (3) the encoding system, (4) the variable format and type, (5) the units of measurement, (6) the requirement level, (7) the variable-level validation rules, (8) the variable properties (observed/calculated), and (9) 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 and 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. A model description per entity should be provided if different entities (e.g., person, area, test, vaccination dose) are needed to cover the requirements captured in the DAG. In addition, a variable capturing information on a particular (co)morbidity might demand the specification of crosswalks (i.e., mapping to different classification systems). This ensures the coverage of the definition by different Participant Nodes who may use 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 in developing 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 [1,35], controlling for selection bias [36]) and handle missing data [37–39]. 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, certain elements common to any federated study should be contained within the pipeline, irrespective of these specific analytical methods. The first step in the analytical pipeline consists of a comprehensive Data Quality Assessment (DQA), including information on the completeness, uniqueness, and integrity [40]. 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 [41,42]), 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 [12,43]. 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) [15,16], DataSHIELD [13,14], and VANTAGE6 [17] 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 Coordination Node should pool the aggregated non-sensitive statistics produced as local outputs of the analytical pipeline 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 initial phase of the methodological framework, preparing for the subsequent implementation of the proposed methodology to respond to a policy-relevant research question. Only open-source software, such as DAGitty [30], R and DuckDB [44], was used to develop the workflow.

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.

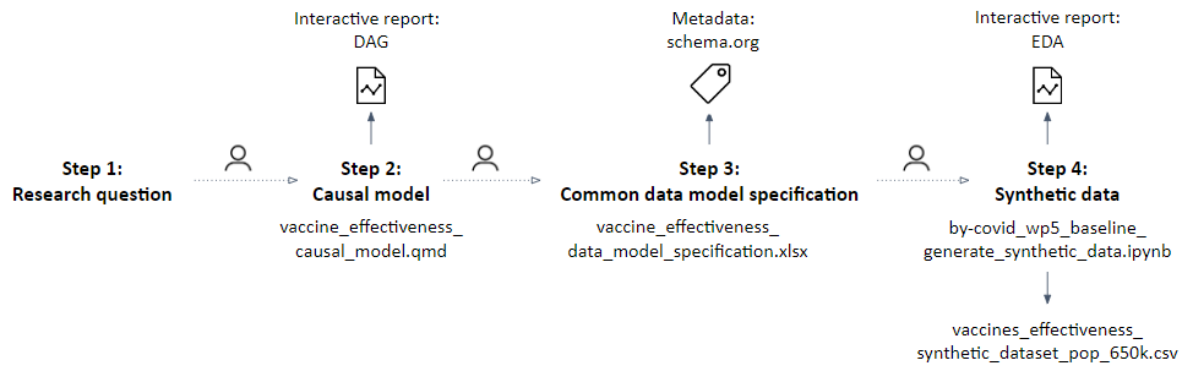


Figure 2. Overview of the executed steps and produced research objects during the implementation of the proposed methodological approach, step 1 to 4, preparing for the assessment of the real-world effectiveness of a primary vaccination schedule as compared to partial or no vaccination in preventing 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 [45], to evaluate 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 [30] (*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 [46]), 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 the study's secondary objectives 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 consists 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 [46]) 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 variables' required encoding and 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.

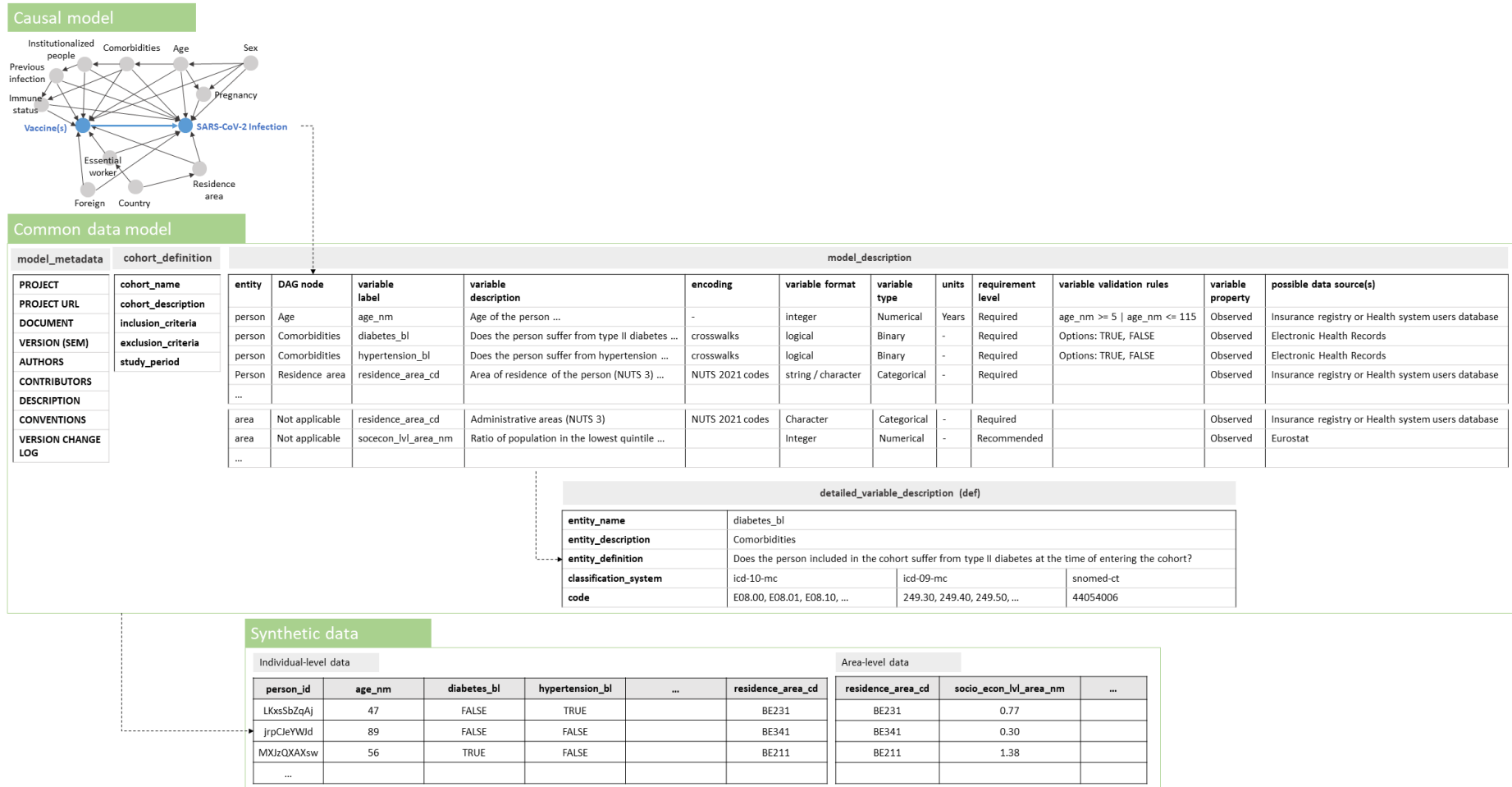


Figure 3. The causal model (using a DAG), Common Data Model (CDM) and synthetic data, and how they relate to each other. The DAG, capturing assumptions on factors and relationships when assessing the real-world effectiveness of a primary vaccination schedule as compared to partial or no vaccination in 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.

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 [46]) 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 [46]). Within this Python script we made use of the Python package ‘Faker’ [47]. 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 as available in the Zenodo publication [46]). Based on the EDA, we observe that the generated synthetic data correctly captures 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: thus, (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, and (6) a survival analysis in the matched study population (see the GitHub repository for methodological details [48]). A graphical overview of the analytical pipeline is presented in Figure 4. Each module of the analysis produces an interactive report, including documentation allowing one to trace back to decisions made along the way and interpret 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 [49,50], 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 [48]. 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.

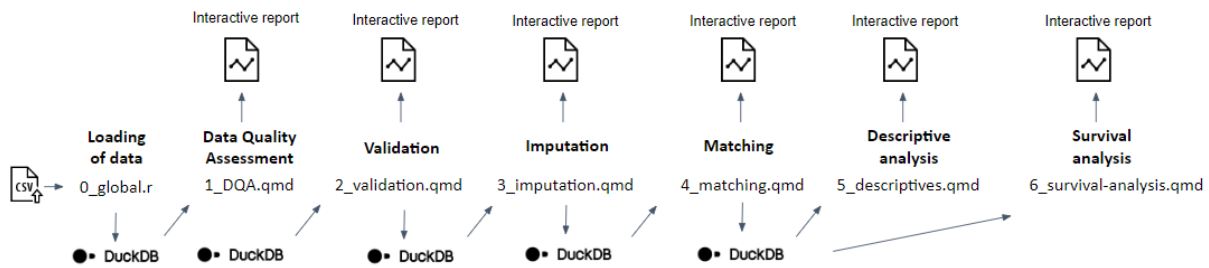


Figure 4. Graphical overview of the developed analytical pipeline, consisting of different subsequent modules, each producing an interactive report. Implementation of step 5 of the proposed methodological approach to assess the real-world effectiveness of a primary vaccination schedule as compared to 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 [46]. Further, the latter plus additional digital research objects produced along the way (i.e., a study protocol [45], a data management plan (DMP) [51], and an interoperable analytical pipeline) were collected together in a GitHub Repository [48]. A Research Object Crate (RO-Crate) [52] 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 [53]. 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 by 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. Different challenges and limitations of the proposed methodology were identified by going through these steps. 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 et al. (2022) [6], 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 [54]. Theoretical identification of confounding paths and other biases was based on the researchers' assumptions 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 residual confounding bias in the obtained causal effect estimates, hampering both the study's internal and external validity [55–57]. 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 [58], 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 [59] 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 granted in many European countries 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 [60]. For the development and publishing of a DMP, use of Argos' services [61,62] 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 [59], 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 [63,64]. A diversity of theoretical backgrounds of researchers in the federated network was observed, resulting in the need to build 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, the Coordination Node gave a theoretical overview on DAGs 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 [65], 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 allow 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, checking the distributed input data against a set of logical validation rules is recommended, examining compliance with 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 [66,67], parallelizing heavy and

repetitive computations where this increases throughput, and working with data management solutions, such as DuckDB [44].

4.3 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 [53]. In public health, this can be essential in responding to emerging public health threats (e.g., a pandemic). However, processes might not always be well documented, researchers might not know 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 [52] 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 using standards and controlled vocabularies, and Reusable by providing clear documentation (e.g., a README) [53].

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 interoperability layers. 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 implementing 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. DECLARATIONS

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 [39] 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.

Acknowledgements: The described work was conducted within the context of the European BeYond-COVID project, with all authors involved as participants within this project. The authors would like to acknowledge the participation and contribution of additional members of the BeYond-COVID project in the project's use case. Participating or contributing members not directly contributing to the manuscript are: Vasso Kalaitzi (KNAW/DANS), Claudia Habl (GÖG), Gunter Maier (GÖG), Mirjam Knol (RIVM), Chantal Reusken (RIVM), Mariken Tijhuis (RIVM), Leon Schutte (RIVM), Markus Perola (THL), Teemu Paajanen (THL), Kati Kristiansson (THL), Pekka Jousilahti (THL), Jostein Starrfelt (NIPH), Hinta Meijerink (NIPH).

Consortia: For the BeYond-COVID project.

7. REFERENCES

1. Hernán MA, Robins JM. *Causal Inference: What If*. 1st edition. Boca Raton: Chapman & Hall/CRC; 2020.
2. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov*. 2009;6:4.
3. Listl S, Jürges H, Watt RG. Causal inference from observational data. *Community Dentistry and Oral Epidemiology*. 2016;44(5):409–415.
4. Pearce N, Lawlor DA. Causal inference—so much more than statistics. *International Journal of Epidemiology*. 2016;45(6):1895–1903.
5. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. *Epidemiology*. 2004;15(5):615–625.
6. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*. 2022;328(24):2446–2447.
7. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*. 2006;60:578–586.
8. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. *Annual Review of Public Health*. 2013;34:61–75.
9. International Vaccine Access Center (IVAC). VIEW-hub. 13 Feb 2023 [cited 13 Feb 2023]. Available: <https://view-hub.org/covid-19/effectiveness-studies>
10. Directorate-General for Informatics (European Commission). New European interoperability framework: promoting seamless services and data flows for European public administrations. Publications Office of the European Union; 2017. Available: <https://data.europa.eu/doi/10.2799/78681>
11. Croatian Institute of Public Health (CIPH), Instituto Aragonés de Ciencias de la Salud (IACS). LOST* and found: Report on interoperability landscape in Europe. pp. 1–55. Available: <https://www.inf-act.eu/sites/inf-act.eu/files/2020-11/D10.1.pdf>
12. González-García J, Estupiñán-Romero F, Tellería-Orriols C, González-Galindo J, Palmieri L, Faragalli A, et al. Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct. *Archives of Public Health*. 2021;79:221.
13. Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*. 2010;39(5):1372–1382.
14. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*. 2014;43(6):1929–1944.

15. Attema T, Worm D. Technological breakthrough finally, a privacy-friendly way to harness data. 2021 [cited 5 Sep 2023]. Available: <http://resolver.tudelft.nl/uuid:8002b966-7bba-427c-b343-56326c1a587b>
16. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. *Data Intelligence*. 2020;2(1-2):96–107.
17. Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. VANTAGE6: an open source priVAcY preserviNg federaTed leArninG infrastruCTurE for Secure Insight eXchange. *AMIA Annu Symp Proc*. 2021;2020:870–877.
18. BY-COVID. [cited 30 Mar 2023]. Available: <https://by-covid.org/>
19. Spellman BA, Gilbert EA, Corker KS. *Open Science. Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. John Wiley & Sons, Ltd; 2018. pp. 1–47.
20. Foster ED, Deardorff A. Open Science Framework (OSF). *J Med Libr Assoc*. 2017;105(2): 203–206.
21. Abboud LA, Bogaert P, Fehr A, Urbanski D, Tolonen H, Noguer-Zambran I, et al. The new Joint Action on Health Information: information for action (InfAct)! *European Journal of Public Health*. 2018;28(suppl_4): cky213.651.
22. Bogaert P, Schutte N. Towards a Population Health Information Research Infrastructure. *European Journal of Public Health*. 2021;31(Supplement_3):ckab164.572.
23. Nishikawa-Pacher A. Research Questions with PICO: A Universal Mnemonic. *Publications*. 2022;10(3):21.
24. Riva JJ, Malik KMP, Burnie SJ, Endicott AR, Busse JW. What is your research question? An introduction to the PICOT format for clinicians. *J Can Chiropr Assoc*. 2012;56(3):167–171.
25. Lira RPC, Rocha EM. PICOT: Imprescriptible items in a clinical research question. *Arq Bras Oftalmol*. 2019;82(2):1.
26. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*. 2016;183(8):758–764.
27. Staplin N, Herrington WG, Judge PK, Reith CA, Haynes R, Landray MJ, et al. Use of Causal Diagrams to Inform the Design and Interpretation of Observational Studies: An Example from the Study of Heart and Renal Protection (SHARP). *CJASN*. 2017;12(3):546–552.
28. Suzuki E, Shinozaki T, Yamamoto E. Causal Diagrams: Pitfalls and Tips. *J Epidemiol*. 2020;30(4):153–162.
29. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620–632.
30. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package “dagitty.” *Int J Epidemiol*. 2016;45(6):1887–1894.
31. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*. 2022;142:264–267.

32. Kasza J, Wolfe R, Schuster T. Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *International Journal of Epidemiology*. 2017;46(4):1303–1311.
33. Using eye tracking to study variable naming conventions and their effect on code readability. [cited 8 Sep 2023]. Available: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1337810&dswid=1132>
34. DiLeo C. Naming Things. In: DiLeo C, editor. *Clean Ruby: A Guide to Crafting Better Code for Rubyists*. Berkeley, CA: Apress; 2019. pp. 9–32.
35. Li L, Kleinman K, Gillman MW. A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity. *J Dev Orig Health Dis*. 2014;5(6):435–447.
36. Bareinboim E, Pearl J. Controlling Selection Bias in Causal Inference. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. PMLR; 2012. pp. 100–108. Available: <https://proceedings.mlr.press/v22/bareinboim12.html>
37. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5): 402–406.
38. Haukoos JS, Newgard CD. Advanced Statistics: Missing Data in Clinical Research—Part 1: An Introduction and Conceptual Framework. *Academic Emergency Medicine*. 2007;14(7): 662–668.
39. Rubin DB, Little RJA. *Statistical Analysis with Missing Data*. John Wiley & Sons; 2019.
40. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*. 2015;14:2.
41. Bashari Rad B, Bhatti H, Ahmadi M. An Introduction to Docker and Analysis of its Performance. *IJCSNS International Journal of Computer Science and Network Security*. 2017;17(3):228-235.
42. Boettiger C. An introduction to Docker for reproducible research. *SIGOPS Oper Syst Rev*. 2015;49(1): 71–79.
43. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. *GigaScience*. 2016;5:30.
44. Raasveldt M, Mühleisen H. DuckDB: an Embeddable Analytical Database. *Proceedings of the 2019 International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1981–1984.
45. Meurisse M, Van Goethem N, Estupiñán-Romero F, González-Galindo J, Royo-Sierra S, Martínez-Lizaga N, et al. BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol. 2023 [cited 31 Jan 2023]. doi:10.5281/zenodo.7560731
46. Estupiñán-Romero F, Van Goethem N, Meurisse M, González-Galindo J, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Common Data Model Specification [Internet]. 2023 [cited 22 Feb 2023]. Available: doi:10.5281/zenodo.7572373
47. Faraglia D. Welcome to Faker’s documentation! — Faker 18.13.0 documentation [Internet]. [cited 12 Sep 2023]. Available: <https://faker.readthedocs.io/en/master/>

48. Meurisse M, Estupiñan-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E. BY-COVID_WP5_T5.2_baseline-use-case [Internet]. GitHub; 2024. Available: https://by-covid.github.io/BY-COVID_WP5_T5.2_baseline-use-case/
49. ydata-profiling [Internet]. YData; 2023. Available: <https://github.com/ydataai/ydata-profiling>
50. Welcome - YData Profiling [Internet]. [cited 13 Sep 2023]. Available: <https://docs.profiling.ydata.ai/4.5/>
51. Martínez-Lizaga N, Meurisse M, Estupiñan-Romero F, Goethem NV, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Data Management Plan [Internet]. 2023 [cited 2 May 2023]. Available: doi:10.5281/zenodo.7625783
52. Sefton P, Ó Carragáin E, Soiland-Reyes S, Corcho O, Garijo D, Palma R, et al. RO-Crate Metadata Specification 1.1.3 [Internet]. 2023 [cited 3 May 2023]. Available: doi:10.5281/zenodo.7867028
53. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
54. Lee PH, Burstyn I. Identification of confounder in epidemiologic data contaminated by measurement error in covariates. *BMC Med Res Methodol*. 2016;16:54.
55. Andrade C. Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation. *Indian J Psychol Med*. 2018;40(5): 498–499.
56. Findley MG, Kikuta K, Denly M. External Validity. *Annual Review of Political Science*. 2021;24: 365–393.
57. Grimes DA, Schulz KF. Bias and causal associations in observational research. *The Lancet*. 2002;359(9302):248–252.
58. Quarto [Internet]. [cited 14 Jun 2023]. Available: <https://quarto.org/>
59. Dube K, Gallagher T. Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use. In: Gibbons J, MacCaull W, editors. *Foundations of Health Information Engineering and Systems*. Berlin, Heidelberg: Springer; 2014. pp. 69–86.
60. Al-Jundi A, Sakka S. Protocol Writing in Clinical Research. *J Clin Diagn Res*. 2016;10(11): ZE10–13.
61. OpenAIRE [Internet]. Argos. [cited 9 May 2023]. Available: <https://argos.openaire.eu/splash/index.html>
62. Papadopoulou E. ARGOS: plan and follow your data [Internet]. 2021. Available: https://www.um.edu.mt/library/oar/bitstream/123456789/70269/1/ARGOS_plan_and_follow_your_data_2021.pdf
63. Margariti V, Stamati T, Anagnostopoulos D, Nikolaidou M, Papastilianou A. A holistic model for assessing organizational interoperability in public administration. *Government Information Quarterly*. 2022;39(2):101712.
64. Weichhart G. Learning for Sustainable Organisational Interoperability. *IFAC Proceedings Volumes*. 2014;47(3):4280–4285.

65. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. *Health Technol.* 2022;12(2): 255–272.
66. Gillespie C, Lovelace R. *Efficient R Programming: A Practical Guide to Smarter Programming*. First Edition. O'Reilly Media, Inc.; 2016.
67. Lutz M. *Learning Python: Powerful Object-Oriented Programming*. Fifth Edition. O'Reilly Media, Inc.; 2013.

8. APPENDICES

Appendix A – Customizable template for building a Common Data Model

Appendix B – List of participating or contributing members of the BeYond-COVID project

Appendix A – Customizable template for building a Common Data Model

Excel file was generated, representing a customizable template for building a Common Data Model.

You can download the Excel file from the following publication:

Meurisse M, Estupiñán-Romero F, González-Galindo J, Martínez-Lizaga N, Royo-Sierra S, Saldner S, et al. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. *BMC Med Res Methodol.* 2023;23:248. <https://doi.org/10.1186/s12874-023-02068-3>

Appendix B – List of participating or contributing members of the BeYond-COVID project

The described work was conducted within the context of the European BeYond-COVID project, with all authors involved as participants within this project. The authors would like to acknowledge the participation and contribution of additional members of the BeYond-COVID project in the project's use case.

Participating or contributing members are: Nina Van Goethem¹, Enrique Bernal-Delgado³, Francisco Estupiñán-Romero³, Marjan Meurisse^{1,2}, Natalia Martínez-Lizaga³, Santiago Royo-Sierra³, Javier González-Galindo³, Simon Saldner⁴, Lorenz Dolanski-Aghamanoukjan⁵, Alexander Degelsegger-Marquez⁵, Stian Soiland-Reyes^{6,7}, Vasso Kalaitzi⁴, Claudia Habl⁸, Gunter Maier⁹ (GÖG), Mirjam Knol¹⁰, Chantal Reusken¹⁰, Mariken Tijhuis¹⁰, Leon Schutte¹⁰, Markus Perola¹¹, Teemu Paaanen¹¹, Kati Kristiansson¹¹, Pekka Jousilahti¹¹, Jostein Starrfelt¹², Hinta Meijerink¹².

¹ Department of Epidemiology and Public Health, Sciensano, Brussels, Belgium

² IREC – EPID, Université Catholique de Louvain, Brussels, Belgium

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

⁴ Data Archiving and Networked Services, Royal Netherlands Academy of Arts & Sciences, Amsterdam, The Netherlands

⁵ International Affairs, Policy, Evaluation and Digitalisation, Gesundheit Österreich GmbH (GÖG), Vienna, Austria

⁶ Department of Computer Science, The University of Manchester, Manchester, UK

⁷ Informatics Institute, Universiteit van Amsterdam, Amsterdam, The Netherlands

⁸ International Affairs and Subsidiaries, Gesundheit Österreich GmbH (GÖG), Vienna, Austria

⁹ Planning and system development, Gesundheit Österreich GmbH (GÖG), Vienna, Austria

¹⁰ National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

¹¹ Finnish Institute for Health and Welfare (THL), Helsinki, Finland

¹² Norwegian Institute of Public Health (NIPH), Oslo, Norway

CHAPTER 4

Real-world comparative effectiveness of SARS-CoV-2 vaccination campaigns against SARS-CoV-2 infections

Estupiñán-Romero F[†], **Meurisse M[†]**, Perola M, Paajanen T, González-Galindo J, Van Goethem N*, Bernal-Delgado E*. Real-world comparative effectiveness of SARS-CoV-2 primary vaccination campaigns against SARS-CoV-2 infections: a federated observational study emulating a target trial in three nations. [†]Equal contribution *Equal contribution as senior authors. *In review at PLOS ONE.*

ABSTRACT

Background: To assess the impact of large-scale SARS-CoV-2 vaccination campaigns in uncontrolled settings across regions, we performed a reproducible cross-border comparison of the real-world effectiveness of primary vaccination in preventing SARS-CoV-2 infections across three sites: Aragon (Spain), Brussels and Wallonia (Belgium), and Finland.

Methods: This observational study emulated a target trial by daily matching primary vaccinated individuals 1:1 to un- or partially vaccinated individuals based on a set of confounders from January to September 2021. Matched individuals were followed up until a SARS-CoV-2 infection was contracted or a censoring event occurred. Vaccine effectiveness in preventing infections was estimated by the difference in Restricted Mean Survival Time (RMST).

Findings: Primary vaccination extended the mean time until contracting a SARS-CoV-2 infection on average by 35.9 (95% CI [34.9 - 37.0]), 59.6 (95% CI [59.3 - 60.0]), and 1.6 (95% CI [1.1 - 2.0]) days over 365 days in the population cohort of Aragon (Spain), Brussels and Wallonia (Belgium), and Finland, respectively.

Interpretation: This federated population-based observational study showed the effectiveness of the SARS-CoV-2 primary vaccination campaign in prolonging the mean time to infection in the Aragon (Spain) and Brussels & Wallonia (Belgium) population cohorts. Only a minor difference over this time frame was found in Finland's population cohort.

Funding: Horizon Europe-funded European BeYond-COVID project (101046203)

1. INTRODUCTION

Since its identification in December 2019 in Wuhan, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), causing coronavirus disease 2019 (COVID-19), has resulted in a global public health crisis. Less than one year after the official identification of the virus, the first SARS-CoV-2 vaccine, i.e., the BNT162b2 vaccine (Comirnaty®, Pfizer–BioNTech), was scientifically evaluated and authorized in the European Union (EU) by the European Medicines Agency (EMA), shortly followed by the authorization of the mRNA-1273 (Spikevax®, Moderna), ChAdOx1-S (Vaxzevria®, Oxford–AstraZeneca), and Ad26.COV2.S (COVID-19 vaccine Janssen®, Johnson & Johnson) vaccines [1]. At the end of 2020, following the authorization of the first vaccines, large-scale SARS-CoV-2 vaccination campaigns were initiated across Europe, resulting in high coverage of primary vaccination across the EU [2]. For example, in Spain [49] and Finland [50], the first dose was administered on the 27th of December 2020, and Belgium started its vaccination campaign of December 28th 2020 [51].

Post-authorization vaccine effectiveness (VE) studies have been conducted in real-world settings in many European countries and showed that having received one or two vaccine doses protects against (symptomatic) infection [3–7]. Comparative analysis of these single-center studies is challenging given that they generally focus on a subpopulation with specific characteristics (e.g., within one country or region), use different methodologies or exposure/outcome measures, or consider different confounding factors during the analysis. More recently, a Vaccine Effectiveness, Burden and Impact Studies (VEBIS) study by Fontán-Vela et al. (2024) [8] did offer a cross-country assessment of real-world VE in preventing COVID-19 hospitalization following a common protocol.

A sensible approach to inform policy-making based on causal estimates inferred from real-world observational data is to design the observational study to emulate a target trial, which represents the hypothetical randomized clinical trial that could have been conducted to answer the research question, as a way to limit avoidable biases and improve quality of the assessment [9]. The current study extends existing research by combinationally (1) emulating a target trial, (2) using virtually all the target population (allowing us to evaluate the vaccination campaign), and (3) performing a cross-border comparison using a reproducible methodology. In this work, conducted as part of the Horizon Europe-funded BeYond-COVID (BY-COVID) project [10], we performed a cross-border comparison of the real-world effectiveness of primary vaccination campaigns in preventing SARS-CoV-2 infections by leveraging data from existing COVID-19 registries from Aragon (Spain), Brussels and Wallonia (Belgium) and Finland, and adopting a federated analysis methodology [11].

2. METHODS

2.1 Study design

The current observational study, covering a period of 20 months (from the start of the primary vaccination campaigns, January 1, 2021, up to one year after the start of the booster campaign, September 1, 2022), was designed to emulate a target trial assessing the efficacy of SARS-CoV-2 primary vaccination as compared to no or partial vaccination in preventing SARS-CoV-2 infection. The methodological framework for approaching federated causal inference by reusing real-world observational data has been described elsewhere [11], and the specification and subsequent emulation of the target trial for the current research question are explicitly described in Appendix A and B. The corresponding causal model constructed using a Directed Acyclic Graph (DAG) is presented in Appendix C, and the subsequent translation into data requirements in the form of a Common Data Model (CDM) can be consulted elsewhere [12]. The analysis code and execution instructions of the reproducible analysis pipeline, including a daily matching algorithm and survival analysis, are provided within a GitHub repository [13] (see documentation provided within the repository for more information [14]). The study was conducted in three sites, namely Aragon (Spain), Brussels and Wallonia (Belgium), and Finland, engaging three ‘*Participant Nodes*’ (Instituto Aragonés de Ciencias de la Salud or IACS, the Belgian Institute for Health or Sciensano, and the Finnish Institute for Health and Welfare or THL) as research institutions being able acquire access to the required individual-level data and participating in the BY-COVID project.

2.2 Study participants

The study population was composed of all individuals eligible to be vaccinated (5 to 115 years old), that were vaccinated with at least one dose of the SARS-CoV-2 vaccine (any of the available brands). Information was additionally extracted for all individuals eligible to be vaccinated who experiences a confirmed SARS-CoV-2 infection (irrespective of the type of test used for confirmation) during the data extraction period (i.e., starting from the date of the first documented SARS-CoV-2 infection until the most recent date at which data is available at the time of analysis) registered within Aragon (Spain), Brussels and Wallonia (Belgium), or Finland. Individuals eligible to be vaccinated, but with neither a registered vaccine dose nor a positive COVID-19 test during the data extraction period, could not be included. Extracting data from the date of the first documented SARS-CoV-2 infection in a site onwards allowed us to identify and exclude individuals who experienced a documented confirmed infection before completing the primary vaccination schedule or before January 1, 2021, for those not having completed a primary vaccination schedule. In addition, individuals non-compliant with predefined

validation rules (see documentation provided within the GitHub repository [14]), based on the specified eligibility criteria (see Appendix B) and data requirements detailed in the CDM [12], were excluded from the study population. The data extraction periods in the respective sites are presented in Appendix G.

2.3 Data sources

We applied the eligibility criteria to the data by linking the individual-level COVID-19 case data and SARS-CoV-2 vaccine registry data within the jurisdictional borders of the individual Participant Nodes based on unique pseudonymized individual-level identifiers, to build site-specific population cohorts. Patient administrative information (e.g., data from insurance registries or health system user databases), information on patient comorbidities (e.g., data from Electronic Health Records or EHR), and mortality data were additionally linked to allow compliance with the data requirements as defined in the CDM [12]. Pseudonymized individual-level data were accessed by authorized researchers of IACS, Sciensano and THL on respectively April 11, 2023, March 26, 2024, and April 19, 2024, in their respective secured processing environment and following their own governance rules and regulatory restrictions. More details on the leveraged data sources and the linkage and transformation procedures within the different sites can be found elsewhere (i.e., the Data Management Plan or DMP [15]). The planned methodology was previously described in a study protocol [16], facilitating data access applications by the individual Participant Nodes.

2.4 Exposure groups and assignment

Within each site, conditional exchangeability was targeted by iteratively (i.e., daily) matching exposed to unexposed individuals. On each day of the enrollment period (i.e., between January 1, 2021 and September 1, 2021), every eligible individual that completed a primary vaccination schedule on that day (defined as two doses of BNT162b2, mRNA-1273, or ChAdOx1-S or one dose of Ad26.COV2.S; individual in the *exposed group*) was matched to an individual who had not (yet) completed the primary vaccination schedule (either not received any vaccine dose or partially vaccinated, i.e., one dose of BNT162b2, mRNA-1273, or ChAdOx1-S; individual in the *unexposed group*). A ‘minimal sufficient adjustment set’ (i.e., the minimal set of covariates to condition on to block non-causal backdoor paths and approach a causal association) was identified based on the constructed causal model (see Appendix C) and by applying the ‘backdoor criterion’ [17]: age, gender, residence area, country, foreign, essential worker, institutionalized, individual-level socio-economic status (SES), presence of comorbidities, immune status and pregnancy. A detailed covariate description can be found in the CDM [12]. As missing values in variables considered for matching (i.e., those identified in

the ‘minimal sufficient adjustment set’) hinder matching, missing values were imputed (one fold, 50 iterations to reach algorithm convergence) with predictive mean matching using the ‘mice’ R package [18], listwise deletion was performed and/or variables were excluded from the list of variables to perform matching on, depending on characteristics of the data (according to a prespecified decision algorithm, see Appendix D). Hence, matching was executed based on this set of covariates, contingent on the availability of sufficient information, i.e., excluding covariates with high or complete missingness in the local analysis: *individual-level SES* for Aragon (Spain), *foreign* and *pregnancy* for Brussels and Wallonia (Belgium), and *institutionalized*, *foreign* and *pregnancy* for Finland. Matching involved exact matching whenever possible (i.e., whenever exact matches could be found) or was based on a distance-based algorithm (i.e., greedy nearest neighbor propensity-score matching, with propensity scores estimated using a generalized linear model) using the ‘MatchIt’ R package [19] (see Appendix E). The covariate balance between matched exposure groups was assessed by evaluating Standardized Mean Differences (SMD).

2.5 Follow-up period and outcome of interest

The outcome is defined as experiencing a documented SARS-CoV-2 infection. Matched individuals are followed up until either diagnosis of a SARS-CoV-2 infection (i.e., the outcome of interest), completion of a primary vaccination schedule (for individuals in the *unexposed group*), completion primary vaccination of the matched unexposed comparator (for individuals in the *exposed group*), administration of a booster dose (for individuals in the *exposed group*), administration of a booster dose of the matched primary vaccinated individual (for individuals in the *unexposed group*), death, death of matched individual, or the end of the study period (i.e., September 1, 2022). The timeline of the sequentially emulated target trial is displayed in Appendix F.

2.6 Statistical analyses

To understand who was participating in the study and explore the study’s internal validity, baseline characteristics of the matched study participants were reported by exposure group, with categorical variables presented both as count numbers and proportions. A time-to-event analysis was performed for each site, and cumulative incidence (CInc) curves were generated ($CInc(t) = 1 - S(t)$, with $S(t)$ the survival probability at time t , calculated using the Kaplan-Meier estimate). Differences in the mean time to the occurrence of a SARS-CoV-2 infection (i.e., Restricted Mean Survival Time Difference, RMSTD) between the exposed and unexposed group (as an estimate of the average treatment effect or ATE of interest) were calculated over one year (365 days), using the ‘survRM2’ R package [20]. The Average Treatment Effect (ATE) reflects the expected effect of an intervention (e.g., a vaccination campaign) if applied to everyone in the population, not to specific individuals. The ATE does not

capture how an individual will respond to the treatment, as individual effects may vary widely. Finally, RMST is suggested as an alternative measure in survival analyses when the assumption of Proportional Hazards does not hold [21], which we considered was the case here due to the potential waning of vaccine-induced immunity over time. Time-to-event outcomes were collected from each site to allow comparative analysis.

2.7 Sensitivity analyses

To explore the robustness of the obtained results to the study assumptions and the impact of possible unmeasured confounding, E-values were calculated using the R package ‘EValue’ [22]. An E-value quantifies the minimum strength of unmeasured confounding, required to entirely negate an estimated exposure-outcome relationship, given the covariates that have been accounted for [23]. In the context of the exposure-outcome relationship and controlled confounding, a low E-value suggests low robustness of the results, with only little unmeasured confounding required to negate the observed effect estimate.

Differences in SES between the exposure groups were accounted for by matching on the residence area and, in addition, on individual-level SES for those Participant Nodes having the data available. A sensitivity analysis was conducted in Brussels and Wallonia (Belgium) to investigate the presence of potential residual confounding introduced when only controlling for area-level differences and not controlling for individual-level SES. We compared the collapsible (therefore assuming the absence of a nonlinearity/non-collapsibility effect [24]) unadjusted RMST Difference (RMSTD) estimate (i.e., obtained by matching on all earlier considered matching factors excluding individual-level SES) with the adjusted RMSTD estimate (i.e., obtained by matching on all earlier considered matching factors including individual-level SES).

2.8 Role of funding source

The study's funder did not participate in the study design, data gathering, analysis, interpretation of results, or manuscript drafting.

3. RESULTS

Executing the analysis pipeline within each site produced local outputs for each analytical step in the form of HTML reports accessible elsewhere [25].

3.1 Study Population Characteristics

Initially, 1,143,805, 3,948,815 and 5,308,428 individuals were included in the population cohorts within Aragon (Spain), Brussels and Wallonia (Belgium), and Finland, respectively. Characteristics of the population before matching can be found in Appendix H. After performing 1:1 matching, respectively, 755,055, 2,123,328 and 2,588,318 matched exposed–unexposed pairs were obtained. The selection procedures to obtain the matched population are presented in Figure 1.

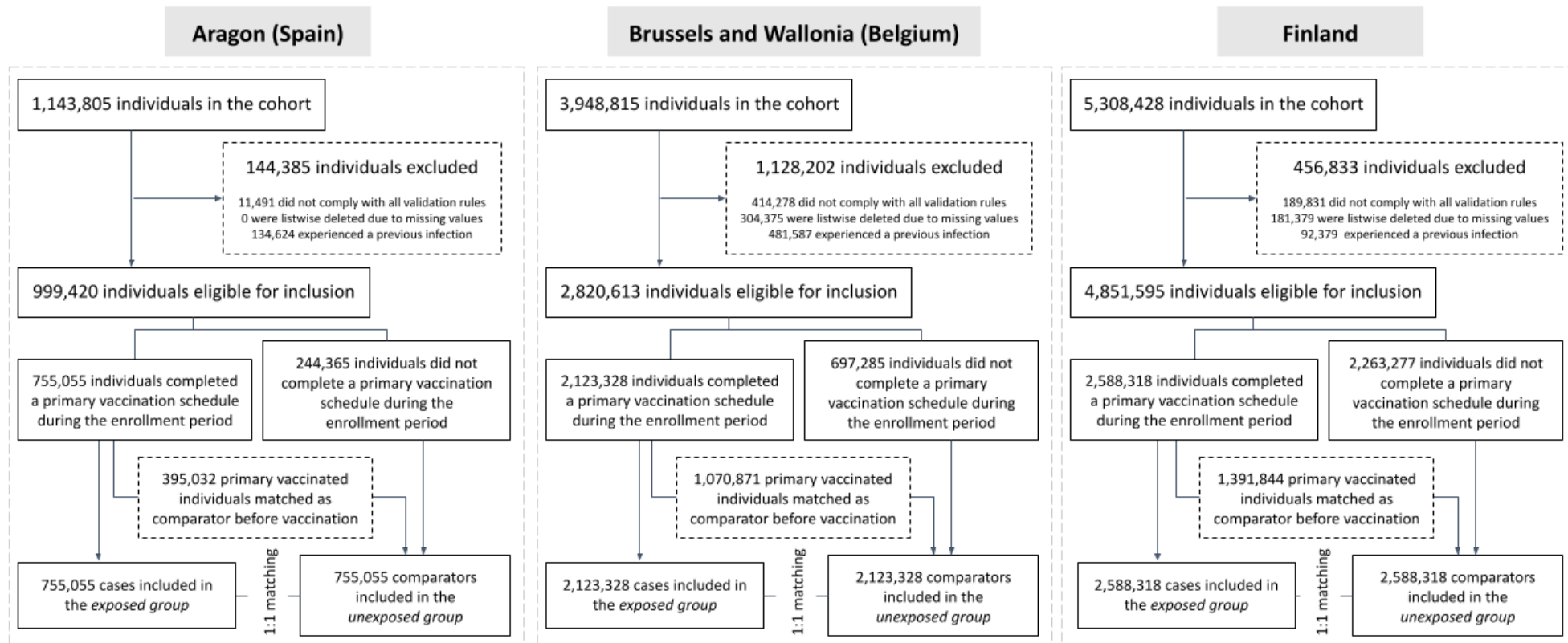


Figure 1. A flow diagram representing the selection process of obtaining the matched study population considered for time-to-event analysis at the different sites: Aragon (Spain), Brussels and Wallonia (Belgium), and Finland. Enrollment period: January 1, 2021, to September 1, 2021. Each day of the enrollment period, newly primary vaccinated individuals (exposed) were 1:1 matched to un- or partially vaccinated individuals (unexposed) based on a set of confounders.

Characteristics of the matched study population by exposure group and for the different population cohorts (sites) are presented in Table 1. Propensity score distributions in contrasted exposure groups before and after matching are visualized in Appendix I. SMDs before and after matching between the contrasted exposure groups for covariates considered for matching are visualized in Appendix J and presented in Appendix K. After matching was completed, baseline characteristics were found to be comparable between the groups.

Table 1. Characteristics of the matched study population by site (Aragon (Spain), Brussels and Wallonia (Belgium), and Finland) and by exposure group. Each day of the enrollment period (January 1, 2021, to September 1, 2021), newly primary vaccinated individuals (exposed) were 1:1 matched to un- or partially vaccinated individuals (unexposed) based on a set of confounders, resulting in the matched study population.

	Aragon (Spain)		Brussels and Wallonia (Belgium)		Finland	
	Exposed group (N = 755,055)	Unexposed group (N = 755,055)	Exposed group (N = 2,123,328)	Unexposed group (N = 2,123,328)	Exposed group (N = 2,588,318)	Unexposed group (N = 2,588,318)
Sex, n (%)						
1 (male)	360,616 (47.8%)	360,813 (47.8%)	1,019,290 (48.0%)	1,019,188 (48.0%)	1,209,685 (46.7%)	1,209,663 (46.7%)
2 (female)	394,439 (52.2%)	394,242 (52.2%)	1,104,038 (52.0%)	1,104,140 (52.0%)	1,378,633 (53.3%)	1,378,655 (53.3%)
Age (category), n (%)						
2 (5-9 yo)	0 (0%)	0 (0%)	35 (0.0%)	34 (0.0%)	0 (0.0%)	1 (0.0%)
3 (10-14 yo)	318 (0.0%)	318 (0.0%)	10,926 (0.5%)	10,895 (0.5%)	2,242 (0.1%)	2,242 (0.1%)
4 (15-19 yo)	16,314 (2.2%)	16,313 (2.2%)	81,026 (3.8%)	80,966 (3.8%)	37,872 (1.5%)	37,869 (1.5%)
5 (20-24 yo)	20,863 (2.8%)	20,857 (2.8%)	118,530 (5.6%)	118,468 (5.6%)	55,185 (2.1%)	55,181 (2.1%)
6 (25-29 yo)	24,610 (3.3%)	24,609 (3.3%)	120,852 (5.7%)	120,777 (5.7%)	79,327 (3.1%)	79,328 (3.1%)
7 (30-34 yo)	36,643 (4.9%)	36,628 (4.9%)	135,461 (6.4%)	135,357 (6.4%)	99,915 (3.9%)	99,910 (3.9%)

8 (35-39 yo)	50,711 (6.7%)	50,703 (6.7%)	141,127 (6.6%)	141,099 (6.6%)	140,313 (5.4%)	140,320 (5.4%)
9 (40-44 yo)	72,485 (9.6%)	72,461 (9.6%)	153,492 (7.2%)	153,417 (7.2%)	184,719 (7.1%)	184,717 (7.1%)
10 (45-49 yo)	79,881 (10.6%)	79,920 (10.6%)	172,450 (8.1%)	172,491 (8.1%)	195,696 (7.6%)	195,716 (7.6%)
11 (50-54 yo)	76,979 (10.2%)	77,022 (10.2%)	187,065 (8.8%)	187,339 (8.8%)	246,049 (9.5%)	246,075 (9.5%)
12 (55-59 yo)	74,495 (9.9%)	74,525 (9.9%)	195,884 (9.2%)	196,027 (9.2%)	270,896 (10.5%)	270,930 (10.5%)
13 (60-64 yo)	66,025 (8.7%)	66,019 (8.7%)	195,720 (9.2%)	195,937 (9.2%)	280,746 (10.8%)	280,777 (10.8%)
14 (65-69 yo)	5,885 (7.5%)	56,873 (7.5%)	181,337 (8.5%)	181,386 (8.5%)	296,567 (11.5%)	296,490 (11.5%)
15 (70-74 yo)	53,710 (7.1%)	53,708 (7.1%)	159,808 (7.5%)	159,781 (7.5%)	300,864 (11.6%)	300,835 (11.6%)
16 (75-79 yo)	45,802 (6.1%)	45,795 (6.1%)	111,346 (5.2%)	111,268 (5.2%)	179,299 (6.9%)	179,299 (6.9%)
17 (80-84 yo)	33,447 (4.4%)	33,428 (4.4%)	76,969 (3.6%)	76,862 (3.6%)	125,315 (4.8%)	125,315 (4.8%)
18 (>=85 yo)	45,887 (6.1%)	45,876 (6.1%)	81,300 (3.8%)	81,224 (3.8%)	93,313 (3.6%)	93,313 (3.6%)
Pregnancy, n (%)						
Yes	13,395 (1.8%)	13,337 (1.8%)	-	-	-	-
No	741,660 (98.2%)	741,718 (98.2%)	-	-	-	-
Essential worker, n (%)						
Yes	4,991 (0.7%)	4,958 (0.7%)	117,847 (5.6%)	117,817 (5.5%)	221,514 (8.6%)	221,465 (8.6%)
No	750,064 (99.3%)	750,097 (99.3%)	2,005,481 (94.4%)	2,005,511 (94.5%)	2,366,804 (91.4%)	2,366,853 (91.4%)
Institutionalized, n (%)						

Yes	6,661 (0.9%)	6,589 (0.9%)	17,562 (0.8%)	17,279 (0.8%)	-	-
No	748,394 (99.1%)	748,466 (99.1%)	2,105,766 (99.2%)	2,106,049 (99.2%)	-	-
Socio-economic level, n (%)						
Higher	-	-	620,238 (29.2%)	620,062 (29.2%)	376,578 (14.5%)	376,517 (14.5%)
Intermediate	-	-	643,550 (30.3%)	643,516 (30.3%)	570,719 (22.0%)	570,751 (22.1%)
Lower	-	-	859,540 (40.5%)	859,750 (40.5%)	1,641,021 (63.4%)	1,641,050 (63.4%)
Foreign, n (%)						
Yes	81,524 (10.8%)	81,495 (10.8%)	-	-	-	-
No	673,531 (89.2%)	673,560 (89.2%)	-	-	-	-
Comorbidities, n (%)						
Yes	326,172 (43.2%)	326,163 (43.2%)	549,014 (25.9%)	549,024 (25.9%)	1,109,074 (42.8%)	1,109,051 (42.8%)
No	428,883 (56.8%)	428,892 (56.8%)	1,574,314 (74.1%)	1,574,304 (74.1%)	1,479,244 (57.2%)	1,479,267 (57.2%)
Immune status, n (%)						
Yes	22,630 (3.0%)	22,621 (3.0%)	60,697 (2.9%)	60,477 (2.8%)	575,741 (22.2%)	575,687 (22.2%)
No	732,425 (97.0%)	732,434 (97.0%)	2,062,631 (97.1%)	2,062,851 (97.2%)	2,012,577 (77.8%)	2,012,631 (77.8%)

3.2 Vaccine effectiveness

At least one SARS-CoV-2 infection was documented during the study period for respectively 520,169 (out of 1,510,110; 34.4%), 1,129,404 (out of 4,246,656; 26.6%) and 818,831 (out of 5,176,636; 15.8%) individuals included in the matched study population in Aragon (Spain), Brussels and Wallonia

(Belgium), and Finland. Among individuals in the exposed group, 258,743 (out of 755,055; 34.3%), 465,876 (out of 2,123,328; 21.9%), and 413,184 (out of 2,588,318; 16.0%) contracted a documented SARS-CoV-2 infection after completing a SARS-CoV-2 primary vaccination schedule in the respective sites. On the other hand, among individuals in the unexposed group, 261,426 (out of 755,055; 34.6%), 663,528 (out of 2,123,328; 31.2%), and 405,647 (out of 2,588,318; 15.7%) contracted an infection during the study period in the respective sites.

The cumulative incidence of documented SARS-CoV-2 infections for contrasted exposure groups obtained from the different sites is presented in Figure 2. Cumulative incidence curves of contrasted groups diverged for time-to-event analysis in the population cohort of Aragon (Spain) and Brussels and Wallonia (Belgium), accumulating more SARS-CoV-2 cases in the unexposed group compared to the exposed group, while the compared cumulative incidence curves for analysis in the population cohort of Finland did not diverge considerably over a one year time period.

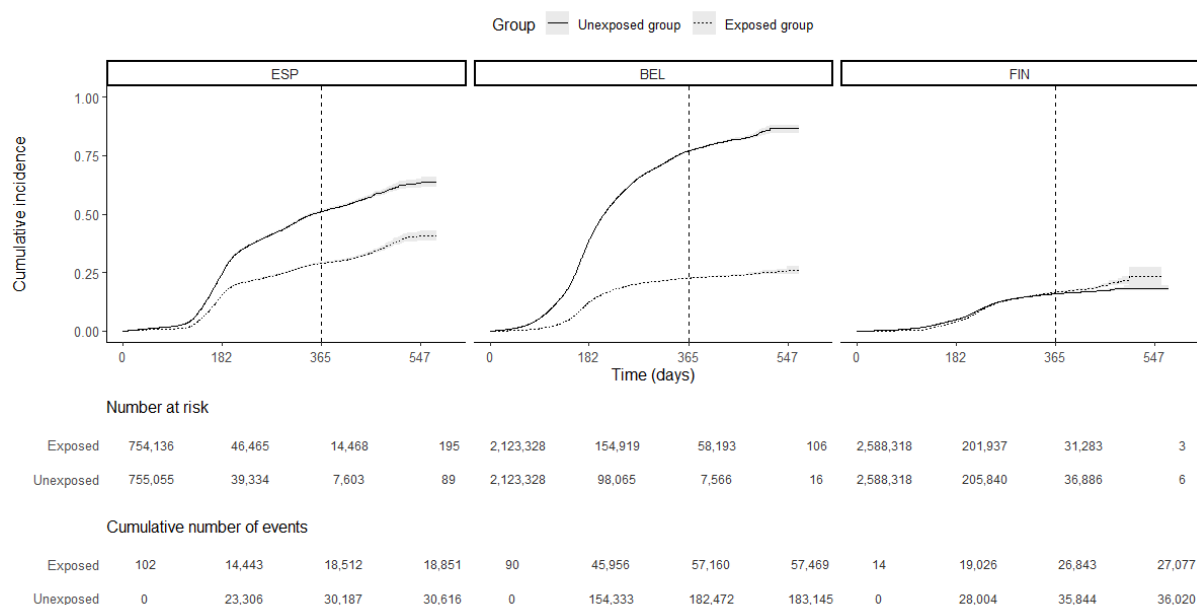


Figure 2. Cumulative incidence (Cinc) curves of documented SARS-CoV-2 infections for both exposure groups (Unexposed group versus Exposed group), obtained from local analysis in the different sites: Aragon, Spain (ESP); Brussels and Wallonia, Belgium (BEL); Finland (FIN). $Cinc(t) = 1 - S(t)$, with $S(t)$ the survival probability at time t . Each day of the enrollment period (January 1, 2021, to September 1, 2021), newly primary vaccinated individuals (exposed) were 1:1 matched to un- or partially vaccinated individuals (unexposed) based on a set of confounders, resulting in the matched study population. Study period: January 1, 2021 to September 1, 2022. A vertical dotted line at 365 days indicates the time up to which the Restricted Mean Survival Time is estimated.

The RMST up to 365 days for the exposed and unexposed group and corresponding RMSTD in the different sites are presented in Table 2. Completing a primary vaccination schedule extended the time

until contracting a (documented) SARS-CoV-2 infection on average with 35.9 (95% CI [34.9 - 37.0], p-value < 0.01), 59.6 (95% CI [59.3 - 60.0], p-value < 0.01) and 1.6 (95% CI [1.1 - 2.0], p-value < 0.01) days over a one year time period (365 days) as compared to not completing the schedule in Aragon (Spain), Brussels and Wallonia (Belgium), and Finland respectively.

Table 2. Restricted Mean Survival Times (RMSTs) at 365 days in the contrasted exposure groups (Unexposed group versus Exposed group) and Restricted Mean Survival Time Difference (RMSTD, RMST exposed group - RMST unexposed group) resulting from local analysis in the different sites: Aragon (Spain), Brussels and Wallonia (Belgium), and Finland. Each day of the enrollment period (January 1, 2021, to September 1, 2021), newly primary vaccinated individuals (exposed) were 1:1 matched to un- or partially vaccinated individuals (unexposed) based on a set of confounders, resulting in the matched study population. Study period: January 1, 2021 to September 1, 2022.

Measure	Aragon (Spain)	Brussels and Wallonia (Belgium)	Finland
RMST [95% CI], unexposed group (days)	280.758 [279.943; 281.573]	214.444 [214.144; 214.745]	340.580 [340.268; 340.893]
RMST [95% CI], exposed group (days)	316.696 [316.035; 317.357]	274.070 [273.862; 274.278]	342.145 [341.807; 342.483]
RMSTD [95% CI] (days)	35.938 [34.889; 36.988]	59.626 [59.260; 59.991]	1.565 [1.104; 2.025]
P-value RMSTD	<0.001	<0.001	<0.001

3.3 Sensitivity analyses

According to the E-values sensitivity analysis, negating the estimated effect (observed HRs: 0.532 [0.52; 0.544] in Aragon (Spain), 0.226 [0.223; 0.228] in Brussels and Wallonia (Belgium) and 0.768 [0.754; 0.782] in Finland) would require unmeasured confounders to be independently associated to the exposure (i.e., primary vaccination) and the outcome (i.e., contracting a SARS-CoV-2 infection) with a magnitude of 2.464 in Aragon (Spain), 4.879 in Brussels and Wallonia (Belgium) and 1.691 in Finland (HR scale).

The strength of confounding by individual-level SES was investigated by contrasting the unadjusted RMSTD estimate (obtained by matching on all earlier considered matching factors excluding individual-level SES; see Appendix I: RMSTD = 57.734 [57.378; 58.090]) and adjusted RMSTD estimate (obtained

by matching on all earlier considered matching factors including individual-level SES; see Table 2: RMSTD = 59.626 [59.260; 59.991]) obtained for the Brussels and Wallonia population cohort (Belgium).

4. DISCUSSION

Vaccination campaigns have had a pivotal role in the global response to the COVID-19 pandemic, and evaluating their effectiveness allows us to review past strategies and inform future health policies. In this federated population-based observational study, the effectiveness of SARS-CoV-2 primary vaccination in preventing SARS-CoV-2 infections was investigated across different cross-border sites (i.e., Aragon (Spain), Brussels and Wallonia (Belgium) and Finland), by leveraging existing real-world observational data within their premises. Results of this study suggest that the SARS-CoV-2 primary vaccination campaign was effective in prolonging the mean time to a SARS-CoV-2 infection in the population cohort of Aragon (35.9, 95% CI [34.9 - 37.0] days over a one year period) and Brussels and Wallonia (59.6, 95% CI [59.3 - 60.0] days over a one year period). Only a minor difference of 1.6 days [1.1 - 2.0] over this time frame was found in Finland's population cohort.

Local results: accuracy and potential biases

In populations within Aragon (Spain), Brussels & Wallonia (Belgium), and Finland, primary vaccination has been estimated to effectively prevent SARS-CoV-2 infections in the current but also prior studies. A population-based study conducted among residents of Aragon (Spain), spanning from December 2020 to May 2021, revealed that the BNT162b2, ChAdOx1-S, and mRNA-1273 vaccines were effective in preventing SARS-CoV-2 infection [5]. In a study conducted in Belgium by Braeye et al. (2021) [4] using test data from contact-tracing, completing a primary vaccination schedule with EMA-authorized vaccines (BNT162b2, ChAdOx1-S, mRNA-1273 and Ad26.CoV2.S) was found to be related to protection against infection. In Finland, SARS-CoV-2 VE estimates were derived for several specific target groups, including elderly people, those chronically ill, and healthcare workers [6,26]. Findings reported by Baum et al. (2021) [6] and Poukka et al. (2021) [26] suggest the effectiveness of primary vaccination in preventing infections among healthcare workers and populations of elderly and chronically ill in Finland.

The accuracy of these estimates may be influenced by various factors and data limitations that could introduce potential biases. Substantial imbalances in confounding factors across exposure groups can limit the comparability of the contrasted groups and introduce bias in the obtained effect estimates in observational studies. By matching exposed to unexposed individuals based on identified confounders in the current study, we aimed to obtain comparable exposure groups with similar distributions for

confounding factors. For all three considered sites, baseline characteristics before matching were observed to differ considerably between eligible individuals who completed a primary vaccination schedule during the enrollment period and those who did not (e.g., a higher proportion of individuals with comorbidities for the former compared to the latter; Appendix H), while after matching these characteristics were observed to be comparable between the contrasted groups (see Table 1). Further, SMDs between contrasted groups for considered covariates (Appendix J and K) showed a lack of balance ($\text{SMD} \geq 0.05$ or $\text{SMD} \leq -0.05$) for certain covariates (e.g., age, comorbidities) prior to matching and a good covariate balance (i.e., $-0.05 < \text{SMD} < 0.05$) post-matching. These results indicate that the local matching processes successfully produced matched study populations with considered baseline characteristics balanced between the exposure groups, hence substantiating these groups' comparability or conditional exchangeability. Matching was, however, limited to controlling for confounders that had been identified and were available within the SPEs of the Participant Nodes. Covariates with high or complete missingness were excluded as matching factors in the local analyses, and as such, residual confounding might persist. For example, information on individual-level SES was lacking for the Aragon (Spain) population cohort. SES was considered to be associated with both SARS-CoV-2 vaccine uptake and the risk of contracting SARS-CoV-2 based on previous studies [27–30] and, therefore, recognized as a potential confounder in the exposure-outcome relationship. It was hypothesized that adjusting for residence area (NUTS3 level) could achieve comparability in area-level socio-economic differences between the exposure groups. However, differences might remain at a smaller area or individual level. Therefore, additionally, we presume the need to adjust for lower-level socio-economic differences, approached here by matching on *individual-level SES*. To assess the potential impact of residual confounding by *individual-level SES*, the magnitude of confounding by this covariate was estimated for the Brussels and Wallonia population cohort (Belgium). The limited percentage change (3.2%) in the estimate (see Appendix L) that was obtained does not clearly indicate confounding by *individual-level SES* in this analysis, already controlling for area-level differences; however, extrapolation to other populations should be done with caution. For the population cohort of Brussels and Wallonia (Belgium), information on the covariates *foreign* and *pregnancy* was lacking, while for the population cohort of Finland, there was limited information on the covariates *institutionalized*, *foreign* and *pregnancy*. In addition, if relevant factors were not included in the causal model, this might additionally introduce unmeasured confounding. We computed the E-values for each regional model to evaluate potential unmeasured confounding, with E-values 2.46 and 4.88 obtained for Aragon (Spain) and Brussels and Wallonia (Belgium) models, respectively. Although a small E-value implies only little unmeasured confounding would be needed to explain away an effect estimate, both E-values are higher than the highest HR for any of the factors introduced in the survival

models, namely *essential worker* and *institutionalized* (HR 1.68 [1.62; 1.75] and 1.26 [1.20; 1.33] respectively) suggesting that any unmeasured confounder would need to have a stronger association than those, which seems unlikely. This suggests that unmeasured confounders are unlikely to challenge the assumption of causality for these models. The E-value for the model for the Finish population cohort was 1.691. This E-value, lower than certain confounders' associations introduced in the survival model (e.g., HR of 2.89 [2.86; 2.93] for the factor *age, category 5*, indicates the plausibility of unmeasured confounding to explain away the effect estimate. Moreover, the factor *institutionalized* was missing and not adjusted for in the Finnish population cohort, potentially introducing residual confounding given its strong association with the outcome in the other sites.

Further, due to reliance on the availability of SARS-CoV-2 test data, which is influenced by characteristics of the testing program and test-seeking behavior, bias could potentially be introduced in the obtained estimates. Testing directed or restricted to individuals with specific characteristics could result in misclassification of the outcome and misapplication of exclusion/inclusion criteria (e.g., exclusion of individuals with a previous SARS-CoV-2 infection) in our study. Although testing strategies were adjusted multiple times during the study period in the considered sites, testing mainly targeted individuals with a high probability of infection (i.e., symptomatic individuals suspected of having COVID-19) or was performed for screening purposes [31–34]. However, unlike testing strategies in Aragon (Spain) and Brussels and Wallonia (Belgium), Finland focused on testing individuals with symptoms indicative of COVID-19, overlooking asymptomatic high-risk contacts contacted through contact tracing. Seroprevalence studies conducted in the considered regions indicate that reported COVID-19 cases represent only a fraction of the actual circulation (approached through the prevalence of viral antibodies) during 2020 and 2021 [35–38]. As such, excluding individuals with a prior SARS-CoV-2 infection to eliminate the element of natural immunity in our study, solely based on registered PCR tests, might not be sufficient. Kuitunen et al. (2022) [39] conducted a retrospective observational study in the Southern Savonia region of Finland and compared COVID-19 testing rates between vaccinated and unvaccinated populations. Notably, they reported higher testing rates among vaccinated individuals, underscoring the potential bias introduced by testing frequency. Bias due to non-exhaustive or non-random testing might be present in all obtained VE estimates, with a need for cautious interpretation of the VE estimate obtained for the Finnish population cohort due to this limitation.

Future research should prioritize minimizing or eliminating potential biases while striving to understand genuine effect modification in comprehending real-world VE.

Comparative assessment

It is crucial to recognize that the comparability of findings across different sites might be limited due to variations in methodological approaches (e.g., considered confounding factors, the adopted approach for confounding adjustment), exposure definitions (e.g., the definition of ‘primary vaccination’, considered vaccination schedules), outcome definitions (e.g., SARS-CoV-2 infections, COVID-19 hospitalization, severe clinical outcomes), and/or target populations (e.g., the entire population of a region or country *versus* individuals with certain characteristics such as nursing home residents). In the current study, due to the adopted reproducible approach (e.g., specifying data requirements in a Common Data Model, distributing deployment of a reproducible analysis pipeline), the estimates obtained from the different sites were considered internally accurate and methodologically, semantically and analytically comparable across jurisdictions. Observed differences in VE across the geographically and jurisdictionally separated sites can be attributable to variations in contextual factors. A contributing factor could be the varying virus circulation in the local populations, leading to differences in exposure. Previous research has suggested that higher disease prevalence can enhance the effectiveness of vaccination in preventing infections [40]. During the study period, Aragon (Spain), and Brussels and Wallonia (Belgium) appeared to exhibit higher virus circulation compared to Finland, as indicated by incidence measures within the study population. This observation aligns with publicly available COVID-19 incidence rate statistics [41–44] and seroprevalence studies [35,37,38]. Differences in population density across the included countries or regions may have influenced the spread of the virus (e.g., through increased transmission opportunities) [45–47]. Notably, the sites with the highest observed VE estimates were those with the highest measured population density in 2021: 28.0 inhabitants/km² in Aragon, 3889.0 inhabitants/km² in Brussels and Wallonia, and 18.2 inhabitants/km² in Finland [48]. However, it is essential to recognize that not all individuals within the population undergo testing at any given time. As such, the observed incidence can only serve as a proxy for the real incidence, and testing strategies might importantly impact key indicators. Additionally, prevailing viral variants at the time of the study period might differ between the sites, with certain variants being more efficient in evading vaccine-induced immunity and, in this way, reducing the observed effectiveness of vaccines. Linking research data from viral gene sequencing collections to real-world data poses challenges due to limited coverage in existing research datasets and their lack of representativeness.

Strengths and limitations

A key strength of the study lies in acquiring population-wide estimates of VE across the participating local sites. These estimates were derived by working with large population cohorts and leveraging

detailed individual-level information sourced from high-quality, routinely collected population health registries. Linkage of multiple individual-level observational data sources allowed Participant Nodes to comply with the data requirements and to control for potential confounding bias. Rigorous daily matching was performed in the local analyses, resulting in exposure groups with balanced baseline characteristics across the contrasted groups and, in this way, also balancing contextual factors changing over time. Additionally, by following the framework for approaching federated causal inference described by Meurisse & Estupiñán-Romero et al., 2023 [11], we aimed to generate reproducible and comparable results from the different sites by the distributed deployment of reproducible statistical analyses.

While large population cohorts were obtained in the study, it is important to note that certain groups (e.g., marginalized communities, those with limited access to healthcare and therefore not reflected in administrative and clinical data sources) may not be adequately represented, which could influence the overall estimate and the generalizability of the findings. Another critical constraint of the study lies in the imperfect adherence to the specified data requirements by the different Participant Nodes. This is primarily attributed to the reliance on population health registries, which often contain missing or incomplete data. Consequently, the precision and reliability of the study's findings may be affected. Further, censoring individuals on the occurrence of a booster vaccination, resulted in the loss of information and a resulting reduced statistical power. Another limitation of our study is the lack of accounting for competing risks, particularly the risk of death, in our survival analysis, although matched individuals were censored upon one dying. While this could influence the interpretation of local results, we assume that due to the similar death rates across countries, the impact on our cross-country comparisons is limited. Finally, we need to acknowledge that assumptions on our exposure-outcome relationship as defined in the predefined causal model may not hold over time, and we highlight the need to continuously monitor trends in VE to assess the impact on population health and to guide and adapt policy with time.

The results of the study should be interpreted considering these limitations and the potential residual biases discussed earlier. The study highlights the challenges that are present when attempting to perform an accurate cross-border comparison of public health interventions, and indicates the potential for future improvement.

Implications

This study showcased a methodological framework for federated causal inference, which offers step-by-step guidance described within the BY-COVID project. Through this demonstration, we highlight its

feasibility while exposing several limitations. Linking different individual-level data registries to meet the data requirements can be complicated due to legal restrictions arising from privacy and data protection regulations (e.g., coming from the General Data Protection Regulation, GDPR) and the lack of a standardized procedure for requesting data access in Europe. Harmonizing procedures, addressing the differential interpretation of the GDPR, encouraging collaboration between stakeholders (e.g., data controllers, data processors and policy decision makers) and educating on the benefits of data sharing for research may advance policy-relevant research requiring linking sensitive individual-level population, health and care data. In the future, Standard Operating Procedures (SOPs) for data collection and management may play a crucial role in ensuring the quality, consistency, accuracy and timeliness of population, health and care data.

Assessing population-wide real-world SARS-CoV-2 VE is of great value, providing crucial insights for evidence-driven policies and fortifying our joint efforts in combating COVID-19 and other infectious diseases. It enables continuous monitoring of vaccine performance in real-world scenarios, extending beyond clinical trials, and facilitates the evaluation of public health interventions. We can shape impactful public health strategies and optimize cross-border vaccination campaigns by acquiring such evidence. Furthermore, acquiring accurate estimates plays a crucial role in effectively communicating vaccine benefits to the public and addressing concerns regarding vaccination.

5. CONCLUSION

Assessing the impact of vaccination campaigns across regions or countries is essential to evaluate past efforts and to shape future local and cross-border policy decisions. This federated population-based observational study indicated the effectiveness of the SARS-CoV-2 primary vaccination campaign in Aragon (Spain) and Brussels and Wallonia (Belgium) in preventing SARS-CoV-2 infections, while the observed effect was limited in Finland; however, questions remain about the causal nature of the estimated effects. The application of this federated methodological approach has demonstrated its effectiveness in assessing population-based policies that rely on sensitive population, health and care data.

6. DECLARATIONS

Ethics approval and consent to participate: Ethical approval was granted for linking different individual-level data sources to comply with the data requirements specified in the study in the sites Aragon (Spain) and Brussels and Wallonia (Belgium). In Aragon (Spain), the Research Ethics Committee

of Aragon (CEICA) granted approval for linking the required data sources (the COVID-19 public health surveillance and monitoring information system of COVID-19 cases, the SARS-CoV-2 vaccination population registry, the health system users databases, and data from electronic health records), via BIGAN (<https://bigan.iacs.es/en/home>), in its ruling of 22 February 2023 (reference number C.I. EPA 23/008). In Brussels and Wallonia (Belgium), activities of this study and thus linkage of the required data sources (the vaccine registry VACCINET+, the COVID-19 test result database, the COBRHA database, the Intermutualistic Agency (IMA) database, the STATBEL database) are covered by the existing deliberation of the Information Security Committee Social Security and Health (ISC-SSH) of the LINK-VACC project. The LINKVACC project was approved by the Medical Ethics Committee UZ Brussels–VUB on 3 February 2021 (reference number 2020/523), and authorization was granted from the ISC Social Security and Health (reference number IVC/KSZG/21/034). Individual data sources within the LINK-VACC project are kept in the pseudonymized environment of healthdata.be. Access rights to the pseudonymized data in the healthdata.be data warehouse are granted ad nominatum for scientists involved in the surveillance activities at Sciensano. After approval by the LINK-VACC governance board and the respective database owners, MM was granted access to the data warehouse in the context of the BY-COVID project, where linkage/transformation procedures and analyses were performed. THL's ethical approval has not been needed for processing register data for research purposes. The role of Finnish Institute for Health and Welfare (THL) in this work has been limited to acting as a national analyst. A THL researcher (TP) has processed Finnish register data under THL's internal data permit and only aggregated non-sensitive data has been transferred to other Parties in the Consortium.

Informed consent was waived based on Art. 6 and 9 of the GDPR. The collection is allowed based on general interest (art. 6 GDPR) and regarding article 9 § 2 of the GDPR: processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular, professional secrecy.

Data availability: 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 analysis pipeline, and local outputs) are available from Zenodo [12, 15, 16, 25] and GitHub, <https://w3id.org/ro/doi/10.5281/zenodo.6913045>.

The individual-level datasets analyzed during the current study do not fulfil the requirements for open data access. Data access to registry data from Aragon (through the BIGAN platform) is managed by the Aragon Institute of Health Sciences and requires prior approval of the research protocol by the Clinical

Research Ethics Committee of Aragon (CEICA). Data access to registry data from Brussels and Wallonia (through the healthdata.be platform) within the LINKVACC project is granted ad nominatum for the scientists involved in the surveillance activities at Sciensano. External investigators with a request for selected data should fill in the data request form (<https://epistat.sciensano.be/datarequest/>, accessed on 26 May 2024). Depending on the type of desired data (anonymous or pseudonymized), the provision of data will have to be assessed by the Belgian Information Security Committee Social Security & Health based on legal and ethical regulations and is outlined in a data transfer agreement with the data owner (Sciensano). All Finnish registry data is available through the register authorization office FinData (www.findata.fi).

Competing interests: Not applicable.

Funding: This research was supported by funds from the BY-COVID Project. 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.

Acknowledgements: We would like to acknowledge the following researchers who contributed to the development of the methodological framework demonstrated in this research: Natalia Martínez-Lizaga, Santiago Royo-Sierra, Simon Saldner, Lorenz Dolanski-Aghamanoukjan, Alexander Degelsegger-Marquez, Stian Soiland-Reyes.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the author(s) used Microsoft Copilot for rewording. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

7. REFERENCES

1. EMA. COVID-19 vaccines: authorised. In: European Medicines Agency [Internet]. 11 Jan 2021 [cited 13 Feb 2023]. Available: <https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines/vaccines-covid-19/covid-19-vaccines-authorised>
2. European Centre for Disease Prevention and Control. COVID-19 Vaccine Tracker [Internet]. [cited 17 Nov 2021]. Available: <https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html#uptake-tab>
3. Pritchard E, Matthews PC, Stoesser N, Eyre DW, Gethings O, Vihta K-D, et al. Impact of vaccination on new SARS-CoV-2 infections in the United Kingdom. *Nat Med*. 2021;27: 1370–1378.

4. Braeye T, Cornelissen L, Catteau L, Haarhuis F, Proesmans K, De Ridder K, et al. Vaccine effectiveness against infection and onwards transmission of COVID-19: Analysis of Belgian contact tracing data, January-June 2021. *Vaccine*. 2021;39(39):5456–5460.
5. Cura-Bilbao A del, López-Mendoza H, Chaure-Pardos A, Vergara-Ugarriza A, Guimbao-Bescós J. Effectiveness of 3 COVID-19 Vaccines in Preventing SARS-CoV-2 Infections, January–May 2021, Aragon, Spain. *Emerg Infect Dis*. 2022;28(3):591–598.
6. Baum U, Poukka E, Palmu AA, Salo H, Lehtonen TO, Leino T. Effectiveness of vaccination against SARS-CoV-2 infection and Covid-19 hospitalisation among Finnish elderly and chronically ill—An interim analysis of a nationwide cohort study. *PLOS ONE*. 2021;16(11):e0258704.
7. Fabiani M, Ramigni M, Gobetto V, Mateo-Urdiales A, Pezzotti P, Piovesan C. Effectiveness of the Comirnaty (BNT162b2, BioNTech/Pfizer) vaccine in preventing SARS-CoV-2 infection among healthcare workers, Treviso province, Veneto region, Italy, 27 December 2020 to 24 March 2021. *Euro Surveill*. 2021;26(17):2100420.
8. Fontán-Vela M, Kissling E, Nicolay N, Braeye T, Evercooren IV, Hansen CH, et al. Relative vaccine effectiveness against COVID-19 hospitalisation in persons aged ≥ 65 years: results from a VEBIS network, Europe, October 2021 to July 2023. *Eurosurveillance*. 2024;29(1):2300670.
9. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*. 2016;183(8):758–764.
10. Beyond COVID | BY-COVID Project | Fact Sheet | HORIZON | CORDIS | European Commission [Internet]. [cited 26 Apr 2023]. Available: <https://cordis.europa.eu/project/id/101046203>
11. Meurisse M, Estupiñán-Romero F, González-Galindo J, Martínez-Lizaga N, Royo-Sierra S, Saldner S, et al. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. *BMC Medical Research Methodology*. 2023;23: 248.
12. Estupiñán-Romero F, Van Goethem N, Meurisse M, González-Galindo J, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Common Data Model Specification [Internet]. 2023 [cited 22 Feb 2023].
13. Meurisse M, Estupiñán-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E. BY-COVID_WP5_T5.2_baseline-use-case [Internet]. GitHub; Available: https://by-covid.github.io/BY-COVID_WP5_T5.2_baseline-use-case/
14. Meurisse M, Estupiñán-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E. BY-COVID_WP5_T5.2_baseline-use-case/vaccine_effectiveness_analytical_pipeline/documentation at main · by-covid/BY-COVID_WP5_T5.2_baseline-use-case [Internet]. [cited 17 Jun 2024]. Available: https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case/tree/main/vaccine_effectiveness_analytical_pipeline/documentation
15. Martínez-Lizaga N, Meurisse M, Estupiñán-Romero F, Goethem NV, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Data Management Plan [Internet]. 2023 [cited 2 May 2023]. Available: doi:10.5281/zenodo.7625783
16. Meurisse M, Van Goethem N, Estupiñán-Romero F, González-Galindo J, Royo-Sierra S, Martínez-Lizaga N, et al. BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol [Internet]. 2023 [cited 31 Jan 2023]. Available: doi:10.5281/zenodo.7560731

17. Hernán MA, Robins JM. *Causal Inference: What If*. 1st edition. Boca Raton: Chapman & Hall/CRC; 2020.
18. Buuren S van, Groothuis-Oudshoorn K, Vink G, Schouten R, Robitzsch A, Rockenschaub P, et al. mice: Multivariate Imputation by Chained Equations. 2022. Available: <https://CRAN.R-project.org/package=mice>
19. Ho D, Imai K, King G, Stuart E, Whitworth A, Greifer N. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference [Internet]. 2023. Available: <https://cran.r-project.org/web/packages/MatchIt/index.html>
20. Uno H, Tian L, Horiguchi M, Cronin A, Battiou C, Bell J. survRM2: Comparing Restricted Mean Survival Time [Internet]. 2022. Available: <https://cran.r-project.org/web/packages/survRM2/index.html>
21. Han K, Jung I. Restricted Mean Survival Time for Survival Analysis: A Quick Guide for Clinical Researchers. *Korean J Radiol*. 2022;23(5):495–499.
22. Mathur MB, Smith LH, Ding P, VanderWeele TJ. EValue: Sensitivity Analyses for Unmeasured Confounding and Other Biases in Observational Studies and Meta-Analyses [Internet]. 2021. Available: <https://cran.r-project.org/web/packages/EValue/index.html>
23. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*. 2017;167(4):268–274.
24. Ni A, Lin Z, Lu B. Stratified Restricted Mean Survival Time Model for Marginal Causal Effect in Observational Survival Data. *Annals of epidemiology*. 2021;64:149-154.
25. Meurisse M, Estupiñán-Romero F, Van Goethem N, Perola M, Paaanen T, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Local outputs and comparative analysis [Internet]. 2024 [cited 11 Jun 2024]. Available: doi:10.5281/zenodo.11209362
26. Poukka E, Baum U, Palmu AA, Lehtonen TO, Salo H, Nohynek H, et al. Cohort study of Covid-19 vaccine effectiveness among healthcare workers in Finland, December 2020 - October 2021. *Vaccine*. 2022;40(5):701–705.
27. Cavillot L, Loenhout JAF van, Devleeschauwer B, Wyndham-Thomas C, Oyen HV, Ghattas J, et al. Sociodemographic and socioeconomic disparities in COVID-19 vaccine uptake in Belgium: a nationwide record linkage study. *J Epidemiol Community Health*. 2024;78(3):176–183.
28. Bastola K, Nohynek H, Lilja E, Castaneda AE, Austero S, Kuusio H, et al. Incidence of SARS-CoV-2 Infection and Factors Associated With Complete COVID-19 Vaccine Uptake Among Migrant Origin Persons in Finland. *Int J Public Health*. 2023;68:1605547.
29. Saarinen S, Moustgaard H, Remes H, Sallinen R, Martikainen P. Income differences in COVID-19 incidence and severity in Finland among people with foreign and native background: A population-based cohort study of individuals nested within households. *PLoS Med*. 2022;19:e1004038.
30. Fernández-Martínez NF, Ruiz-Montero R, Gómez-Barroso D, Rodríguez-Torronteras A, Lorusso N, Salcedo-Leal I, et al. Socioeconomic differences in COVID-19 infection, hospitalisation and mortality in urban areas in a region in the South of Europe. *BMC Public Health*. 2022;22:2316.

31. Meurisse M, Lajot A, Dupont Y, Lesenfants M, Klamer S, Rebolledo J, et al. One year of laboratory-based COVID-19 surveillance system in Belgium: main indicators and performance of the laboratories (March 2020–21). *Archives of Public Health*. 2021;79:188.
32. Ministry of Social Affairs and Health. National strategy for COVID-19 testing and tracing updated to target testing and tracing more accurately. In: Finnish Government [Internet]. 14 Sep 2021 [cited 24 Apr 2024]. Available: <https://valtioneuvosto.fi/en/-/1271139/national-strategy-for-covid-19-testing-and-tracing-updated-to-target-testing-and-tracing-more-accurately>
33. Ministry of Social Affairs and Health. Strategy for COVID-19 testing and contact tracing updated to prevent spread of disease particularly among risk groups. In: Ministry of Social Affairs and Health [Internet]. 2 Oct 2022 [cited 24 Apr 2024]. Available: <https://stm.fi/en/-/strategy-for-covid-19-testing-and-contact-tracing-updated-to-prevent-spread-of-disease-particularly-among-risk-groups>
34. COVID-19 Health System Response Monitor (HSRM). Monitoring and surveillance. 14 Apr 2022 [cited 24 Apr 2024]. Available: <https://eurohealthobservatory.who.int/monitors/hsrm/all-updates>
35. Pollán M, Pérez-Gómez B, Pastor-Barriuso R, Oteo J, Hernán MA, Pérez-Olmeda M, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *The Lancet*. 2020;396(10250):535–544.
36. Leclercq V, Van den Houte N, Gisle L, Roukaerts I, Barbezange C, Desombere I, et al. Prevalence of Anti-SARS-CoV-2 Antibodies and Potential Determinants among the Belgian Adult Population: Baseline Results of a Prospective Cohort Study. *Viruses*. 2022;14(5):920.
37. Herzog SA, De Bie J, Abrams S, Wouters I, Ekinci E, Patteet L, et al. Seroprevalence of IgG antibodies against SARS-CoV-2 - a serial prospective cross-sectional nationwide study of residual samples, Belgium, March to October 2020. *Euro Surveill*. 2022;27(9):2100419.
38. Solastie A, Nieminen T, Ekström N, Nohynek H, Lehtonen L, Palmu AA, et al. Changes in SARS-CoV-2 seroprevalence and population immunity in Finland, 2020–2022. *Emerg Microbes Infect*. 2023;12(2):2222849.
39. Kuitunen I, Uimonen M, Seppälä SJ, Ponkilainen VT. COVID-19 vaccination status and testing rates in Finland—A potential cause for bias in observational vaccine effectiveness analysis. *Influenza Other Respir Viruses*. 2022;16(5):842–845.
40. Lau CS, Aw TC. Considerations in Understanding Vaccine Effectiveness. *Vaccines (Basel)*. 2022;11(1):20.
41. Gobierno De Aragon. Datasets - Datos abiertos Aragón Open Data - Portal del Gobierno de Aragón. In: Datos abiertos Aragón Open Data [Internet]. [cited 12 Apr 2024]. Available: <https://opendata.aragon.es/datos/catalogo/dataset/publicaciones-y-anuncios-relacionados-con-el-coronavirus-en-aragon>
42. Gobierno De Aragon. Casos confirmados de COVID-19 Aragón [Internet]. [cited 12 Apr 2024]. Available: <https://datacovid.salud.aragon.es/covid/>
43. Sciensano. Belgium COVID-19 Dashboard - Sciensano. In: Belgium COVID-19 Epidemiological Situation [Internet]. [cited 12 Apr 2024]. Available: <http://lookerstudio.google.com/reporting/c14a5cfc-cab7-4812-848c-0369173148ab/page/tpRKB?feature=opengraph>

44. THL. COVID-19 cases in the infectious diseases registry [Internet]. [cited 12 Apr 2024]. Available: <https://www.thl.fi/episeuranta/tautitapaukset/coronamap.html>
45. Md Iderus NH, Lakha Singh SS, Mohd Ghazali S, Yoon Ling C, Cia Vei T, Md Zamri ASS, et al. Correlation between Population Density and COVID-19 Cases during the Third Wave in Malaysia: Effect of the Delta Variant. *Int J Environ Res Public Health*. 2022;19(12):7439.
46. Wong DWS, Li Y. Spreading of COVID-19: Density matters. *PLOS ONE*. 2020;15(12):e0242398.
47. Abd Rasid ASB, Dom NC, Abdullah S, Salim H. Population density as a factor in the spread of covid-19 cases in peninsular Malaysia. *Malaysian Journal of Medicine and Health Sciences*. 2021;17:7–13.
48. Statistics | Eurostat [Internet]. [cited 23 Apr 2024]. Available: https://ec.europa.eu/eurostat/databrowser/view/demo_r_d3dens/default/table

8. APPENDICES

Appendix A – Target trial specification

Appendix B – Target trial protocol and emulation using observational data

Appendix C – The constructed causal model, using a Directed Acyclic Graph

Appendix D – Decision algorithm to guide decisions on how to manage missing values

Appendix E – Matching algorithm

Appendix F – Timeline of the sequentially emulated target trial

Appendix G – Data extraction periods

Appendix H – Characteristics of the unmatched population eligible for matching

Appendix I – Propensity score distribution in contrasted exposure groups before and after matching in the considered sites

Appendix J – Graphical representation of Standardized Mean Differences (SMDs) before and after matching between contrasted exposure groups for each covariate considered for matching in the different sites

Appendix K – Standardized Mean Differences (SMDs) before and after matching between contrasted exposure groups for each covariate considered for matching, in the different sites.

Appendix L – Change-in-estimate analysis to estimate the magnitude of confounding by individual-level SES

Appendix A – Target trial specification

In the target trial, individuals eligible to be vaccinated (aged 5 to 115 - included) in Aragon (Spain), Brussels or Wallonia (Belgium), or Finland, excluding individuals that had experienced a SARS-CoV-2 infection before potential enrolment, would have been eligible for inclusion in the study population. Eligible individuals would have been enrolled in the study from the start of the SARS-CoV-2 vaccination campaign (January 1, 2021) until the beginning of the SARS-CoV-2 booster campaign (1 September, 2021), randomly assigning them to one of the exposure arms: (a) completing a primary vaccination schedule (defined as two doses of BNT162b2, mRNA-1273, or ChAdOx1-S or one dose of Ad26.COV2.S), *versus* (b) not completing a primary vaccination schedule (either not receiving any vaccine dose or being only partially vaccinated, i.e., one dose of BNT162b2, mRNA-1273, or ChAdOx1-S). Thus, the specified enrollment period covered nine months. Enrolled individuals would have been followed up from enrollment (i.e., at randomization to one of the vaccination strategies, baseline t_0) until either (1) the end of the follow-up period, 365 days after baseline, (2) the experience of a diagnosed SARS-CoV-2 infection (i.e., the outcome of interest), (3) death (all causes), (4) loss to follow-up, (5) the date of receiving a booster dose, whichever occurs first. Data could have been collected until September 1, 2022, as such allowing a potential follow-up of 365 days (12 months) for individuals enrolled (baseline t_0) at the end of the enrollment period. The timeline of the specified hypothetical target trial is displayed in Figure A1.

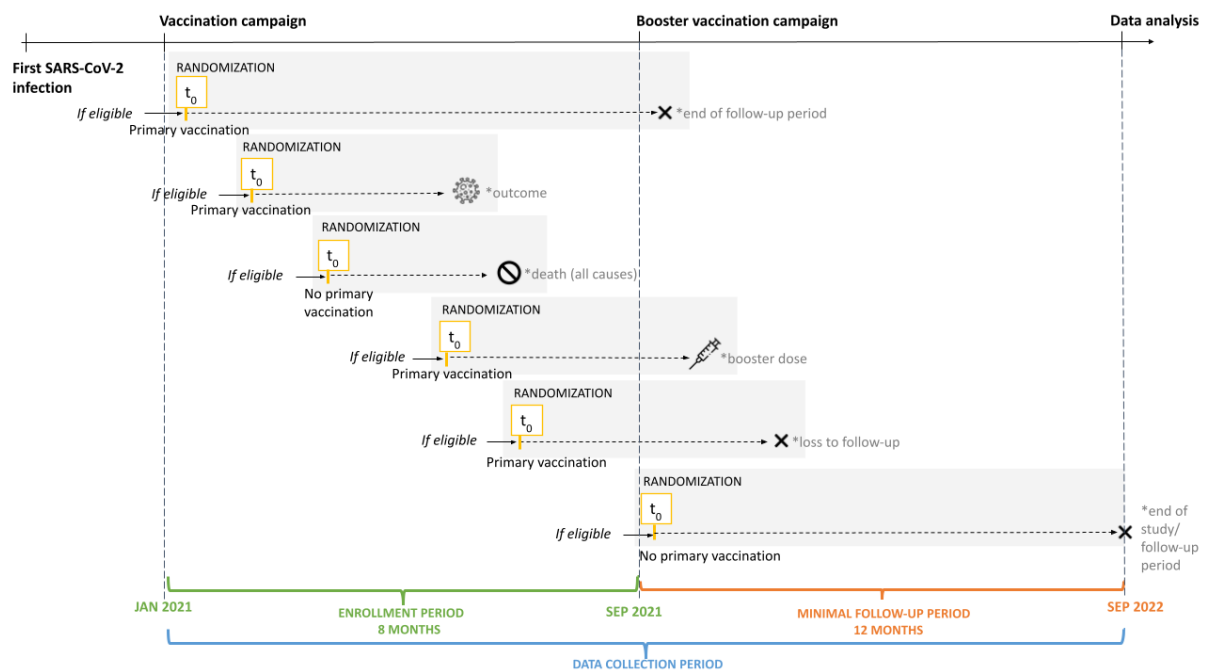


Figure A1. Timeline of the target trial. t_0 : baseline, at which individuals would have been randomised to one of the intervention arms.

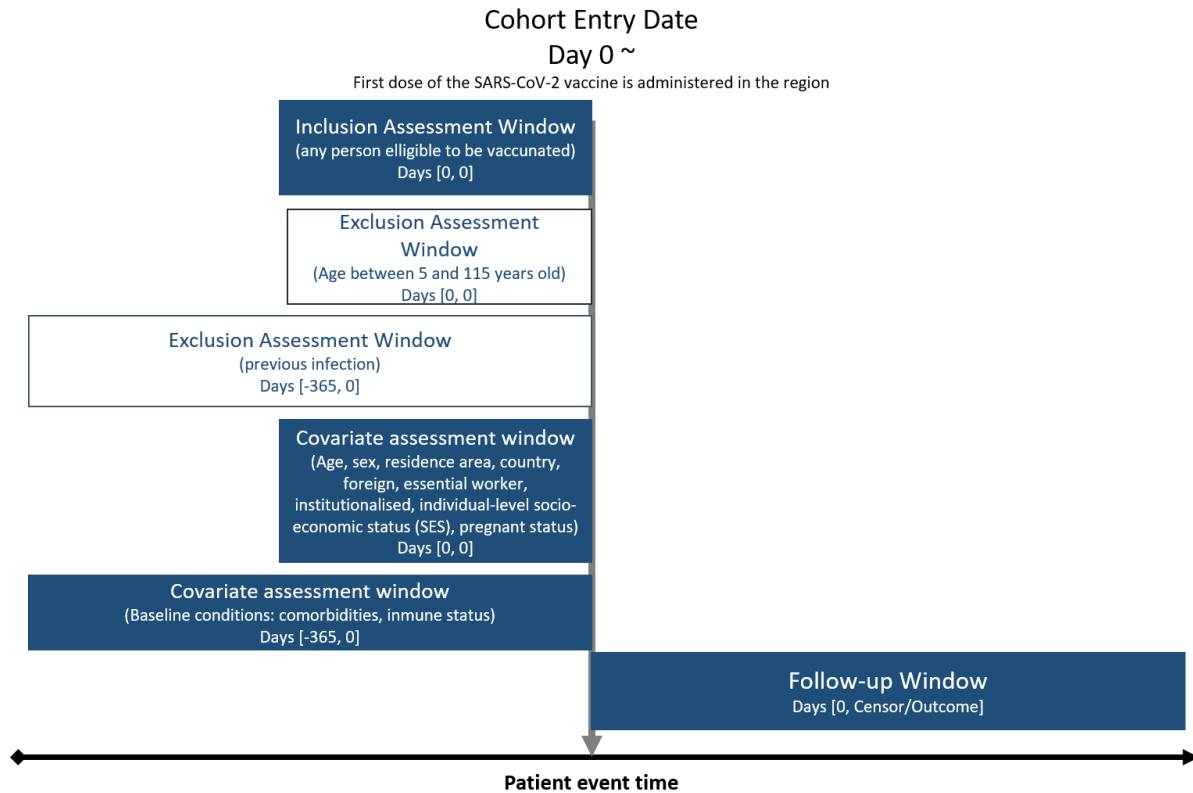


Figure A2. Study design diagram accounting for patient event time from the cohort entry date to the end of the follow-up window (i.e., first booster dose administration in the region).

Appendix B – Target trial protocol and emulation using observational data

Table A1. Protocol of the target randomized control trial and target trial emulation by using observational data.

Protocol Component	Target Trial Specification (<i>Ideal Hypothetical Trial</i>)	Target Trial Emulation (<i>Actual Study Design reusing Real World Data</i>)
Eligibility criteria	• Study population: All individuals eligible to be vaccinated. • Enrollment period: From the start of the SARS-CoV-2 vaccination campaign (January 1, 2021) to the start of the SARS-CoV-2 booster vaccination campaign (September 1, 2021). • Age: 5 to 115 years old, included (<i>all individuals eligible for vaccination</i>). • Previous infection: Individuals without a SARS-CoV-2 infection before enrolment. • Country: Individuals living in the respective country	• Study population: All individuals vaccinated with at least one dose of the SARS-CoV-2 vaccine (any of the available brands) and all individuals eligible to be vaccinated with a documented positive diagnosis (irrespective of the type of test) for SARS-CoV-2 infection during the data extraction period. • Enrollment period: From the start of the SARS-CoV-2 vaccination campaign (January 1, 2021) until the beginning of the SARS-CoV-2 booster vaccination campaign (September 1, 2021). • Age: 5 to 115 years old, included (<i>all individuals eligible for vaccination</i>). • Previous infection: Individuals without 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). • Country: Resident of the respective country We applied these eligibility criteria to the data obtained by fully joining the registered individuals in the COVID-19 cases and vaccination datasets. Therefore, individuals with neither a registered vaccine dose nor a positive diagnosis during the data extraction period were therefore excluded.
Treatment strategies	<u>Primary analysis</u> 1) Exposed group: Completing a primary vaccination schedule (defined as two doses of BNT162b2, mRNA-1273, or ChAdOx1-S or one dose of Ad26.COV2.S) 2) Unexposed group: Not completing a primary vaccination schedule (either	Same as in Target Trial Specification. Vaccination status ascertainment (including brand, dose and date) based on the vaccination registry. The date of being primary vaccinated (completing a primary vaccination schedule) was considered 14 days after receiving the final dose of the primary vaccination schedule.

	not receiving any vaccine dose or being only partially vaccinated, i.e., one dose of BNT162b2, mRNA-1273, or ChAdOx1-S). Pragmatic trial: strategies compared under the usual conditions.	
Assignment procedures	Participants will be randomly assigned, at baseline, to either complete the primary vaccination schedule (<i>exposed group</i>) or to not complete the primary vaccination schedule (<i>unexposed group</i>). The participants will be aware of the strategy to which they have been assigned (i.e., without blind assignment).	To mimic randomization, we needed to adjust for all confounding factors required to ensure comparability (conditional exchangeability) of the groups defined by vaccination (<i>as identified in the Directed Acyclic Graph</i> [1]). We assumed that individuals were randomly assigned within levels of baseline covariates: age, gender, residence area, country, foreign, essential worker, institutionalized, individual-level socio-economic status (SES), presence of comorbidities, immune status and pregnancy. A detailed covariate description could be found in the specified Common Data Model (CDM) [1]. The adjustment was performed via matching.
Follow-up period	- Starts, for each individual, at randomization to one of the vaccination strategies (baseline) - Ends at diagnosis of SARS-CoV-2 infection (i.e., outcome of interest), death, loss to follow-up, the date of receiving a booster dose, end of the follow-up period (365 days after baseline), whichever occurs first.	Sequential trial emulation: - Starts, for each individual: - Exposed group: time of completing a primary vaccination schedule - Unexposed group: first eligible time (time of completing a primary vaccination schedule of the matched pair) - Ends at diagnosis of SARS-CoV-2 infection (i.e., outcome of interest), death (all causes), death (all causes) of matched individual, completion of a primary vaccination schedule (for not/partially vaccinated individuals, <i>unexposed group</i>), completion primary vaccination of the matched control (for primary vaccinated individuals, <i>exposed group</i>), administration of a booster dose (for primary vaccinated individuals, <i>exposed group</i>), administration of a booster dose of the matched primary vaccinated individual (for not/partially vaccinated individuals, <i>unexposed group</i>), the end of the study period (i.e., September 1, 2022, one year

		after the end of the enrollment period). Loss to follow-up due to emigration or ‘disenrollment’ from the healthcare system was not taken into account. Newly vaccinated individuals (completing a primary vaccination schedule) were eligible for inclusion in the <i>exposed group</i> of the study, even if they had previously been selected in the <i>unexposed group</i> (as controls).
Outcome	Laboratory-confirmed SARS-CoV-2 infection	A documented laboratory-confirmed SARS-CoV-2 infection, excluding self-reported test results.
Causal contrast of interest	Per-protocol effect.	Observational analogue of the per-protocol effect.

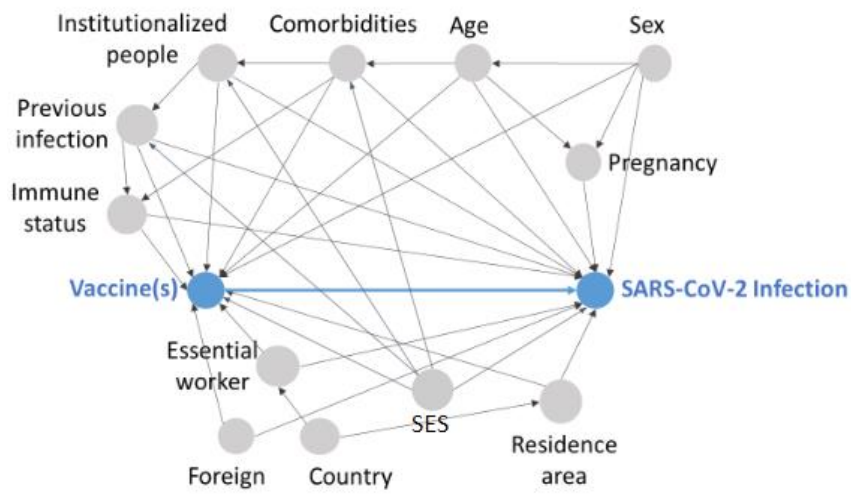
Appendix C – The constructed causal model, using a Directed Acyclic Graph

Figure A3. The defined causal model, constructed using a Directed Acyclic Graph (DAG), capturing assumptions on variables and their relationships. SES: Socio-Economic Status, measured at the individual level.

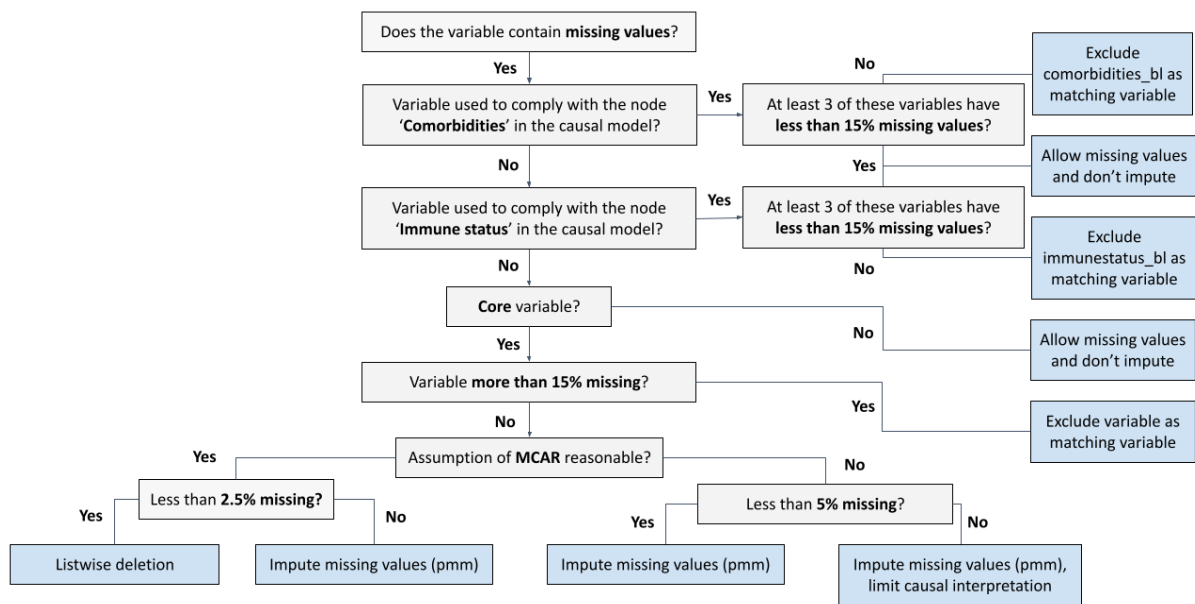
Appendix D – Decision algorithm to guide decisions on how to manage missing values

Figure A4. Decision tree for the managing of missing values for further analytical steps. For each variable, different checks are conducted, based on which a decision is made on how to handle missing values. Variables used to comply with the node ‘Comorbidities’ in the causal model (i.e., used to compute ‘comorbidities_bl’ in the analytical pipeline): *diabetes_bl*, *obesity_bl*, *heart_failure_bl*, *copd_bl*, *solid_tumor_without_metastasis_bl*, *chronic_kidney_disease_bl*, *sickle_cell_disease_bl*, *hypertension_bl*, *chronic_liver_disease_bl*. Variables used to comply with the node ‘Immune status’ in the causal model (i.e., used to compute ‘immunestatus_bl’ in the analytical pipeline): *blood_cancer_bl*, *transplanted_bl*, *hiv_infection_bl*, *primary_immunodeficiency_bl*, *immunosuppression_bl*. Core variables: *age_nm*, *sex_cd*, *residence_area_cd*, *pregnancy_bl*, *essential_worker_bl*, *institutionalized_bl*, *foreign_bl* and *socecon_lvl_cd*. Missing values in core variables will obstruct subsequent analysis steps. MCAR: missing completely at random; Is it reasonable to assume MCAR in the subset of variables used to predict missing values (core variables plus variables used to comply with the node ‘Comorbidities’ and ‘Immune status’ in the causal model, with less than 15% missing values)? Reasonability of MCAR is tested using Little’s test [2]. Pmm: predictive mean matching (matching method).

Appendix E – Matching algorithm

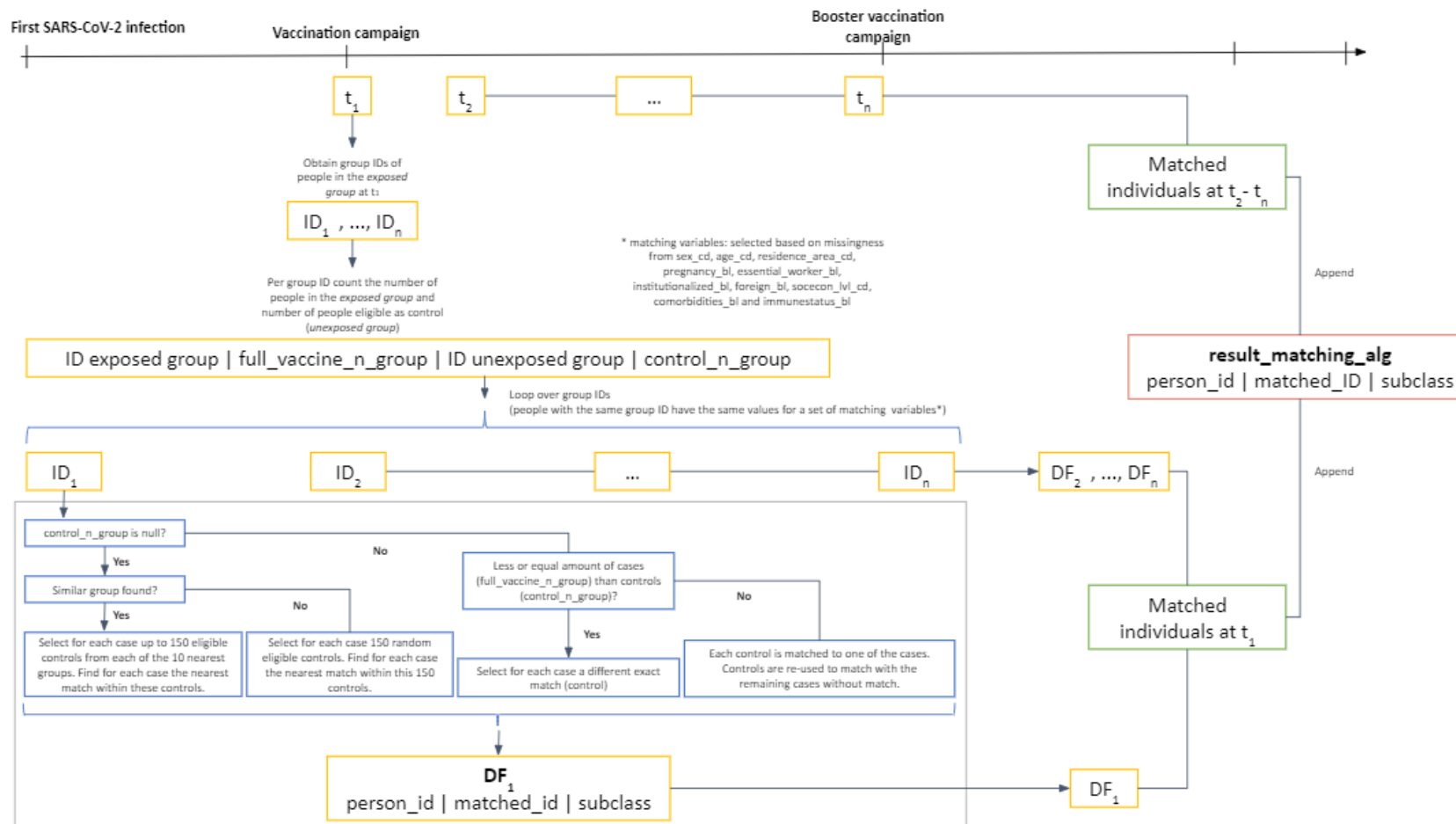


Figure A5. Graphical overview of the automated matching process. t_1 - t_n : days considered within the enrollment period. Group ID: ID of groups identified based on combinations of considered covariates (i.e., matching variables). People in the exposed group at t_x : individuals completing their primary vaccination schedule at the day t_x . 'result_matching_alg': DuckDB database table, used to store matching results.

Appendix F – Timeline of the sequentially emulated target trial

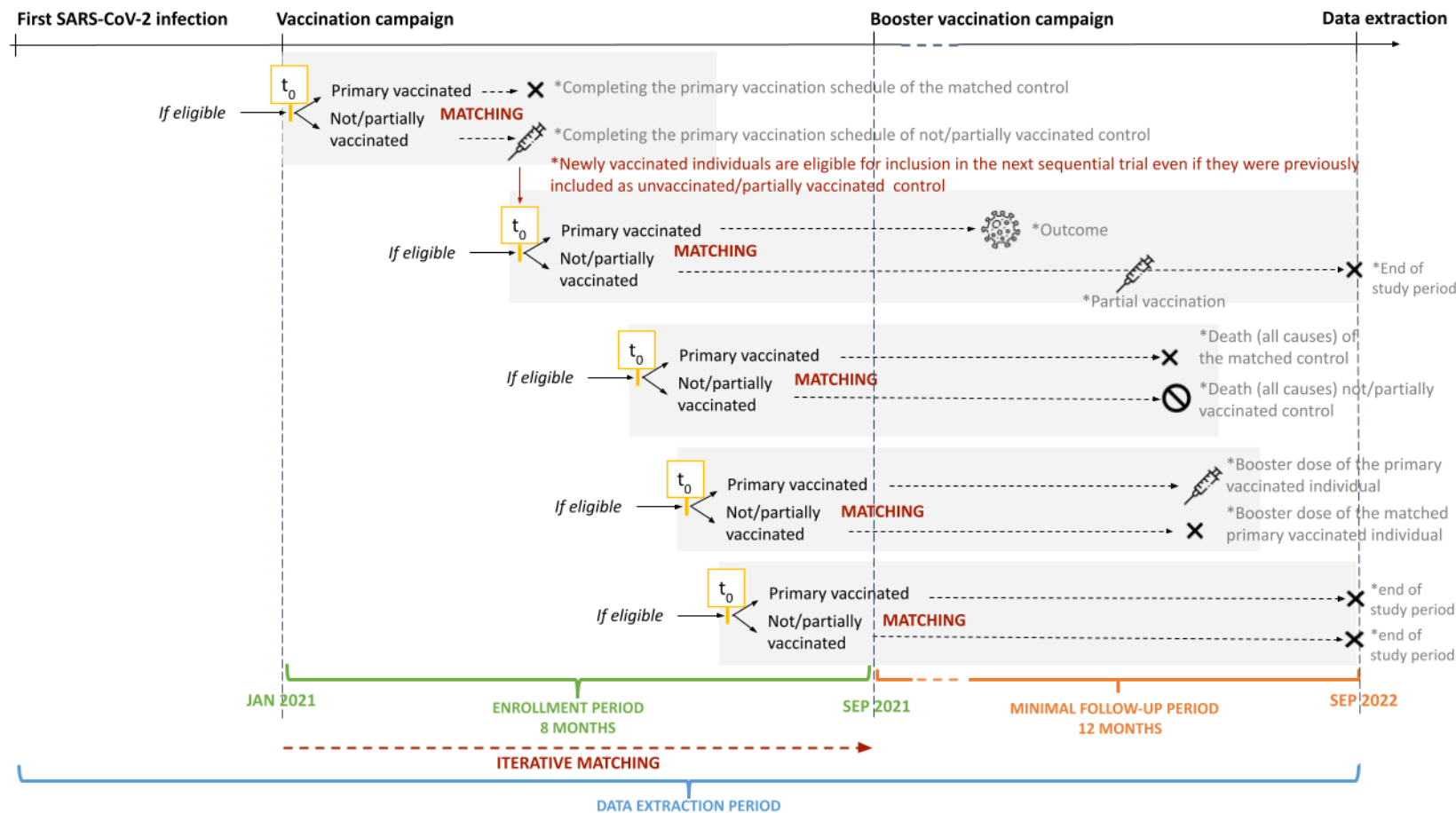


Figure A6. Timeline of the sequentially emulated target trial, by reusing observational data within each participating region or country. t_0 : baseline, at which the vaccination status of individuals is assessed and matching is performed. Enrollment period: January 1, 2021 to September 1, 2021. Data extraction period: starting from the date of the first documented SARS-CoV-2 infection until the most recent date at which data is available at the time of analysis. End of the study period: end of the data extraction period.

Appendix G – Data extraction periods

Table A2. *Data extraction periods (i.e., starting from the date of the first documented SARS-CoV-2 infection until the most recent date at which data is available at the time of analysis) in the different sites: Aragon (Spain), Brussels and Wallonia (Belgium), and Finland*

Site	Start data extraction period	End data extraction period
Aragon (Spain)	2020-03-01	2023-04-18
Brussels and Wallonia (Belgium)	2020-09-01	2024-04-05
Finland	2020-01-01	2022-12-20

Appendix H – Characteristics of the unmatched population eligible for matching**Table A3.** Characteristics of the unmatched population eligible for matching by Participant Node (Aragon (Spain), Brussels and Wallonia (Belgium), and Finland) and by intervention group.

	Aragon (Spain)		Brussels and Wallonia (Belgium)		Finland	
	Primary vaccinated* (N = 755,055)	Not primary vaccination** (N = 244,365)	Primary vaccinated* (N = 2,123,328)	Not primary vaccination** (N = 697,285)	Primary vaccinated* (N = 2,588,318)	Not primary vaccination** (N = 2,263,277)
Sex, n (%)						
1 (male)	360,616 (47.8%)	126,054 (51.6%)	1,019,290 (48.0%)	343,475 (49.3%)	1,209,685 (46.7%)	1,184,459 (52.3%)
2 (female)	394,439 (52.2%)	118,311 (48.4%)	1,104,038 (52.0%)	353,810 (50.7%)	1,378,633 (53.3%)	1,078,818 (47.7%)
Age (category), n (%)						
2 (5-9 yo)	0 (0.0%)	37,945 (15.5%)	11 (0.0%)	77,059 (11.1%)	0 (0%)	273,060 (12.1%)
3 (10-14 yo)	318 (0.0%)	52,280 (21.4%)	10,611 (0.5%)	121,717 (17.5%)	2,242 (0.1%)	274,985 (12.1%)
4 (15-19 yo)	16,314 (2.2%)	33,824 (13.8%)	80,797 (3.8%)	86,070 (12.3%)	37,872 (1.5%)	224,984 (9.9%)
5 (20-24 yo)	20,863 (2.8%)	24,530 (10.0%)	118,337 (5.6%)	53,752 (7.7%)	55,185 (2.1%)	217,002 (9.6%)
6 (25-29 yo)	24,610 (3.3%)	23,541 (9.6%)	119,579 (5.6%)	53,672 (7.7%)	79,327 (3.1%)	233,172 (10.3%)
7 (30-34 yo)	36,643 (4.9%)	16,814 (6.9%)	134,279 (6.3%)	56,726 (8.1%)	99,915 (3.9%)	210,038 (9.3%)
8 (35-39 yo)	50,711 (6.7%)	13,937 (5.7%)	140,687 (6.6%)	48,514 (7.0%)	140,313 (5.4%)	180,610 (8.0%)
9 (40-44 yo)	72,485 (9.6%)	10,592 (4.3%)	152,447 (7.2%)	40,357 (5.8%)	184,719 (7.1%)	126,999 (5.6%)
10 (45-49 yo)	79,881 (10.6%)	8,180 (3.3%)	170,927 (8.1%)	32,306 (4.6%)	195,696 (7.6%)	90,088 (4.0%)

11 (50-54 yo)	76,979 (10.2%)	5,995 (2.5%)	185,092 (8.7%)	27,729 (4.0%)	246,049 (9.5%)	80,071 (3.5%)
12 (55-59 yo)	74,495 (9.9%)	4,556 (1.9%)	195,044 (9.2%)	23,020 (3.3%)	270,896 (10.5%)	72,732 (3.2%)
13 (60-64 yo)	66,025 (8.7%)	4,653 (1.9%)	194,349 (9.2%)	18,367 (2.6%)	280,746 (10.8%)	61,802 (2.7%)
14 (65-69 yo)	56,885 (7.5%)	2,458 (1.0%)	179,502 (8.5%)	14,377 (2.1%)	296,567 (11.5%)	51,760 (2.3%)
15 (70-74 yo)	53,710 (7.1%)	1,276 (0.5%)	158,875 (7.5%)	12,431 (1.8%)	300,864 (11.6%)	43,919 (1.9%)
16 (75-79 yo)	45,802 (6.1%)	1,013 (0.4%)	110,856 (5.2%)	9,621 (1.4%)	179,299 (6.9%)	30,545 (1.3%)
17 (80-84 yo)	33,447 (4.4%)	888 (0.4%)	76,632 (3.6%)	7,730 (1.1%)	125,315 (4.8%)	34,565 (1.5%)
18 (>=85 yo)	45,887 (6.1%)	1,883 (0.8%)	81,108 (3.8%)	9,649 (1.4%)	93,313 (3.6%)	56,945 (2.5%)
Pregnancy, n (%)						
Yes	13,395 (1.8%)	9,341 (3.8%)	-	-	-	-
No	741,660 (98.2%)	235,024 (96.2%)	-	-	-	-
Essential worker, n (%)						
Yes	4,991 (0.7%)	621 (0.3%)	117,847 (5.6%)	21,567 (3.1%)	221,391 (8.6%)	172,063 (7.6%)
No	750,064 (99.3%)	243,744 (99.7%)	2,005,481 (94.4%)	675,718 (96.9%)	2,365,934 (91.4%)	2,048,942 (90.5%)
Institutionalised, n (%)						
Yes	6,661 (0.9%)	557 (0.2%)	17,562 (0.8%)	1,498 (0.2%)	-	-
No	748,394 (99.1%)	243,808 (99.8%)	2,105,766 (99.2%)	695,787 (99.8%)	-	-
Socio-economic level, n (%)						

Higher	-	-	620,180 (29.2%)	133,678 (19.2%)	376,469 (14.5%)	311,272 (13.8%)
Intermediate	-	-	643,492 (30.3%)	187,060 (26.8%)	570,270 (22.0%)	633,469 (28.0%)
Lower	-	-	859,436 (40.5%)	376,455 (54.0%)	1,640,586 (63.4%)	1,276,264 (56.4%)
Foreign, n (%)						
Yes	81,524 (10.8%)	5,225 (22.2%)	-	-	-	-
No	673,531 (89.2%)	190,140 (77.8%)	-	-	-	-
Comorbidities, n (%)						
Yes	326,172 (43.2%)	28,768 (11.8%)	549,014 (25.9%)	65,621 (9.4%)	1,109,074 (42.8%)	430,287 (19.0%)
No	428,883 (56.8%)	215,597 (88.2%)	1,574,314 (74.1%)	631,664 (90.6%)	1,479,244 (57.2%)	1,832,990 (81.0%)
Immune status, n (%)						
Yes	22,630 (3.0%)	2,016 (0.8%)	60,697 (2.9%)	8,303 (1.2%)	575,741 (22.2%)	194,678 (8.6%)
No	732,425 (97.0%)	242,349 (99.2%)	2,062,631 (97.1%)	688,982 (98.8%)	2,012,577 (77.8%)	2,068,599 (91.4%)

* Eligible individuals who completed a primary vaccination schedule during the enrollment period (i.e., January 1, 2021 to September 1, 2021)

** Eligible individuals who did not complete a primary vaccination schedule during the enrollment period (i.e., January 1, 2021 to September 1, 2021)

Appendix I – Propensity score distribution in contrasted exposure groups before and after matching in the considered sites

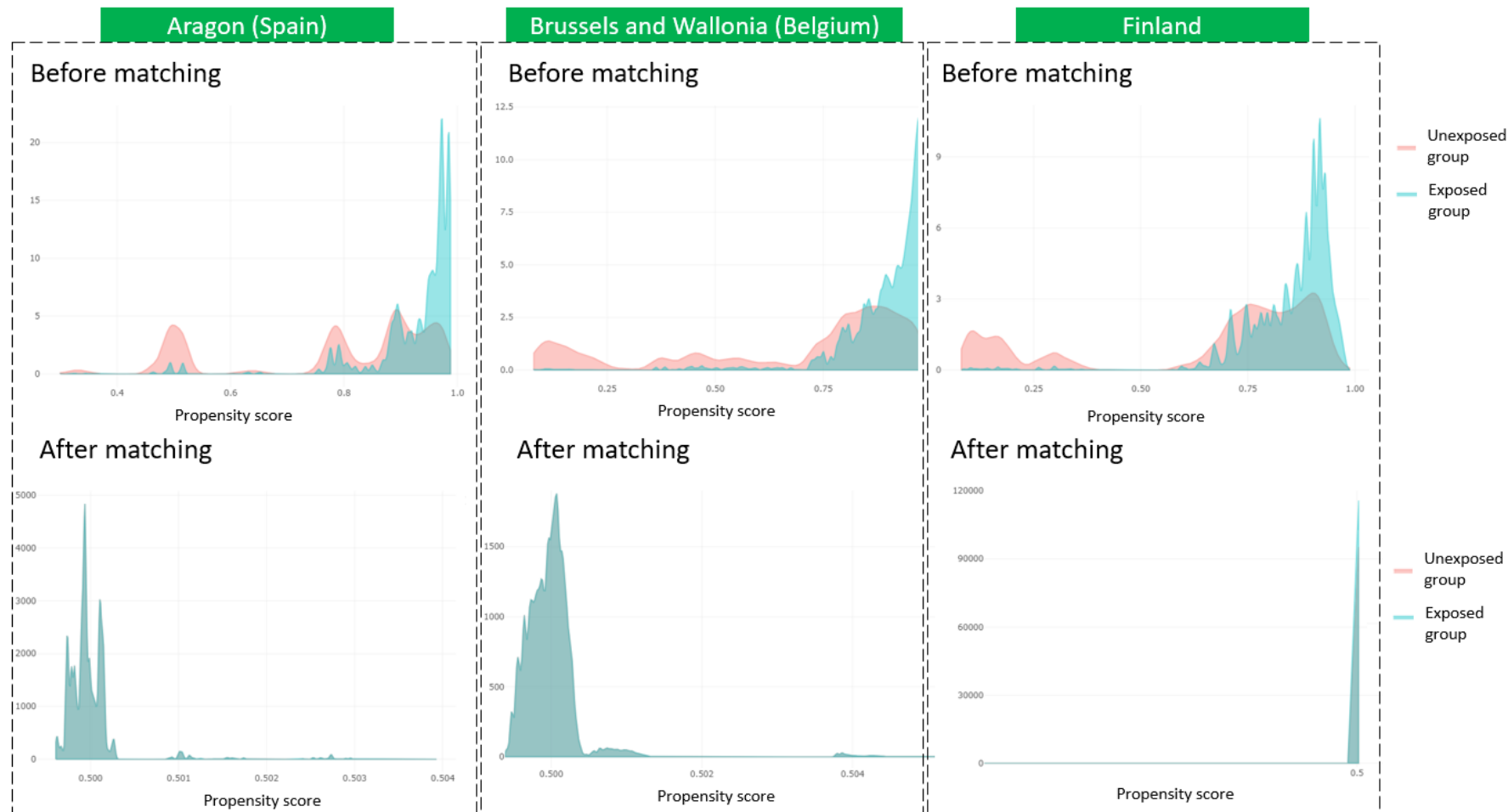


Figure A7. Propensity score distribution plots in contrasted exposure groups before and after matching in the considered sites: Aragon (Spain), Brussels and Wallonia (Belgium), Finland.

Appendix J – Graphical representation of Standardized Mean Differences (SMDs) before and after matching between contrasted exposure groups for each covariate considered for matching in the different sites

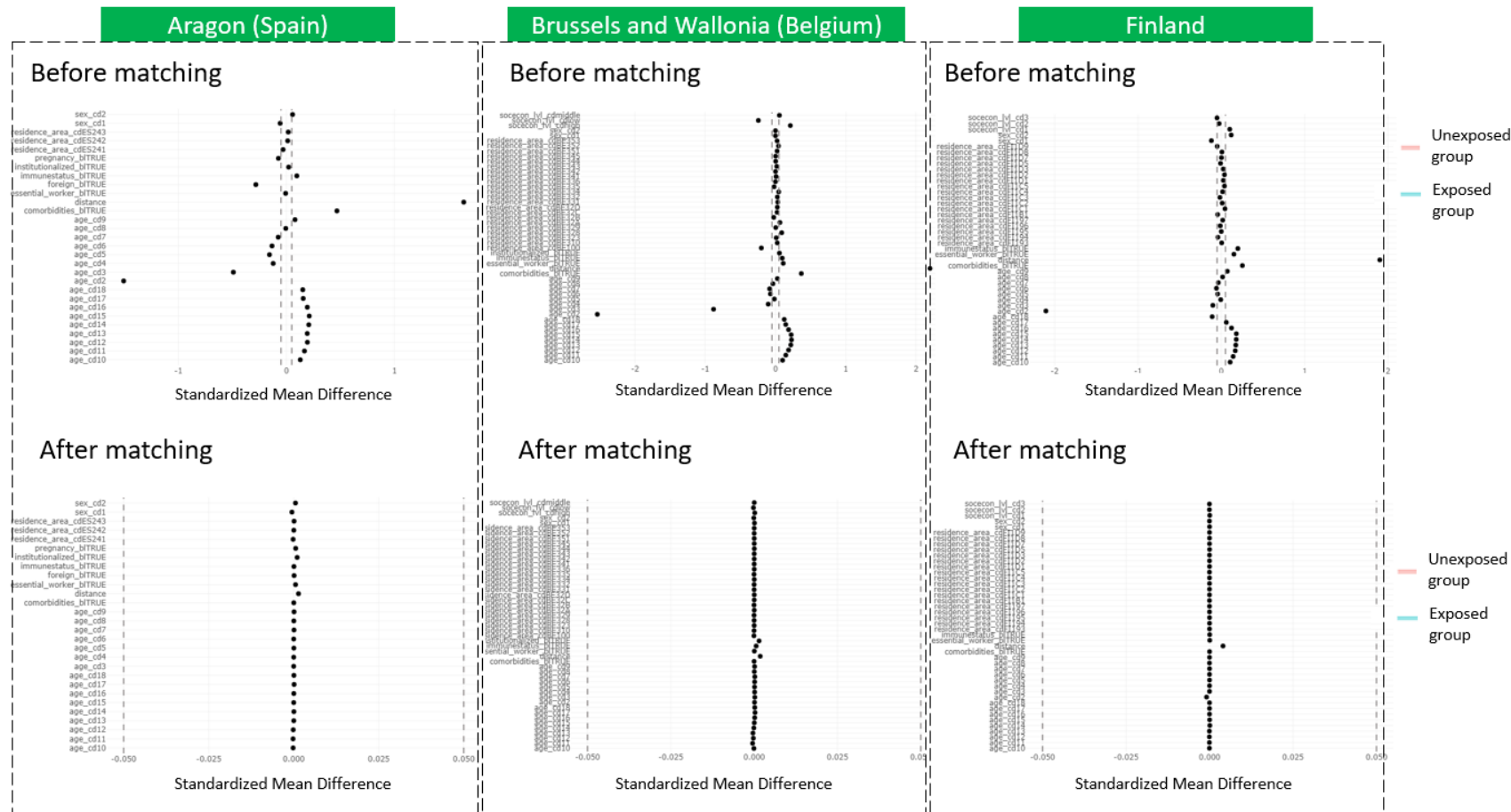


Figure A8. Graphical representation of Standardized Mean Differences (SMDs) before and after matching between contrasted exposure groups for each covariate considered for matching in the different sites: Aragon (Spain), Brussels and Wallonia (Belgium), Finland.

Appendix K – Standardized Mean Differences (SMDs) before and after matching between contrasted exposure groups for each covariate considered for matching, in the different sites

Table A4. Standardized Mean Differences before and after matching between contrasted exposure groups, for each covariate considered for matching, Aragon (Spain).

	Standardized Mean Difference before matching	Standardized Mean Difference after matching
Distance	1.641725020	1.412596e-03
sex_cd1	-0.057432835	-5.223406e-04
sex_cd2	0.057432835	5.223406e-04
age_cd2	-1.506233112	/
age_cd3	-0.490890385	0.000000e+00
age_cd4	-0.122138493	9.109062e-06
age_cd5	-0.156618157	4.847949e-05
age_cd6	-0.132791923	7.458482e-06
age_cd7	-0.075861162	9.245043e-05
age_cd8	-0.005404281	4.232983e-05
age_cd9	0.079957452	1.078980e-04
age_cd10	0.128666317	-1.679326e-04
age_cd11	0.167351509	-1.882102e-04
age_cd12	0.194435150	-1.332369e-04
age_cd13	0.193346383	2.813053e-05
age_cd14	0.207429394	6.021460e-05
age_cd15	0.211616943	1.030473e-05
age_cd16	0.193917382	3.883785e-05
age_cd17	0.156043095	1.222994e-04
age_cd18	0.152091086	6.097802e-05
residence_area_cdES241	-0.030511385	-1.641985e-04

residence_area_cdES242	0.012192087	3.559674e-05
residence_area_cdES243	0.017448286	1.143012e-04
pregnancy_bITRUE	-0.073902226	5.819078e-04
essential_worker_bITRUE	-0.006835831	5.393504e-04
institutionalized_bITRUE	0.021122908	1.019759e-03
foreign_bITRUE	-0.282783266	1.237589e-04
comorbidities_bITRUE	0.468086506	2.406300e-05
immunestatus_bITRUE	0.096718282	6.990668e-05

Table A5. Standardized Mean Differences between contrasted intervention groups, for each covariate considered for matching, Brussels and Wallonia (Belgium).

	Standardized Mean Difference before matching	Standardized Mean Difference after matching
Distance	2.1992597402	1.828475e-03
sex_cd1	-0.0001550932	9.615221e-05
sex_cd2	0.0001550932	-9.615221e-05
age_cd2	-2.5354036348	1.160008e-04
age_cd3	-0.8810198701	2.040529e-04
age_cd4	-0.1037882622	1.474956e-04
age_cd5	-0.0160622132	1.271869e-04
age_cd6	-0.0720882998	1.524582e-04
age_cd7	-0.0831480991	2.004161e-04
age_cd8	-0.0368123851	5.293941e-05
age_cd9	0.0267192980	1.363965e-04
age_cd10	0.1012464192	-7.068662e-05
age_cd11	0.1466775278	-4.552732e-04
age_cd12	0.1865733535	-2.327264e-04
age_cd13	0.2181298788	-3.532912e-04
age_cd14	0.2274606522	-8.257133e-05

age_cd15	0.2231181058	4.819998e-05
age_cd16	0.1865803011	1.647955e-04
age_cd17	0.1458589159	2.696093e-04
age_cd18	0.1249091665	1.865252e-04
residence_area_cdBE100	-0.2009285148	-5.906023e-05
residence_area_cdBE310	0.0252622841	-2.544553e-05
residence_area_cdBE323	0.0106126766	8.193615e-06
residence_area_cdBE328	0.0885347426	-4.255294e-06
residence_area_cdBE329	0.0038294871	-2.408440e-05
residence_area_cdBE32A	0.0635141327	3.229259e-05
residence_area_cdBE32B	-0.0251878781	-3.402086e-05
residence_area_cdBE32C	0.0227596552	2.658916e-05
residence_area_cdBE32D	0.0263528162	3.787469e-05
residence_area_cdBE331	0.0325206280	0.000000e+00
residence_area_cdBE332	0.0262957873	-5.503277e-06
residence_area_cdBE334	0.0458176653	3.477750e-05
residence_area_cdBE335	-0.0194912089	-2.916835e-05
residence_area_cdBE336	-0.0002299111	2.795237e-05
residence_area_cdBE341	0.0084663934	1.044831e-05
residence_area_cdBE342	0.0007837180	2.460222e-05
residence_area_cdBE343	0.0173921956	2.009669e-05
residence_area_cdBE344	0.0032082133	5.086846e-05
residence_area_cdBE345	0.0023742138	3.290944e-05
residence_area_cdBE351	0.0229490948	2.887334e-05
residence_area_cdBE352	0.0426394233	1.622087e-05
residence_area_cdBE353	0.0227093472	2.235767e-05
essential_worker_bITRUE	0.1102706999	6.170960e-05
institutionalized_bITRUE	0.0563316624	1.471616e-03

socecon_lvl_cdhigh	0.2112528812	1.822808e-04
socecon_lvl_cdlow	-0.2438007833	-2.014880e-04
socecon_lvl_cdmiddle	0.0559786835	3.484091e-05
comorbidities_bITRUE	0.3676178677	-1.075629e-05
immunestatus_bITRUE	0.0963132700	6.217679e-04

Table A6. Standardised Mean Differences between contrasted intervention groups, for each covariate considered for matching, Finland.

	Means control group	Standardized Mean Difference
Distance	1.9020165178	4.034454e-03
sex_cd1	-0.1185661440	1.703579e-05
sex_cd2	0.1185661440	-1.703579e-05
age_cd2	-2.1044574284	-8.790352e-04
age_cd3	-0.1015375309	0.000000e+00
age_cd4	-0.0066189695	9.652817e-06
age_cd5	-0.0437475031	1.069844e-05
age_cd6	-0.0581121245	-2.241506e-06
age_cd7	-0.0358672901	1.002754e-05
age_cd8	0.0158830199	-1.194381e-05
age_cd9	0.0730296968	3.001534e-06
age_cd10	0.1061768421	-2.922819e-05
age_cd11	0.1424234383	-3.424871e-05
age_cd12	0.1674116536	-4.291163e-05
age_cd13	0.1736996983	-3.851477e-05
age_cd14	0.1792646857	9.339973e-05
age_cd15	0.1803915762	3.495720e-05
age_cd16	0.1224049084	0.000000e+00
age_cd17	0.0600141517	0.000000e+00
age_cd18	-0.1093198767	0.000000e+00

residence_area_cdFI193	0.0046732194	3.624173e-06
residence_area_cdFI194	-0.0408667432	8.582977e-06
residence_area_cdFI195	-0.0004519568	6.563667e-06
residence_area_cdFI196	-0.0117511613	0.000000e+00
residence_area_cdFI197	0.0172933014	-3.988977e-06
residence_area_cdFI1B1	-0.0436668134	-9.252919e-06
residence_area_cdFI1C1	0.0443360061	-2.721600e-06
residence_area_cdFI1C2	0.0106594329	-2.177293e-06
residence_area_cdFI1C3	-0.0164409409	1.938849e-06
residence_area_cdFI1C4	0.0155803032	0.000000e+00
residence_area_cdFI1C5	0.0377920931	-1.233891e-05
residence_area_cdFI1D1	0.0245363444	9.299805e-06
residence_area_cdFI1D2	0.0380839970	0.000000e+00
residence_area_cdFI1D3	0.0238659883	-2.253008e-06
residence_area_cdFI1D5	-0.0090660344	2.024242e-05
residence_area_cdFI1D7	0.0020039128	0.000000e+00
residence_area_cdFI1D8	0.0072482953	6.790062e-06
residence_area_cdFI1D9	-0.0493560583	1.501142e-06
essential_worker_blTRUE	0.1516587655	6.767283e-05
socecon_lvl_cd1	0.1019365723	6.683983e-05
socecon_lvl_cd2	-0.0259588197	-2.982094e-05
socecon_lvl_cd3	-0.0512514925	-2.325936e-05
comorbidities_blTRUE	0.2535122393	1.795675e-05
immunestatus_blTRUE	0.1991196427	5.016536e-05

Appendix L – Change-in-estimate analysis to estimate the magnitude of confounding by individual-level SES

Table A7. Restricted Mean Survival Times (RMSTs) in contrasted exposure groups, and Restricted Mean Survival Time Difference (RMSTD, RMST intervention group - RMST control group) resulting from local analysis in Brussels and Wallonia (Belgium), with the time-to-event analysis executed on a matched study population obtained by matching on all previously included factors, excluding individual-level SES.

Measure	Brussels and Wallonia (Belgium)
RMST [95% CI], unexposed group	216.120 [215.828; 216.412]
RMST [95% CI], exposed group	273.854 [273.651; 274.058]
RMSTD [95% CI]	57.734 [57.378; 58.090]
P-value RMSTD	<0.001

CHAPTER 5

Experts' experiences and perceptions towards the integration of socio-economic and genomic data for public health research

Including information from:

Van Goethem N, **Meurisse M**, Estupiñán Romero F, Bernal-Delgado E, Saldner S, Kalaitzi V. BY-COVID - WP5 - Baseline Use Case: Integration of socioeconomic data in observational studies using a prototyped workflow standard to federated population health research (1.0.0). 2024. <https://doi.org/10.5281/zenodo.12771703>

Van Goethem N, Bernal-Delgado E, Estupiñán-Romero F, **Meurisse M**. BY-COVID D5.1 Enriched report viral variants and health outcomes. 2024. <https://doi.org/10.5281/zenodo.13870075>

1. INTRODUCTION

Causal inference has an important role in public health research and policy, aiding in understanding the impact of exposures, treatments, or interventions on health outcomes. Using “real-world” observational data offers opportunities for informing health policy by assessing relationships in uncontrolled settings, with broader populations and at lower costs. However, this approach faces challenges due to inherent biases present in observational data. To improve the accuracy of causal estimates obtained from observational studies, emulating a hypothetical randomized experiment or target trial can reduce avoidable biases. Applying such a causal framework often implies the need for detailed individual-level data in order to mitigate confounding.

Health-related data, i.e., data providing information on health status or health determinants of given individuals or populations, are typically obtained from the interaction of an individual or population with the health system, for example administrative information, diagnoses, treatments, lab tests, etc. These data are often originating from multiple heterogeneous data sources, such as Electronic Health Records (EHR), healthcare claims, pharmacy records, laboratory data, and specific disease registries. Combining health-related data with diverse alternative data sources enables adjustment for various confounding factors and a comprehensive analysis of how genetic, social, economic and environmental factors interact to influence public health.

Individual-level data might not always be accessible for mobilization and linkage due to privacy concerns. As an alternative, area-level approximations of individual characteristics can be employed. Area-level approximations can be obtained by linking individuals to a geographically defined area (e.g., based on their area of residence) and assigning a factor measured at this area level to the individuals linked to this respective area. Area-level approximations can potentially capture contextual factors and might be readily available in open data repositories, without requiring extensive data access applications. However, interpretation across area-level groups might be vulnerable to the ecological fallacy.

In the context of a use case within the European BeYond-COVID (BY-COVID) project, we performed a comparative analysis of the real-world effectiveness of SARS-CoV-2 primary vaccination campaigns in preventing SARS-CoV-2 infections by leveraging data from existing COVID-19 registries from Aragon (Spain), Brussels and Wallonia (Belgium) and Finland [1], and adopting a reproducible federated analysis methodology [2]. In this work, the integration of diverse data sources was required to adjust for confounding and to allow for the execution of secondary analyses. Within the premises of the different Participant Nodes (i.e., institutions having access to individual-level COVID-19 and population

registries), health-related data from the local SARS-CoV-2 vaccination population registry, COVID-19 public health surveillance and monitoring information system, insurance registry or health system users database, and EHRs were leveraged.

Enriching this cross-border comparative analysis by integrating additional types of data from other domains beyond those established within the public health sector, such as those data sources originating from the Social Sciences and Humanities (SSH) or molecular life sciences, offers a multidisciplinary approach towards understanding health outcomes and disparities. Despite its importance for public health, input from SSH research into policy-making was often limited during the COVID-19 pandemic compared to other disciplines, like biomedical research [3]. Socio-economic data can provide important insights into social and economic health determinants and contribute to the detection of patterns in health inequality [4]. For many exposure-outcome relationships under investigation in public health research the socio-economic status (SES) of individuals and their surrounding is considered to potentially confound the association. By taking the influence of SES into account, researchers can consider its impact and warrant more accurate and reliable findings. Further, testing whether differences in the effect of interest across individuals with contrasting SES exist (i.e., investigating its role as a moderator), can be of key importance for the further development of intervention programs, strategies and priorities. Integrating socio-economic data into public health policy-making has the potential to enhance research, expertise and evidence diversity in future public health crises.

Integrating genomic data adds another layer of depth for public health research by offering a molecular-level perspective. Taking into account pathogen genomic data allows researchers to obtain a comprehensive view of infectious disease dynamics and host-pathogen interactions [5,6]. In parallel, integrating host (i.e., human) genetic data allows researchers to uncover genetic predispositions and variations that contribute to individual and public health [7].

Data originating from the SSH or molecular life sciences field, presented here as a proof of concept for integrating heterogeneous data sources from different domains, can provide an important source of information for public health research. However, since these data sources are often stored decentralized in isolated environments, integrating multidisciplinary observational data, especially across national borders, is a challenging undertaking. This work aims to gather experiences, perceptions, needs and expectations of multiple stakeholders (e.g., public health professionals, data providers) on integrating socio-economic and pathogen genomic data with other health-related data to address public health needs for pandemic preparedness.

The work described in this chapter was conducted in the context of the European BY-COVID project. As part of this project, we have organized workshop on socio-economic data integration for observational public health studies. The current chapter includes information from the complementary workshop report [8], however, tailored towards the objectives of this doctoral thesis. Further, it describes qualitative analyses conducted in the context of this thesis, of which conclusions were secondly shortly described within a broader BY-COVID project deliverable.

2. METHODS

2.1 Socio-economic data integration

2.1.1 Theoretical background: literature search

A non-systematic literature search was conducted to compile a non-exhaustive overview of available socio-economic indicators and data sources relevant for public health research. A broad search strategy was adopted, consulting an academic database (i.e., PubMed), and governmental or organizational websites and reports to gather diverse literature and documentation on the topic.

2.1.2 Experiences from Participant Nodes

To conduct a cross-border comparison of SARS-CoV-2 vaccine effectiveness in the BY-COVID use case, the different Participant Nodes were requested to locally mobilize and link different data sources in order to comply with the data requirements. This included data from the SSH field, to gain information on the socio-economic status of individuals. After local linkage and transformation of the data, a reproducible analysis pipeline was distributed and executed locally using these integrated data as input, and generating aggregated, non-sensitive and therefore sharable results that allow the execution of a comparative assessment. The analyses were executed in three Participant Nodes: Instituto Aragonés de Ciencias de la Salud (IACS), the Belgian Institute for Health (Sciensano), and the Finnish Institute for Health and Welfare (THL). Gesundheit Österreich GmbH (GÖG) was initially also recruited to function as a Participant Node in this study. However, access to the required data was not reached within the project's time frame.

To gather experiences on the integration of individual-level socio-economic data in the context of the BY-COVID use case, a mapping exercise among the Participant Nodes was executed following the analysis executions, completing information on the local availability of individual-level socio-economic data (i.e., information regarding data sources, data providers, discoverability, and data access), individual-level socio-economic data integration (i.e., information regarding pseudonymization

procedures and integration barriers and facilitators), and individual-level socio-economic data use (e.g., frequently used indicators).

2.1.3 Online workshop

2.1.3.1 Workshop design

The interactive online workshop, entitled “BY-COVID Spring 2024 Baseline Use Case Workshop: Integration of individual-level socio-economic data for infectious diseases research and prevention in Europe”, was conducted on the 6th of June 2024, 9:30-11:30 CET [9,10], after completion of the analyses in three Participant Nodes (IACS, Sciensano, and THL). The workshop was co-organized by the Dutch national center of expertise and repository for research data – Royal Netherlands Academy of Arts and Sciences (DANS – KNAW), Sciensano and IACS as part of the BY-COVID project. Through organizing the workshop, it was aspired to share, evaluate and discuss experiences on integrating individual-level socio-economic data in the context of the BY-COVID use case, and gather expert opinions, needs and expectations on the topic. We aimed to enhance insights on the identification, integration, and analysis of individual-level socio-economic data in Europe, identifying and understanding challenges faced, searching for solutions to overcome them, and generalizing these solutions to various disciplinary and geographical contexts.

2.1.3.2 Participant selection and recruitment

The workshop targeted researchers working with sensitive data and having experience with or being interested in solutions towards the integration of individual-level socio-economic data for public health research. Participants were recruited by creating a workshop invitation on the BY-COVID project website [9], sharing this invitation by the BY-COVID project social media and email channels, and direct invitations of researchers in the organizers’ network.

2.1.3.3 Workshop activities

The workshop consisted of presentations from BY-COVID participants working on the described use case, covering (1) a description of the adopted federated methodology and (2) experiences from the different Participant Nodes (IACS, Sciensano, THL, and GÖG) on the integration of individual-level socio-economic data with health-related data in the context of the use case and the availability of individual-level socio-economic data in their region or country [10]. Following these presentations, an interactive session was organized, engaging participants in a group discussion on different questions related to the topic: (1) “Which challenges did you experience when mobilizing and linking sensitive social science data for health research and how have you faced those challenges?”, (2) “How can we

improve the availability and accessibility of sensitive socio-economic data for EU-level health research?”, and (3) “What are limitations of existing socio-economic data sources?”. During this session, a Miro board was used to facilitate interaction by providing a dynamic and collaborative platform where participants could engage and visually organize and share their ideas in real-time. This way, the workshop aimed to facilitate knowledge exchange and collaboration.

2.1.3.4 Data collection and analysis

Data collected during the workshop include audio recordings of the presentations and the workshop Miro board with participant notes [10]. Notes obtained during the interactive session were analyzed and related ideas were clustered to identify key themes and insights.

2.2 Genomic data integration

2.2.1 Theoretical background: literature search

A non-systematic literature search was executed to obtain a greater knowledge on the availability of open pathogen genomic data. An academic database (i.e., PubMed), and governmental and organizational websites were browsed broadly to obtain a non-exhaustive overview of pathogen genomic data repositories and their characteristics (i.e., year of launch, conditions for data access, hosted data and accepted formats).

2.2.2 Expert consultation: in-depth interviews

To gather expert experiences, opinions, needs and expectations from stakeholders on the integration of pathogen genomic data (from the molecular biology field) with health-related data to address public health research needs for pandemic preparedness in Europe, we conducted interviews with key experts from institutions across different countries on the topic.

2.2.2.1 Interview design

A qualitative approach was employed, using semi-structured interviews to explore the experiences, unmet needs, and future prospects of different experts in the field. An interview guide was developed based on existing literature and the preceding literature review, including a series of open-ended questions focusing on key areas such as their experience (e.g., facilitators and barriers) and their prospects on the re-use of sensitive and open genomic data (see Appendix A). Two pilot interviews were conducted to test the interview questions and pinpoint weaknesses in the interview design,

resulting in minor adjustments to the interview guide. A participant information sheet was additionally prepared, specifying the context and purpose of the interview (see Appendix B).

2.2.2.2 Interview participants

Participants could be researchers working on different domains (e.g., epidemiologists, data scientists, public health surveillance professionals, genomics computational scientists), involved in public health research integrating or aiming to link genomic data with other real-world health-related data for public health research in the World Health Organization (WHO) European region, potentially within the context of a European project. These participants could be BY-COVID project consortium members, individuals referred by BY-COVID project consortium members, or individuals working for institutions involved in European projects. A distributed representation across the European Union was pursued. Potential participants were identified from the network of contributing researchers and through browsing relevant literature. They were invited via email including the participant information sheet.

2.2.2.3 Data collection and analysis

Semi-structured interviews were conducted between January 29th and February 23th 2024. The interviews were conducted online (through video conference) or in person, depending on the participant's preferences and logistics. The interview duration was around 45 to 60 minutes, however the exact duration fluctuated across participants. The interview was conducted following the predesigned interview guide, providing a consistent framework for the interviews, while allowing for flexibility in the responses. Participants received a comprehensive explanation of the study and the informed consent, and there was the opportunity to ask questions before the interview. With the participant's permission, the audio and video of the interview was recorded, based on which transcriptions were subsequently generated verbatim using *Descript's* online tool.

Transcripts were analyzed using thematic analysis to identify key themes and insights. We followed a six-step analysis process [11,12]: (1) familiarizing ourselves with the data (i.e., reading the transcripts several times, writing down early thoughts, selecting relevant phrases or sentences), (2) selecting keywords (i.e., close examination of the data, derive keywords from selected text), (3) open coding (i.e., assigning codes to the selected text segments capturing a core message without starting from pre-defined codes, simplifying complex textual data), (4) searching for themes inductively (i.e., organizing and clustering codes to discern patterns providing insights into the topic, based on the collected data), (5) reviewing, defining, and naming themes (i.e., reviewing initial themes, comparing them with the entire dataset, defining and assigning an understandable name to each theme), and (6)

conceptualization and reporting (i.e., summarize findings into a diagram linking themes, offering an analytic narrative).

2.2.2.4 Ethical considerations

An informed consent form, specifying how their data would be used and processed, was signed by participants preceding the interview (see Appendix C). Anonymity and confidentiality were maintained throughout the data collection, analysis and reporting process.

3. RESULTS

3.1 Socio-economic data integration

3.1.1 State-of-the-art

A non-exhaustive overview of available socio-economic indicators and data sources relevant for public health research is provided in Appendix D (Table A1, A2).

3.1.2 Experiences from Participant Nodes

Tables A3 – A6 in Appendix E present a mapping exercise of the local availability of individual-level socio-economic data among the BY-COVID use case Participant Nodes from different sites: Sciensano (Belgium), IACS (Aragon, Spain), THL (Finland), and GÖG (Austria).

In Belgium, socio-economic information originates from Statistics Belgium (Statbel). Researchers can request microdata from Statbel for scientific purposes. For the use case, individual-level data from Statbel on the net taxable income on household level, education, employment status, household type, and migration background were accessible to designated researchers and available for linkage to other health-related data sources in a pseudonymized secured processing environment (SPE) provided in the context of the LINK-VACC project (i.e., a project that links selected variables from existing national registries for COVID-19 vaccine surveillance to ensure the monitoring of COVID-19 vaccines in the phase following their marketing authorization) [13]. Some challenges were noted, such as the lack of a clear overview of the available datasets and variables or detailed documentation on the methodology for data collection, difficulty in selecting indicators as non-experts in the SSH field, lengthy data access procedures, and infrequent updates. However, effective collaboration between Statbel and governmental and academic institutions facilitates data sharing and access for research.

In Finland, socio-economic information originates from Statistics Finland. Under the Statistics Act, Statistics Finland is authorized to share sensitive data with other statistical authorities, including THL, for the generation of statistics. Further, data that was collected for statistical objectives can be made available for scientific research through a remote access system (to allow secure data processing). Preceding research data access, an application process to obtain a microdata user license is required. Individual-level data was available for linkage in a SPE on occupation using the European Socio-economic Classification (ESeC) for their participation in the use case. Detailed data access applications and stringent data protection regulations were also noted as a barrier in Finland.

In Spain, socio-economic information originates from the National Statistical Institute INE. Public access to microdata from administrative data sources and surveys is available through standard anonymization and can be accessed via their website. However, access to confidential microdata is restricted to institutions that justify the need for statistical analyses for scientific purposes in the public interest. Currently, individual-level socio-economic data is not directly linked to patient data in Spain, limiting its availability for research purposes. Coordination between the Health System and the National or Regional Statistics Institute is required to link socio-economic with patient data. This involves submitting data requests to both entities and processing requests to the National or Regional Statistics Institute to create a common pseudonymization. No individual-level socio-economic data were available for the analyses conducted in the use case context. It was commented that access is restricted to certain types of scientific research and organizations by law. The lack of legal ground for data integration and privacy and data protection concerns, and time consuming procedures were identified as barriers in Spain.

In Austria, obtaining individual-level socio-economic data for secondary use and subsequent linkages are virtually impossible due to data governance issues. These challenges are primarily rooted in significant data protection concerns, institutional barriers and the presence of data silos, which hinder the integration and sharing of data across different entities. Austria has a central service that generates several domain-specific person identifiers. This increases data security by limiting the parties involved and the illicit linkage of different data sources. But it also increases the complexity in coordinating the involved institutions whenever linkage is required. These processes are not yet well-practiced and trusted enough to be efficient and exploit the design's full potential. Consequently, these issues limit the ability to conduct detailed studies that require linking socio-economic information with other types of individual-level data.

3.1.3 Online workshop

3.1.3.1 Participant characteristics

Twelve participants from different institutes across Europe (i.e., Sciensano, IACS, DANS – KNAW, Consortium of European Social Science Data Archives – University of Ljubljana – Slovenian Social Science Data Archives or CESSDA – UL – ADP, GÖG, Dutch National Institute for Public Health and the Environment, University of Catania) took part in the workshop. The participants included researchers from diverse fields such as social, health and data science. The diversity of professional backgrounds and institutional affiliations provided a wide range of perspectives for the interactive session.

3.1.3.2 Thematic analysis

Perspectives were collected during the interactive session using the Miro board. The resulting Miro board, including participant notes, is presented in Appendix F (Figure A1). Through a process of analyzing and clustering comments, several key themes could be identified: ‘Data governance’, ‘Data discovery’, ‘Data quality’, ‘Standardization’, ‘Centralization/fragmentation’, and ‘Representativeness’ (see Appendix G, Table A7 for coding and theme assignment).

Data governance

Various data governance challenges, were brought forward, impeding the efficient and effective integration of (sensitive) individual-level socio-economic data (see Figure A1 in Appendix F and Table A7 in Appendix G). As highlighted by participants of the workshop, the diverse ownership of different data sources and siloed information systems makes data integration a laborious and time-consuming task. Additionally, navigating administrative and legal hurdles, and ensuring compliance with anonymization or pseudonymization policies was indicated to add further layers of complexity. Furthermore, it was mentioned that data governance does not always prioritize research use cases.

As such, the participants of the workshop advocated that researchers might benefit from more support structures. The workshop participants proposed standardization of data application processes, clear data retention policies and relevant documentation, availability of trustworthy research infrastructures at national and European level, and establishing maximum response times from institutions as possible solutions to streamline procedures.

Data discovery

The workshop participants identified several barriers to effective data (see Figure A1 in Appendix F and Table A7 in Appendix G). These barriers were mainly related to inadequate documentation and

metadata. They noted a frequent lack of detailed data information (e.g., variable definitions), codebooks, and data collection methods, which complicates understanding what the data represent and its potential limitations. It was commented that this lack of documentation can be challenging, particularly for non-experts in the SSH field, who may struggle to identify adequate variables, indicators, and data sources.

Data quality

The workshop participants highlighted several issues related to the quality of socio-economic data (see Figure A1 in Appendix F and Table A7 in Appendix G). A first concern that was put forward is the *timeliness* of data, with updates of socio-economic data commonly not aligning the timeliness of other data sources, potentially leading to discrepancies that can affect the accuracy and relevance of analyses. Participants suggested leveraging continuously updated administrative data sources. *Accuracy* of the data was put forward as another critical element, particularly concerning self-reported data, which may suffer from biases or inaccuracies. The re-use of existing statistics was noted to come with methodological limitations that could constrain meaningful research outcomes. It was exemplified by a participant that routinely collected taxable income data do not consider the life context of an individual (e.g., the household). Lastly, it was indicated that the level of *granularity* in available data was sometimes inadequate for specific research needs, rendering it less useful for certain inquiries.

Standardization

Workshop participants brought forward standardization as being important in enhancing the usability, accessibility, and comparability of socio-economic data across institutions (see Figure A1 in Appendix F and Table A7 in Appendix G). They highlighted the need for established standardized data formats. This includes adopting Common Data Models (CDMs) and metadata standards to enable interoperability among federated entities. Further, they emphasized the need to standardize data application processes to streamline data access and integration.

Centralization/Fragmentation

Tension between centralization and fragmentation in data management was noted, particularly within the context of aspiring the creation of a consolidated European data ecosystem (see Figure A1 in Appendix F and Table A7 in Appendix G). The idea of a universal secure virtual repository (VR) at the European level was proposed as a potential solution for centralizing data access, potentially enhancing security and streamlining data management processes. However, participants acknowledged the

current state of scattered and fragmented data, where data sources are siloed, often managed by different data owners or holders.

Representativeness

Concerns about the representativeness of socio-economic data used in public health research, and the resulting lack of generalizable research results, were raised (see Figure A1 in Appendix F and Table A7 in Appendix G). It was described that sampling methods may not be aligned with the purposes of public health research.

3.2 Genomic data integration

3.2.1 Theoretical background

A non-exhaustive overview of pathogen genomic data repositories and their characteristics is provided in Appendix H, Table A8.

3.2.2 Field expert consultation: in-depth interviews

3.2.2.1 Participant characteristics

Seven field experts participated in the in-depth interviews, sharing their experiences and perceptions regarding integrating pathogen genomic data with epidemiological and clinical data sources for public health research. The participating stakeholders represented seven different countries within the WHO European region (i.e., Austria, Portugal, France, Belgium, Switzerland, The Netherlands, Croatia), and different fields, including bioinformatics, platform development, epidemiology, health system planning, and surveillance.

3.2.2.2 Thematic analysis

Transcripts from the interviews were analyzed to identify emerging themes. The thematic analysis of the transcripts revealed three key themes, each reflecting different facets of the participants' viewpoints: 'Feasibility', 'Data lifecycle governance' and 'Utility'. A mind map is shown in Figure 1, representing the identified themes, sub-themes, sub-sub-themes and their interrelationships. A complete list of selected text elements, codes and identified themes is presented in Appendix I (Table A9 and A10).

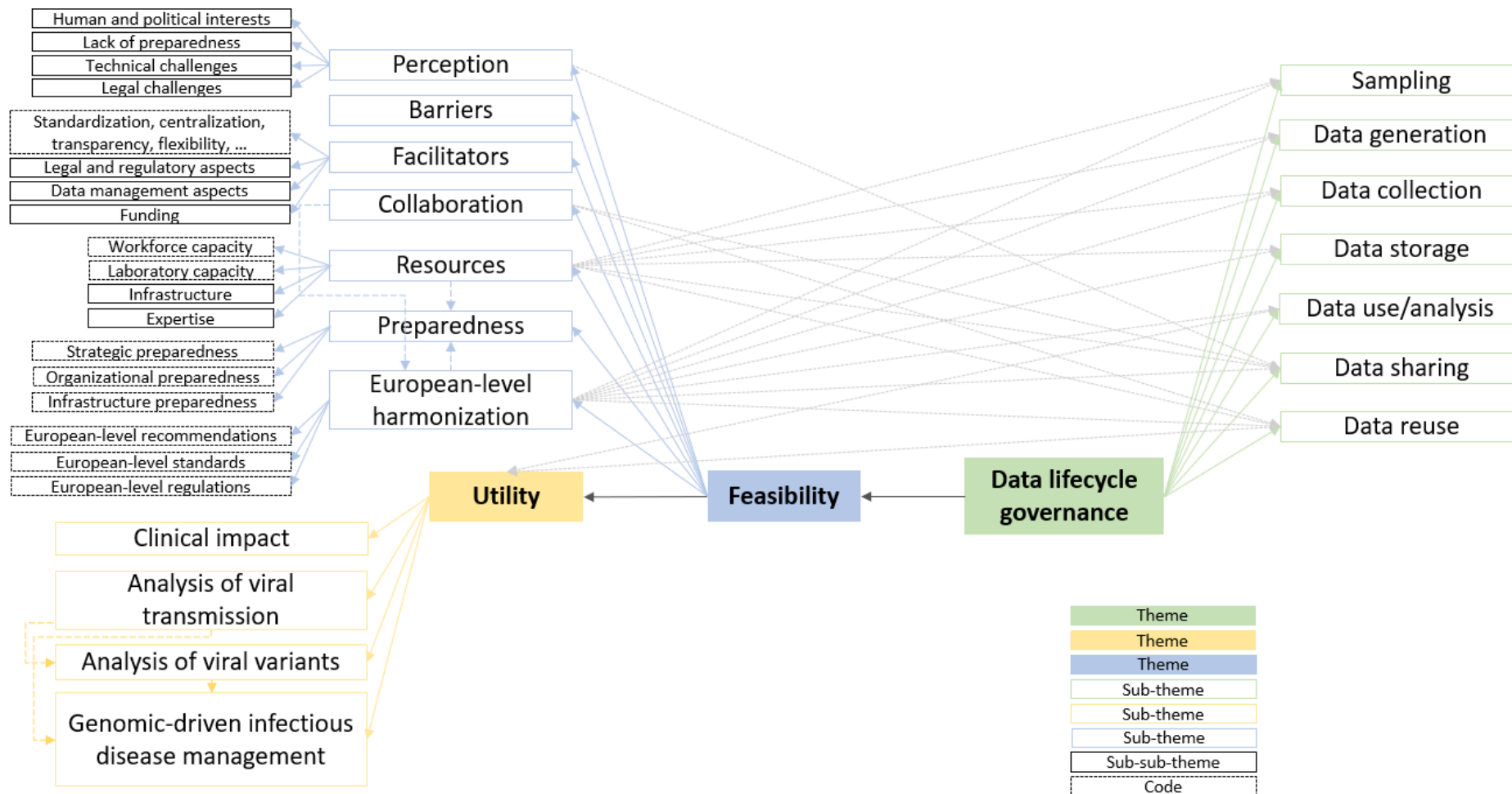


Figure 1. Mind map representing themes, sub-themes, sub-sub-themes, and their interrelationships, as identified from thematic analysis of seven qualitative in-depth interviews regarding the integration of pathogen genomic data with epidemiological and clinical data sources for public health research. The following key themes were identified: ‘Feasibility’ (blue), ‘Utility’ (yellow) and ‘Data lifecycle governance’ (green). For each theme several sub-themes (and potentially sub-sub-themes) could be recognized.

Feasibility

To make the integration of pathogen genomic data with health-related data feasible for public health research, several critical elements were marked and the following sub-themes were identified: ‘Resources’, ‘Preparedness’, ‘European-level harmonization’, ‘Collaboration’, ‘Perception’, ‘Facilitators’, and ‘Barriers’ (see Table A9 and A10 in Appendix I).

Multiple participants highlighted the need for robust and sustainable ‘**Resources**’ at different levels. Investments to develop, improve, and maintain scalable, secure, and user-friendly digital infrastructures were presented as being essential, enabling centralized and standardized data collection, storage, and access. Technological investments and digitalization were brought forward as key drivers in building a robust digital infrastructure that can support the integration and utilization of pathogen genomic data in public health research. The success of these efforts was also mentioned to rely on laboratory and workforce capacity. Participants also emphasized that establishing expertise is crucial to navigate the complexities of this operation successfully. Learning from past experiences, successes and challenges, was an important factor in refining strategies and responses in routine and emergency states.

Besides generating and maintaining a resilient infrastructure (i.e., infrastructure preparedness), interview participants emphasized the need for ‘**Preparedness**’ at other levels, such as organizational and strategic preparedness.

The need for ‘**European-level harmonization**’ at different levels was noted by different interview participants, such as European-level standards (e.g., data standards, metadata standards, standards for data exchange), regulations (e.g., European Health Data Space or EHDS), recommendations and guidelines. Several participants expressed their viewpoint that the standards should be coming from a higher policy level (i.e., European level), hence implying top-down decision-making, to reach large scale harmonization.

The requirement for ‘**Collaboration**’ among a diverse array of stakeholders for the purpose of integrating genomic data was pointed out. This includes collaboration between different research teams, laboratories, departments, and agencies, and collaboration of these stakeholders with public health authorities, clinicians, IT departments, and data owners. Effective collaboration was said to be essential in addressing complex technical and regulatory challenges associated with data integration.

The importance of the ‘**Perception**’ towards the integration of genomic data with health-related data for public health research was pointed out, with participants remarking that building trust among

stakeholders (including researchers, policy-makers, the general public) is essential to allow data integration for research. It was suggested that establishing a positive narrative, where the potential benefits of integration, such as improved disease surveillance and public health interventions, are communicated, could be beneficial.

Different '**Facilitators**' were identified by the interview participants, focusing on legal and regulatory aspects (e.g., emergency provisions, legal obligation for case notification, providing unique identifiers or reporting), data management aspects (e.g., transmittable unique identifiers, limited administration), and human and political interests (e.g., having a sense of purpose, political willingness, lack of competing interests). The need for (sustainable) funding to foresee in the required resources (e.g., infrastructure, workforce) was emphasized by several interview participants.

When asked about '**Barriers**', aspects such as technical or legal challenges, the lack of preparedness, and human and political interests were noted. Compliance anxiety (i.e., the stress that researchers or institutions experience or concerns related to adhering to legal and regulatory requirements) was identified to restrict the integration of these different data sources in a timely manner. Competing research interests were found to hinder effective collaboration and data sharing. Further, the lack of political recognition of the importance of this work was mentioned as an important barrier.

Data lifecycle governance

Different critical aspects were brought up related to elements of the data lifecycle, consisting of 'Sampling and Data generation', 'Data collection', 'Data storage', 'Data use or analysis', 'Data sharing', and 'Data re-use' (sub-themes) (see Table A9 and A10 in Appendix I).

Interview participants made mention of different '**Sampling**' strategies to select clinical samples for sequencing pathogen genomes (e.g., exhaustive, random or targeted sampling, different coverages), with these strategies indicated to potentially change over time and with the availability of resources (e.g., laboratory capacity, funding). The importance of finding a balance between ensuring a high sampling coverage and the cost of sequencing was indicated. Wastewater surveillance was brought up as a complementary strategy to sequencing viral isolates from clinical samples, allowing to gain information about the contextual situation.

It was pointed out that performing sequencing and analysis (i.e., '**Data generation**') in a centralized place is of importance. This centralization is seen as a facilitator of efficiency, reducing the complexities that often arise when tasks are distributed across multiple locations.

Elements related to '**Data collection**' were raised. The ambition towards and importance of collecting high quality (e.g., timely, complete, accurate, reliable) genomic data and metadata was highlighted. However, it was exemplified that technical and administrative challenges (e.g., retrospective updates, validation processes) can complicate data collection and potentially cause overhead, amplify workload, and/or lower the reliability of the data. Concerns related to privacy and data protection in data collection were brought up. Participants acknowledged the importance of data controllers limiting the collection of personal information to the minimum that is directly relevant and necessary to achieve a specific authorized purpose (i.e., according to the data minimization principle). However, they underscored the need for collecting common and unique sample or person identifiers aiding linkage to other heterogeneous data sources and avoiding data silos.

Different participants highlighted the requirement for standardized, centralized and pathogen or disease-independent '**Data storage**'. The need for substantial data storage capacity to store large volumes of genomic data was also highlighted. Further, ethical and legal clearance to host genomic data together with (pseudonymized) unique identifiers and/or sensitive metadata should be available.

The importance of '**Data sharing**' (e.g., of sequencing and genetic variation data, and resulting statistics) to allow re-use and foster transparency, was underscored by different interview participants. Participants mentioned that genomic data generated within their institution were deposited on genomic repositories like GISAID, ENA, and GenBank, and that non-sensitive data, derived statistics, visualizations, and reports were additionally shared on their institutional websites. Emphasis was put on the timeliness of data sharing. While slight delays in the sharing of large raw genomic data files were remarked (through challenges in data transfer), different interview participants communicated a clear commitment to quick data sharing.

Next, aspects related to '**Data re-use**' of real-world data sources were brought up. As the utility of pathogen genomic data on its own for public health research is limited, linkage to individual-level personal host data is often required. It was stressed that when accessing (sensitive) personal data and linking it to (non-personal) pathogen genomic data, obtaining the necessary legal and ethical clearances is imperative to ensure compliance with privacy and data protection regulations. Access to (sensitive) personal data can be obtained by going through thorough application processes. In addition, the importance of establishing a secure infrastructure was highlighted, as well as facilitating safe data linkage and providing controlled access to researchers. This infrastructure would serve as a safeguard against potential breaches while allowing for the effective use of valuable data. In cases where individual-level data is not accessible, it was suggested that regional correlation can be employed as an alternative method to derive meaningful insights. An interview participant acknowledged that the

discoverability of data for re-use could be improved, by enhancing effective data discovery tools, providing the ability to perform advanced queries on databases to better understand the availability of data. This participant expressed a clear aspiration for a centralized portal, enabling researchers to easily access metadata and identify the appropriate contacts for gaining access to specific datasets across Europe.

Utility

Participants underscored the utility of the integration of pathogen genomic data for public health research, enabling the monitoring of viral variants, their transmission dynamics, and their impact on elements such as clinical severity and vaccine effectiveness (see Table A9 and A10 in Appendix I). Overall, studying these subjects was noted to be critical for informing public health responses and allowing pathogen genomic-informed public health practice.

4. DISCUSSION

In this work, we aimed to gather insights (i.e., experiences, perceptions, needs and expectations) from experts across various fields on the integration of socio-economic and pathogen genomic data with other health-related data to address public health research questions and thereby supporting evidence-driven public health policy. To achieve this, we organized a workshop focused on gathering experiences with and perceptions towards the integration of socio-economic data, in parallel to in-depth interviews conducted with several field experts focused on the integration of genomic data for public health research. These data sources, from the SSH or molecular life sciences field, served as illustrations to demonstrate the integration of heterogeneous data sources from diverse domains to generate relevant insights for public health research. The qualitative analyses from workshop participant notes focusing on socio-economic data integration, alongside the in-depth interview transcripts concerning the integration of pathogen genomic data, revealed several overlapping elements. These shared (sub-)themes indicate common challenges and considerations across different domains. By examining the perspectives of experts from both fields, we identified key factors, essential for the successful integration of diverse data sources (as presented hereunder).

4.1 Key themes and recommendations for future preparedness

4.1.1 Data governance

Data governance poses various challenges that impede the efficient and effective integration and analysis of diverse (sensitive) personal data, with each step of the data lifecycle, from collection to re-use, requiring careful consideration.

Given that data sources are often stored decentralized in isolated environments, conducting interdisciplinary causal research, especially across national borders, faces several challenges. Even when following a federated approach, which offers a potential solution by distributing data management responsibilities, issues related to data governance persist within individual nodes, as exemplified by the experiences of Participant Nodes in the BY-COVID use case and commented by workshop and interview participants. The diverse ownership of different data sources and siloed information systems makes data integration laborious and time-consuming.

Navigating administrative and legal hurdles adds a layer of complexity, requiring careful consideration to balance ethical standards and legal obligations with the utility of data for the public good, i.e., protecting and advancing public health. During the collection of personal data, privacy concerns can arise. The principle of data minimization poses the need to limit the collection of personal information to what is strictly necessary and relevant for the purpose [14]. Inadequate data minimization increases the risk of data breaches and misuse. Data minimization commands proper data governance and enhances public trust. However, analysis of interview transcripts revealed the concerns about creating data silo's.

Recommendation 1 In alignment with the data minimization principle, we propose to gather no or minimal personal information alongside pathogen genomic data while ensuring the collection of unique identifiers that enable linkage to other data sources. By limiting data collection to the minimal required, organizations can restrict the amount of data to store and minimize risk of data breaches. Collecting linkable pseudonymized identifiers enables integration of different data sources when required, enabling valuable research without compromising the privacy of individuals. This approach provides a way to balance the need for effective data analysis with the requirement to minimize data collection, safeguarding privacy while supporting the organization's data-driven goals.

As personal data moves to storage, warranting data security becomes highly relevant. Further, hosting data with unique identifiers requires an ethical and legal clearance, and alignment with the storage

minimization principle (i.e., stressing the limitation of storage to the time required to fulfill a specific purpose). Due to ethical and privacy considerations to protect individuals' identity and confidentiality, access to personal data is restricted. Human genomic data are recognized as a particularly sensitive form of personal data by the current privacy regulations in Europe (General Data Protection Regulation or GDPR) [15,16]. Pathogen genomic data collected and stored together with person-associated host information (e.g., health status) and certain socio-economic information like the migration background of an individual, are also considered as sensitive personal data by this legislation [17]. Access to and approval for linkage of different (sensitive) personal data sources to re-use for research purposes (i.e., secondary use) under the GDPR needs to be granted by an authorizing body (typically an ethical or privacy committee), and also needs to comply with the principles of purpose and data minimization [14]. In most countries of the European Union (EU), evaluation of a research protocol, including information on the planned data management procedures, is required to assess the request. However, no standard data access procedure across Europe is available at this moment in time. Participants reported challenges in the workshop and in-depth interviews related to lengthy and demanding data access applications and navigating complex procedures. Additionally, adequate data documentation (e.g., codebooks, data collection methods) and metadata are often lacking or challenging to obtain before going through these burdensome data access applications. The time and resources demanded to maneuver through these complex regulations often slows down research efforts, delaying crucial insights into pressing subjects like outbreak response and treatment development in a public health crisis.

Recommendation 2 Centralized coordination, streamlined procedures, standardization of data access applications at the European level, as envisioned in EHDS legislation [18,19], could accelerate data access and, in time, responses to policy-relevant research questions. Establishing maximum response times for institutions to respond to applications could speed up procedures. Proper documentation of procedures, data content, data collection and data processing activities, enforced by efficient data discovery tools to improve discoverability, would enable researchers to locate data and explore whether they should request access. A centralized European discovery portal is also proposed, allowing researchers to analyze availability across Europe and identify contact points.

Different interview participants highlighted that personal data sharing faces challenges due to differences in the interpretation and implementation of the GDPR across institutions and countries, leading to inconsistent data sharing practices, compliance uncertainties, and uneven compliance burdens. This variability complicates the establishment of standardized protocols for data exchange,

creating legal and administrative barriers. Compliance anxiety, stemming from complex data privacy and security regulations, also considerably contributes to a general cautiousness and reluctance among stakeholders to integrate and share data. Institutions and researchers often face uncertainty about legal boundaries imposed by regulations like the GDPR. The fear of unintentionally violating regulations, and consequential legal repercussion or reputation damage, can result in overly protected data silos and hindered collaboration efforts, impeding the (re-)use of valuable data. Further, data re-use, while valuable for advancing research, brings ethical dilemmas about the original consent that was provided. On the other hand, quickly opening non-personal data (e.g., publication of anonymized pathogen genomic data in repositories like GISAID or ENA) ensures that valuable information is promptly accessible to the broader research community, facilitating immediate re-use and advancing scientific progress. Several interview participants expressed their commitment towards this.

Recommendation 3 Continuing the efforts to achieve a uniform implementation of the GDPR across Europe would eliminate discrepancies, thereby making it easier for organizations and researchers across Europe to adhere to these regulations, and resulting in the equal protection of individuals' privacy rights within the EU. Establishing clear, standardized processes for processing and sharing data, ensures that compliance is built into everyday operations. Clearer guidance from the European Data Protection Board (EDPB, i.e., an independent European body warranting the consistent application the GDPR [20]) and consistent enforcement practices are crucial to achieve this. Further, additional support systems for stakeholders, education (e.g., education of stakeholders on GDPR requirements, clarifying what is allowed and what not), regular reviews and privacy impact assessments (allowing to identify potential issues before they escalate) and leveraging technological tools and infrastructures (e.g., data encryption, anonymization, secure data sharing platforms) can play an important role in reducing fear and uncertainty, and thus in overcoming compliance anxiety. The European Health Data Space (EHDS), of which the creation was approved in April 2024 and seeking to establish a clearer legal structure the secondary use of data [21,22] , can play an important role in these matters.

Finally, despite mandates for data provision, data holders are often ill-prepared for the demands of researchers and may prioritize other tasks, particularly after the urgency of the pandemic wanes. This highlights the need for greater alignment between governance practices and research needs. Budget constraints exacerbate these challenges, as data holders often lack the resources to facilitate data requests, leaving researchers struggling to access data from official statistical offices and other data sources.

Recommendation 4 In cases where individual-level data is not accessible due to privacy concerns or other restrictions, it was suggested that regional correlation could be employed as an alternative method to derive meaningful insights. Although this approach might introduce some biases into the research, it allows researchers to draw significant conclusions without compromising individual privacy, thereby expanding the possibilities for data analysis while maintaining ethical standards.

4.1.2 Data quality and representativeness

The quality and representativeness of data in sources utilized in these data integrations for public health research remain critical factors that can significantly influence the validity, reliability, and generalizability of research findings.

The timeliness of data is critical in these settings, as it influences the ability to directly respond to health threats. For instance, surveys allow collecting specific socio-economic information required for constructing socio-economic indicators in a standardized comparable way. However, data coming from surveys are likely not updated regularly enough to provide current information and to match the timeliness of other registries, reducing their utility for policy research.

Recommendation 5 We propose using continuously updated administrative data sources to enhance the timeliness and relevance of socio-economic information. Besides of their timeliness, administrative data benefit from greater coverage and acquisition at a lower cost. However, administrative data might not be collected in a standardized way and/or might not contain all information required to adequately build the indicators of interest.

Comparably, delays in collecting, processing, and sharing pathogen genomic data can lead to outdated information, which may not reflect ongoing shifts in pathogen evolution and public health threats. Ensuring the timeliness of pathogen genomic data, through streamlined workflows and efficient data-sharing mechanisms is essential for its effective integration in real-time public health decision-making.

Further, there is an important risk of bias and misinterpretation when data sets are incomplete or not representative of the broader target population. For example, pathogen genomic surveillance is generally non-exhaustive, relying on isolates obtained from clinical testing – which are potentially not representative due to segments of the population potentially having limited access to the healthcare system, testing bias, or asymptomatic infections [23]. Therefore, obtaining representative genomic data can be challenging, hampering epidemiological and clinical interpretation of the obtained results [23]. Concurrently, missing data can intrude at various stages, from sample collection to sequencing

and data processing, due to technical limitations or resource constraints [24]. Ensuring data completeness requires detailed data management practices to capture all information required to obtain accurate insights for public health practice. In addition, ensuring comprehensive sample coverage is essential to accurately reflect the diversity and distribution of pathogens within the population, limiting the chance to overlook emerging or rare variants. However, it was mentioned that it is essential to find a balance between the cost of genomic surveillance and the obtained societal benefits (i.e., cost-benefit analysis). A study in Switzerland, a country that during the COVID-19 demonstrated a highly comprehensive surveillance system, showed that lowering the sampling up to one-third of the actual effort did not strongly influence certain surveillance outcomes like the first detection of Variants Of Concern (VOCs) [25]. However, other surveillance outcomes were affected and the results were found to be dependent on the lineage of circulation. They highlighted the need for predefining targets for genomic surveillance at a national level based on which sequencing coverage and efforts [25].

As a contrast to pathogen genomic surveillance through clinical samples, wastewater surveillance offers a valuable approach to capture genetic material from a wide and diverse population, including individuals who may not seek medical attention [26]. This makes it an effective early warning system for detecting circulating pathogens and variants, providing timely data that can inform public health responses. Additionally, it offers a cost-effective and scalable method for continuous monitoring, particularly in resource-limited settings or during periods of reduced testing capacity [26]. However, it is important to point out the limitations of wastewater surveillance, such as the lack of individual-level information (making it impossible to link the findings directly to specific individuals) and elements affecting the accuracy of the results [26,27].

These challenges emphasize the need to carefully consider data quality aspects in research planning and execution. Data validation can play a key role here, involving checks and procedures to verify the quality of collected data.

4.1.3 Standardization

Another considerable challenge in the secondary use of diverse real-world data coming from different sources, is the lack of standardization (e.g., use of different formats and standards), especially for research crossing multiple countries or regions. Standardization efforts are essential to enable seamless integration and analysis across systems, institutions, and sources. In the BY-COVID use case, harmonization across Participant Nodes for federated analysis was obtained by specifying crosswalks when defining data requirements in a CDM, mapping variable categories (or values) to different classification systems. However, further harmonization in data collection and data processing across

institutional, regional or national borders, ideally following international (European) standards, would greatly improve the consistency, comparability and reliability of studies. Standardization of procedures (e.g., for data sharing, data access, data security) would additionally facilitate timely and effective re-use of data.

Recommendation 6 CDMs provide a structured framework for organizing and representing real-world data consistently across different sources, ensuring that data attributes, relationships and semantics are uniformly defined and understood. Bespoke CDMs can be specified specifically for a research question (building on a specific causal model if aiming to approach causal inference) as in the BY-COVID use case [28], or a priori defined CDMs, like the OMOP CDM [29,30], can be provided as a community standard, designed to standardize the general content and structure of observational data. Mapping data to a CMD allows for interoperable analytics in different sites complying to the specifications. Meanwhile, metadata standards can define descriptive details of data regarding the organization, content and background of data, encompassing definitions, formats, and conventions. Following metadata standards for describing real-world data enables re-use by increasing findability [30,31].

Recommendation 7 Standardized data storage ensures that data, such as sequence data, is stored in consistent formats, facilitating interoperability and collaboration across different research institutions and countries.

Recommendation 8 Decisions on SES indicators could be made on a European level to determine the variables most crucial for research purposes. Establishing a common set of SES indicators across federated entities would help streamline data collection processes, ensure data quality, facilitate comparability, and data integration. Such an approach supports data minimization principles, limiting data collection to what is necessary. This could be accomplished by defining a set of ‘target’ variables required to answer research questions of interest (i.e., agreeing on a minimal dataset) and that are universally recognized and accepted within the research community. For this, researchers and domain experts need to collaborate to identify SES indicators that capture key dimensions of socio-economic status and that have been widely used and validated in previous studies, demonstrating their relevance, reliability, and validity in measuring socio-economic status across diverse populations and contexts. Collaborative projects involving interdisciplinary teams and perspectives from different countries can provide valuable insights, recognizing that SES factors may vary in relevance across different cultural contexts. Expanding the skill set of the research team to encompass a

broader understanding of SES indicators and cross-cultural perspectives is essential for effective data integration and analysis.

4.1.4 Centralization/fragmentation

Centralization of data and resources within an consolidated system or platform can enhance consistency and accessibility of data, and simplifies coordination efforts. Centralization facilitates standardization, collaboration, and the implementation of broad comparative studies with generalizable findings.

Recommendation 9 Regarding genomic surveillance, institutes can ensure better coordination, quicker turnaround times, and more consistent quality control by housing both sequencing and analysis under one roof. This approach enhances operational efficiency and promotes a more cohesive and integrated approach to research, allowing for more robust and reliable outcomes.

Recommendation 10 Centralized data storage systems are critical for consolidating genomic data from multiple sources, enabling more efficient analysis and comparison essential for tracking disease outbreaks and identifying genetic variants.

While centralization offers benefits like improved data accessibility and consistency, it must be balanced with considerations for privacy and data security, the autonomy of data holders, and the unique requirements of different datasets. This balance between centralization and fragmentation is crucial for fostering collaboration while maintaining the flexibility needed to address diverse research and policy needs across the EU.

4.2 Requirements for effective integration of heterogeneous data sources for public health research

The effective integration of heterogeneous data sources necessitates a robust system that addresses diverse challenges related to data governance, quality, representativeness, and interoperability. Efforts towards building and maintaining these systems should ideally be put in during periods without an urgent crisis, establishing preparedness to respond rapidly when emergency states arise.

System preparedness relies on different levels of capacity, such as robust technological infrastructure (i.e., capable of managing diverse data types, equipped with interoperability standards, including scalable storage solutions, efficient data management systems, high-performance computing resources), physical infrastructure (e.g., laboratories), standardized protocols or working procedures

(e.g., for data collection, data sharing), workforce capacity (e.g., sufficient staff, skilled in fields like laboratory technology, bioinformatics, public health surveillance, data science, or legal aspects of data privacy/security), and support systems. Capacity must be sufficient to maintain routine operations and scalable to support surge demands during health crises. Previous capacity building and establishing expertise provides a foundation for the development of robust methodologies and protocols. The integration process also benefits from diverse involved parties' shared expertise and resources.

Interdisciplinary collaborative networks or partnerships among data scientists, domain experts, and policy-makers at regional, national, European and broader levels are also crucial to mobilize data, contextualize and interpret the integrated data effectively, and to build trust across parties. However, these collaborations are scarce outside the context of national and European projects, usually far from the centers of decision, thus decelerating the responses to urgent public health threats. Further, competing (research) interests were repeatedly mentioned to obstruct collaborative efforts. Setting up these robust collaborative relationships and overcoming these barriers, will facilitate effective and timely data integration to response to emerging policy-relevant research questions.

These efforts and resources largely rely on the availability of sustainable funding. However, a balance between costs and benefits for society should be found, thereby highlighting the importance of the scalability of systems. Efforts towards building and maintaining these systems at a baseline level should ideally be put in during periods without an urgent crisis, establishing preparedness to respond rapidly when emergency states arise. Regular evaluations and updates of infrastructures to adopt emerging technologies and standards, is required to maintain the capacity of the system to handle the complexity and scale of heterogeneous data sources that can be effectively integrated. Further, policy support is crucial to facilitate easy data integration for research purposes. Without strong political support, necessary resources, legislative changes, and cross-sector collaboration are often insufficient, hindering progress. There is a need for a top-down approach, where higher instances (e.g., the European Commission, the World Health Organization or WHO) provide guidance, recommendations, a legal framework, regulations and high-level standards.

By ensuring system preparedness, heterogeneous data sources can be leveraged more efficiently to generate actionable insights, and supporting public health policy. Addressing these elements is crucial for the effective re-use of integrated data for advancing public health research (within and across borders) and developing effective public health interventions.

4.3 Strengths and limitations of the study

In this qualitative study, participants of both the workshop and in-depth interviews were distributed across different sites in the WHO European region and cover diverse fields of expertise. Thematic analysis was employed to analyze transcripts from in-depth interviews. The well-considered selection of participants and the use of a thematic analysis approach resulted in rich and meaningful participant insights and a nuanced understanding of the topic. Thereby, the robustness and insightfulness of the findings was enhanced. The approach enabled to develop a thorough understanding of the subjective meanings that participants assign to their experiences.

However, the number of participants was limited in both initiatives and the representativeness of the participants was not assessed, therefore limiting the generalizability of the obtained observations. Certain fields of expertise, such as the legal field, were not represented among the interview participants. Further research should address these limitations by incorporating a larger and more diverse set of participants, explicitly including experts from the legal domain. Additionally, the subjective, interpretative nature of thematic analyses can introduce bias, and affect the study's reproducibility. There is a risk that subtle but significant themes are overlooked.

5. CONCLUSION

This work gathers and analyses insight from diverse field experts on integrating socio-economic and genomic data with health-related data to address public health research questions. Insights were gathered through an online workshop and in-depth interviews, with qualitative analyses revealing common challenges, concerns, and potential ways forward among participants, such as elements related to data governance, data quality, representativeness, and standardization. This convergence of insights highlights the potential for developing unified strategies that address the complexities inherent in combining heterogeneous datasets for public health research, laying the groundwork for future implementation.

6. ACKNOWLEDGEMENTS

This work was conducted within the context of the European BeYond-COVID project. The author would like to acknowledge members contributing to the workshop organization and/or interview development: Nina Van Goethem (Sciensano), Simon Saldner (DANS – KNAW), Vasso Kalaitzi (DANS – KNAW), Francisco Estupiñán-Romero (IACS), Enrique Bernal-Delgado (IACS), Iris Van Dam (Sciensano).

7. REFERENCES

1. Meurisse M, Estupiñán-Romero F, Perola M, Paaanen T, González-Galindo J, Van Goethem N, et al. Real-World Comparative Effectiveness of SARS-CoV-2 Primary Vaccination Campaigns Against SARS-CoV-2 Infections: A Federated Observational Study Emulating a Target Trial in Three Nations. Rochester, NY; 2024. doi:10.2139/ssrn.4869107
2. Meurisse M, Estupiñán-Romero F, González-Galindo J, Martínez-Lizaga N, Royo-Sierra S, Saldner S, et al. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. *BMC Medical Research Methodology*. 2023;23: 248.
3. Lohse S, Canali S. Follow *the* science? On the marginal role of the social sciences in the COVID-19 pandemic. *Euro Jnl Phil Sci*. 2021;11(4): 99.
4. Adler NE, Ostrove JM. Socioeconomic Status and Health: What We Know and What We Don't. *Annals of the New York Academy of Sciences*. 1999;896:3–15.
5. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol*. 2014;15:538.
6. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen Genomics in Public Health. *New England Journal of Medicine*. 2019;381(26):2569–2580.
7. Roberts MC, Fohner AE, Landry L, Olstad DL, Smit AK, Turbitt E, et al. Advancing precision public health using human genomics: examples from the field and future research opportunities. *Genome Med*. 2021;13:97.
8. Van Goethem N, Meurisse M, Estupiñán-Romero F, Bernal-Delgado E, Saldner S, Kalaitzi V. BY-COVID - WP5 - Baseline Use Case: Integration of socioeconomic data in observational studies using a prototyped workflow standard to federated population health research (1.0.0) [Internet]. Zenodo. 2024 [cited 23 Aug 2024]. Available: doi:10.5281/zenodo.12771703
9. BY-COVID Spring 2024 Baseline Use Case Workshop: Integration of individual-level socioeconomic data for infectious diseases research and prevention in Europe [Internet]. [cited 11 Jul 2024]. Available: <https://by-covid.org/news-events/spring-2024-baseline-usecase-workshop/>
10. Meurisse M, Saldner S, Van Goethem N, Kalaitzi V, Estupiñán-Romero F, Bernal-Delgado E. BY-COVID Spring 24 Baseline Use Case Workshop [Internet]. Zenodo. 2024. Available: doi:10.5281/zenodo.12168494
11. Naeem M, Ozuem W, Howell K, Ranfagni S. A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods*. 2023;22.
12. Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 2006;3(2):77–101.
13. Sciensano. Linking of registers for COVID-19 vaccine surveillance. In: [sciensano.be](https://www.sciensano.be/en/projects/linking-registers-covid-19-vaccine-surveillance) [Internet]. [cited 25 Nov 2021]. Available: <https://www.sciensano.be/en/projects/linking-registers-covid-19-vaccine-surveillance>

14. Biega AJ, Potash P, Daumé H, Diaz F, Finck M. Operationalizing the Legal Principle of Data Minimization for Personalization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020; 399–408.
15. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur J Hum Genet*. 2018;26:149–156.
16. Sariyar M, Suhr S, Schlünder I. How Sensitive Is Genetic Data? *Biopreservation and Biobanking*. 2017;15(6):494–501.
17. What personal data is considered sensitive? - European Commission [Internet]. [cited 13 Aug 2024]. Available: https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en
18. Hussein R, Balaur (Roznovat) I, Burmann A, Ćwiek-Kupczyńska H, Gadiya Y, Ghosh S, et al. Getting Ready for the European Health Data Space (EHDS): The Iderha's Plan to Align with the EHDS Requirements for the Secondary Use of Health Data. Open Research Europe; 2024. Available: 10.2139/ssrn.4755396
19. Marin S, Haaramo P. D7.1: Common data access application form [Internet]. 2023 Dec. Available: https://ehds2pilot.eu/wp-content/uploads/2024/04/HealthData@EU-Pilot-Deliverable-7.1_FIN.docx-1-1.pdf
20. European Data Protection Board. The European Data Protection Board [Internet]. [cited 15 Aug 2024]. Available: https://www.edpb.europa.eu/about-edpb/who-we-are/european-data-protection-board_en
21. European Commission. European Health Data Space [Internet]. 24 Apr 2024 [cited 15 Aug 2024]. Available: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
22. European Commission. Commission welcomes European Parliament's adoption of the European Health Data Space and regulation on substances of human origin [Internet]. [cited 15 Aug 2024]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_24_2250
23. Meurisse M, Van Oyen H, Blot K, Catteau L, Serrien B, Klamer S, et al. Evaluating methodological approaches to assess the severity of infection with SARS-CoV-2 variants: scoping review and applications on Belgian COVID-19 data. *BMC Infectious Diseases*. 2022;22(1): 839.
24. Agiwal V, Chaudhuri S. Methods and Implications of Addressing Missing Data in Health-care Research. *Current Medical Issues*. 2024;22(1):60-62.
25. Wegner F, Cabrera-Gil B, Tanguy A, Beckmann C, Beerenwinkel N, Bertelli C, et al. How much should we sequence? An analysis of the Swiss SARS-CoV-2 surveillance effort. *Microbiology Spectrum*. 2024;12(5):e03628-23.
26. Tiwari A, Radu E, Kreuzinger N, Ahmed W, Pitkänen T. Key considerations for pathogen surveillance in wastewater. *Science of The Total Environment*. 2024;945:173862.
27. Diamond MB, Keshaviah A, Bento AI, Conroy-Ben O, Driver EM, Ensor KB, et al. Wastewater surveillance of pathogens can inform public health responses. *Nat Med*. 2022;28:1992–1995.

28. Estupiñán-Romero F, Van Goethem N, Meurisse M, González-Galindo J, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Common Data Model Specification [Internet]. 2023 [cited 22 Feb 2023]. Available: doi:10.5281/zenodo.7572373
29. OHDSI CDM Working Group. OMOP Common Data Model [Internet]. [cited 20 Aug 2024]. Available: <https://ohdsi.github.io/CommonDataModel/>
30. Hughes N, Kalra D. Data Standards and Platform Interoperability. In: He W, Fang Y, Wang H, editors. Real-World Evidence in Medical Product Development. Cham: Springer International Publishing; 2023. pp. 79–107. doi:10.1007/978-3-031-26328-6_6
31. Schriml LM, Chuvochina M, Davies N, Eloie-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci Data*. 2020;7:188.

8. APPENDICES

Appendix A – Interview guide

Appendix B – Participant information sheet

Appendix C – Informed consent form

Appendix D – Individual-level indicators of socio-economic status

Appendix E – Mapping exercise of the local availability of individual-level socio-economic data among the BY-COVID use case Participant Nodes

Appendix F – Workshop interactive session Miro board

Appendix G – Thematic analysis of workshop participant notes

Appendix H – Theoretical background

Appendix I – Thematic analysis interview transcripts

Appendix A – Interview guide

Introduction

- Shortly introduce yourself (e.g. where do you work, how you are involved in the project).
- Thank the stakeholder for participating.
- Explain the format and estimated duration of the interview.
 - *“This is a semi-structured interview, with open-ended questions, and will take about 45 minutes to 1 hour.”*
- Explain informed consent and ask the stakeholder to sign.
 - *“Did you have a chance to read the information sheet and informed consent form in advance?”*
 - If not, explain the research project.
 - *“I will quickly go over the main points of the informed consent form. To start, your participation is voluntary and you are free to withdraw your participation or any part of your response up to 30 days after the interview date, without having to give a reason or without being penalized or disadvantaged in any way. The interview will be recorded, but the recorder can be paused at any time during the interview. The recording and transcription will be stored locally on a secured file server at Sciensano. Unless specifically requested otherwise (confirmed in writing), any identifying information (e.g., names and characteristics) will be removed from the transcripts. Non-anonymized data, such as the audio and video recordings and a completed informed consent form, are collected as part of the research project. Audio and video recordings will be removed after the interview is transcribed, the latest at 6 months after the interview. Results will only be described and reported at an aggregated level.”*
 - *“Do you still have any questions? Is anything unclear?”*
- Before proceeding, ask if there are any remaining questions regarding the research project, the interview or confidentiality.
- Describe the structure of the interview.
 - *“The interview will consist of some opening questions on your background, some questions related to your own experience with linking genomic data to other data sources for public health research and the facilitators, challenges and unmet needs you experienced. Next, some questions about your prospects related to this type of work and the use of openly available pathogen genomic data in general will follow.”*

Ask permission to turn on the recorder.

Section I: Opening Questions

1. *“Can you describe yourself within the project/organization?”*
2. *“Can you describe your experience with genomics, epidemiology and/or public health research?”*
3. *“What data sources have you used in your research?”*

Section II: Main questions

Experiences

4. *“Can you give an **example of a research question or project** you worked on linking SARS-CoV-2 genomic data with epidemiological and clinical data sources?”*

If the interviewee has no experience, skip to the next main section.

5. *“Which **data sources** were linked in this [research or project]?”*
 1. *“Were these data sources routinely available (i.e., already existing) or were they collected specifically for this [research or project] (or both)?”*
6. *“Can you tell me a bit more about the **SARS-CoV-2 genomic data** that were integrated in this [research or project]?”*
 1. *“What was the source of this information/these data?” (e.g., a database, registry)*
 2. *“What was the sampling strategy for obtaining the SARS-CoV-2 genomic data?”*
7. *“Can you tell me something about the **availability** of the SARS-CoV-2 genomic data that were integrated in the [research or project]?”*
 1. *“What steps were taken to access the SARS-CoV-2 genomic data?”*
 2. *“How would other researchers be able to access and link these SARS-CoV-2 genomic data?”*
8. *“In the [research or project], how did you go about **linking** the SARS-CoV-2 genomic data with the other data sources?”*
 1. *“Which identifier was used for linking the genomic SARS-CoV-2 data to the other data sources?”*
 2. *“What type of SARS-CoV-2 genomic information was integrated?”*
9. *“Which aspects were **facilitating** this research?”*
10. *“What **challenges** related to genomic data linkage did you experience when conducting this research?”*

11. *“Can you think of any specific **unmet needs** you experienced when [doing this research/working on this project]? (What)*

1. *“What was the origin of these unmet needs?”*
2. *“How do you think these unmet needs can be addressed?”*

Prospects

12. *“Can you think of some practical examples of where the analysis of these linked data (i.e., SARS-CoV-2 genomic data linked with epidemiological and clinical data sources) can be relevant for public health policy?”*

Re-use of sensitive data

13. *“In an ideal scenario, how should SARS-CoV-2 genomic data be **collected** to be useful for public health research?”*

1. *“What kind of sampling strategy should ideally be used for SARS-CoV-2 genomic surveillance?”*
2. *“In practice, how do you think sampling bias could be limited?”*
3. *“Which genotyping method do you consider to be most useful, considering costs and benefits?”*

14. *“Keeping this same ideal scenario in mind, how should SARS-CoV-2 genomic data be **linked** with other data sources?”*

15. *“Can you think of any challenges specific to research that is linking SARS-CoV-2 genomic data to other data sources and is shifting from public health research at the national/regional to a **cross-border level**?”*

1. *“What would be the requirements to conduct this type of research?”*

Re-use of open data

16. *“Which open source data repository(ies) for genomic SARS-CoV-2 data do you know and/or work with?”*

17. *“What do you consider as the benefits of having these open SARS-CoV-2 genomic data available?”*

1. *“Can you think of some practical examples of how you could use these openly available SARS-CoV-2 data for public health research?”*

18. *“In what way could an integration of openly available SARS-CoV-2 genomic data with epidemiological and clinical data sources be executed?”*

1. *“Can you think of issues in relation to linking openly available SARS-CoV-2 genomic data to other data sources for public health research?”*
19. *“What requirements should openly available SARS-CoV-2 genomic data ideally comply with to facilitate easy linkage to epidemiological and clinical data sources and be useful for public health research?”*
 1. *“With which corresponding information (metadata) should SARS-CoV-2 genomic data be shared, in repositories such as GISAID or ENA, to facilitate linkage to other data sources?”*
 2. *“Could these requirements have consequences on the (open) accessibility of the data?”*
20. *“With which frequency should SARS-CoV-2 genomic data be shared to allow for timely public health research feeding into policy?”*

Section III: Closing questions

21. *“In your experience, is there anything else important related to the integration of genomic data for public health research that we have not yet touched upon?”*
22. *“Can you think of anyone else who is working in this domain and might be interested to participate in the study?”*
23. *“Do you have my contact details, if any additional questions would come up later?”*

Appendix B – Participant information sheet

Integration of SARS-CoV-2 genomic data with epidemiological and clinical data sources for public health research

We would like to invite you to participate in a research project. Your participation is voluntary; choosing not to take part will not penalize or disadvantage you in any way.

Before you decide whether you want to take part, it is important for you to understand why the research is being executed and what your participation will involve. Please take your time to read the following information carefully and discuss it with others if you wish. Please contact us if anything is not clear or if you would like to receive more information.

The research project

This work is executed in the context of the European BY-COVID project, launched in October 2021. This project aims to connect pre-existing and emerging data sources, and to provide access to linked data across national borders via the COVID-19 data portal. This project aims to drive the mobilization of FAIR (findable, accessible, interoperable, reusable) data across Europe and aims to enable answering research questions to inform public health policies. It is also aimed to move beyond the COVID-19 pandemic, generating a preparedness for potential future pandemics.

In one of the Use Cases, which have been defined within WP5 of the BY-COVID project and are led by **Sciensano** (the National Institute for Health in Belgium) and **IACS** (the Health Sciences Institute in Aragon, Spain), we have described a **methodological framework** to re-use sensitive population, health and care data distributed across national borders in a privacy-preserving and interoperable way to address federated cross-national policy-relevant causal research questions. We are subsequently **demonstrating this framework** by investigating *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*, with the purpose of evaluating the effectiveness of the primary vaccination campaign.

In this context, we came across a **gap in information**, required to investigate one of the additionally formulated secondary analyses, namely, *the evaluation of effectiveness in preventing infection with a specific SARS-CoV-2 variant*, in order to detect potential heterogeneous effects. We noticed difficulties in linking SARS-CoV-2 genomic data to the other data sources required to conduct this secondary analysis.

Here, we will conduct **in-depth interviews** in order to gather experiences (e.g., in terms of facilitators and barriers), perceptions (e.g., in terms of the feasibility or the utility for public health practices),

requirements (e.g., in terms of data standards, repositories, or infrastructure) and expectations (e.g., in terms of timeliness) on the integration of SARS-CoV-2 genomic data with epidemiological and clinical data sources for public health research in the WHO European region. We aim to identify minimal requirements for successful data integration and translate the results into recommendations and action points towards future preparedness.

Why have you been invited to take part?

You have been invited to participate based on your experience and/or expertise relevant to the topic under study. Your participation would be valuable to reach the objectives of the study.

Am I obliged to take part in the interview?

Your participation is voluntary. Before the start of the interview you will be asked to sign an informed consent form. You are free to withdraw your participation or any part of your response up to 30 days after the interview date, without having to give a reason or without being penalized or disadvantaged in any way. If you wish to withdraw, you can contact Marjan Meurisse (Marjan.Meurisse@sciensano.be). No financial benefits or coverage of expenses can be offered.

What will taking part involve?

We will conduct a semi-structured interview. You will be asked to sign an informed consent form before the start of the interview. The interview will take around one hour, however, the exact duration can depend on the interviewee. The interview will be conducted online, via a Webex video call or in person when possible. With your permission, the audio and video of the interview will be recorded. The recorder can be paused at any time during the interview.

What will happen to the information?

The recording will be used to transcribe the interview. To ensure confidentiality, the recording and transcription will be stored locally on a secured file server at Sciensano (ICT services at Sciensano are experts in the handling of sensitive personal data). The folder in which the files will be saved has restricted access, limited to researchers from the “Public health and Genome” unit within Sciensano (i.e., 3 researchers). Unless specifically requested otherwise (confirmed in writing), any identifying information (e.g., names and characteristics) will be removed from the transcripts. Sections of the transcript text may be looked at by responsible individuals from Sciensano and IACS. Non-anonymized data, such as the audio and video recordings and a completed informed consent form, are collected as part of the research project. Audio and video recordings will be deleted after the interview is transcribed, the latest at 6 months after the interview.

What will happen to the results of the study?

Results from the in-depth interviews will be summarized and translated into recommendations and concrete action points to facilitate the integration of pathogen genomic data with epidemiological and clinical data for public health research. Results of this work will feed into a report, serving as a deliverable of the BY-COVID project. Further, we aim to compile the results of this work in a scientific publication. The results might be presented at an international conference. Results will only be described and reported at an aggregated level and in a anonymity-preserving manner (unless specifically requested otherwise, to be confirmed in writing).

Who should you contact for further information?

Any remaining questions or requests for additional information can be directed to Marjan Meurisse, which can be reached at the following email address: Marjan.Meurisse@sciensano.be.

Thank you for taking the time to read and consider this information

Appendix C – Informed consent form

INFORMED CONSENT FORM

Title of project: Integration of genomic data with epidemiological and clinical data sources for public health research

Name of the participant:

Please initial box

1. I confirm that I have read and understand the participant information sheet dated for the above study. I may keep this information sheet for my records and I have had the opportunity to ask questions which have been answered fully. ☐
2. I understand that my participation is voluntary and I am free to withdraw my participation or any part of my response up to 30 days after the interview, without giving any reason and without being penalized or disadvantaged in any way. ☐
3. I understand that sections of my recorded comments and transcript text may be looked at by responsible individuals from Sciensano and IACS. I give permission for these individuals to access this data as relevant to this research ☐
4. I am willing to have this interview audio and video recorded. ☐
5. I understand that the audio and video recordings of my interview will be deleted after the interview is transcribed, the latest at 6 months after the interview. ☐
6. I understand that this consent form will be collected as part of the research project, but will be kept separate from the data. ☐
7. I understand that, unless specifically requested otherwise (to be confirmed in writing), the researchers will maintain my anonymity throughout the project, including in publication. ☐
7. I understand that the transcript data will be kept to enable building on this work in future projects. ☐
8. I agree to take part in the above study. ☐

Name of Participant
(Printed)

Date

Signature

Name of Researcher
(Printed)

Date

Signature

1 copy for subject; 1 copy for researcher

Appendix D – Individual-level indicators of socio-economic status**Table A1.** Individual-level indicators of socio-economic status.

Aspect	Indicator	Encoding(s)	Description	Potential data sources
Occupation	Economic activity	ECONPO9	A person's economic activity reveals whether the person is in paid employment, and if not employed, their employment search efforts. E.g. economically active (full time employee, part time employee, ...) <i>versus</i> economically inactive (seeking employment, retired, student, long-term sick, ...)	CeLSIUS ONS LS data*
	Employment status	EMPST9	A person's employment status indicates the type of work arrangement of the person. E.g. employee, self-employed, or self-employed and employing others	CeLSIUS ONS LS data*
	Employment relations and conditions of occupation	NS-SEC	A measure of a person's employment relations and conditions of occupations. E.g. higher or lower managerial/administrative/professional occupations, intermediate occupations, small employers and own account workers, ...	CeLSIUS ONS LS data*
	Type of business and occupation	International Standard Industrial Classification of all economic activities (ISIC)	Classification according to individuals' economic/productive activity documents industrial sectors, global level.	ILOSTAT**, National statistical offices***
		Statistical classification of economic activities in the European Community (NACE)	Classification of economic activities, derived from ISIC (more detailed than ISIC at lower levels, but the same highest level classification), European level.	ILOSTAT**, National statistical offices***, Eurostat LFS or EU-SILC data****
		National versions of NACE	Classification of economic activities, national level.	National statistical offices***
		International Standard Classification	Classification of occupations	ILOSTAT**, National statistical offices***, Eurostat LFS or EU-SILC data****

		of Occupations (ISCO)		
		Standard Occupational Classification (SOC)	Classification of occupations	US Bureau of Labour Statistics*****
Income or wealth	Personal income	-	Individual net taxable income, indicator of material resources	Eurostat EU-SILC data****
	Household income	-	Household net taxable income, indicator of material resources	Eurostat EU-SILC data****
	Equivalized income	-	Indicator of material resources adjusts household income for family size and its associated costs of living	Eurostat EU-SILC data****
	Car ownership	-	Indicator of material resources can be used when a direct measure of income is not available	Eurostat EU-SILC data****
Education	Years of education	-	Continuous variable	National statistical offices ***,
	Classification of education	International Standard Classification of Education (ISCED)	A global classification system of educational programs by using levels and fields of education. E.g. early/pre-childhood education (ISCED0), primary education (ISCED1), lower secondary education (ISCED2), upper secondary education (ISCED3), ...	National statistical offices ***, Eurostat LFS or EU-SILC data ****
Housing	Housing tenure	-	Whether a household in a private dwelling rents, owns, or holds that dwelling in a family trust, and whether payment is made by the household for the right to reside in that dwelling. E.g. dwelling owned or partially owned, dwelling not owned and not held in a family trust, ...	National statistical offices ***, Eurostat EU-SILC data ****

* CeLSIUS: Centre for Longitudinal Study Information and User Support. LS: Longitudinal Study. Access to data can be requested by accredited researchers through the submission of the required application documents: <https://www.ucl.ac.uk/epidemiology-health-care/research/epidemiology-and-public-health/research/health-and-social-surveys-research-group/studies-14>

** Aggregated statistics can be obtained from the ILOSTAT catalogue: <https://ilostat.ilo.org/data/>. ILO does not provide access to microdata, these requests should be directed to the relevant national statistical offices.

*** Each European country has its own statistical body, such as Statbel (Belgium), ONS (UK), or Destatis (Germany), which collect and publish detailed socio-economic data.

**** LFS: Labour Force Survey. EU-SILC: European Union Statistics on Income and Living Conditions. Access to microdata can be requested at Eurostat by recognized research entities through the submission of one or

multiple research proposals:

[https://ec.europa.eu/eurostat/documents/203647/771732/How to apply for microdata access.pdf](https://ec.europa.eu/eurostat/documents/203647/771732/How_to_apply_for_microdata_access.pdf) .

***** Access to microdata can be requested by eligible researchers at the US Bureau of Labour Statistics:

<https://www.bls.gov/rda/home.htm>

Table A2. Area-level indicators of socioeconomic status.

Indicator	Based on/derived from	Description
Townsend Deprivation Index	Unemployment, household overcrowding, car ownership, home ownership	Measure considering both social and material deprivation
Carstairs (and Morris) Deprivation Index	(Male) unemployment, household overcrowding, car ownership, social class	Measure considering both social and material deprivation
Underprivileged Area (UPA) Score, Jarman Index of Social Disadvantage	Elderly living alone, children under five years old, single parent families, unskilled head of the family, unemployment, overcrowding, mobility (changing address), ethnic minority	Measure of social deprivation, measure of the potential workload or pressure on the services of general practitioners
European Deprivation Index (EDI)	Income, poverty, social exclusion, living conditions	Construction based on data from the EU-SILC survey, for each country a national version of EDI exists
National versions of the Index of Multiple Deprivation (IMD)	-	E.g. Scottish Index of Multiple Deprivation (SIMD), Danish Deprivation Index (DANDEX), German Index of Multiple Deprivation (GIMD), Belgian Index of Multiple Deprivation (BIMD)
At Risk Of Poverty of social Exclusion (AROPE rate)	Poverty, material/social deprivation, work intensity	Share of the total population at risk of poverty or social exclusion

Appendix E – Mapping exercise of the local availability of individual-level socio-economic data among the BY-COVID use case Participant Nodes

Table A3. Individual-level socio-economic data availability in Belgium.

General elements	
Public Health Institute	Sciensano
Country	Belgium
Complying to the Baseline Use Case's Common Data Model: optional variable on individual-level socio-economic status?	Yes
Individual-level socio-economic data availability	
Data source(s)	Administrative data sources (existing data from public or private institutions) • National Register of Natural Persons (RNPP) • Census • IPCAL (Impôt des Personnes physique CALculé, dataset of the federal department of Finance includes the tax information of every Belgian resident) 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, URL: https://statbel.fgov.be/)
Data discoverability	Metadata of Statbel datasets included in the Health Information Portal: • https://www.healthinformationportal.eu/health-information-sources/belgian-national-register-natural-persons • https://www.healthinformationportal.eu/health-information-sources/belgian-mortality-registry • https://www.healthinformationportal.eu/health-information-sources/vital-statistics Codebook of available variables per dataset (in Dutch): • https://statbel.fgov.be/nl/over-statbel/wat-doen-we/microdata-voor-onderzoek/gegevenscatalogus
Data access specifications (who grants access, access procedures, access type, ...)	Researchers can request microdata if the data is necessary for their statistical or scientific research. The sharing of personal data is regulated under the General Data Protection Regulation (Regulation (EU) No 2016/679 of the European Parliament and of the Council of 27 April 2016 and the law of 30 July 2018 on the protection of natural persons with regard to the processing of personal data. Statbel has to respect the conditions set by the Statistical Supervisory Committee, the Information Security Committee or the contract concluded with the data provider (in the case of existing administrative data). For each variable requested, motivation must be given as to why the data is needed for the research. The researcher can submit a formal application using a standardized form (https://statbel.fgov.be/en/about-statbel/what-we-do/microdata-research).

	Each application will be assessed by a multidisciplinary committee within Statbel, and Statbel's Data Protection Officer (DPO) will draw up a DPO advice. The shared microdata should be placed on a secure server located within Europe, with access restricted to the researchers involved in the project.
Data discoverability and data access barriers/facilitators	Barriers: • Difficult as a non-expert in the SSH field to decide, before the data request, which indicators to use and what data sources are available. • There is no clear overview of available datasets and variables, nor detailed documentation on the methodology. • Data access is restricted by stringent data protection regulations. Requesting access to sensitive individual-level socio-economic data can involve lengthy and complex data access procedures that can hinder timely research and analysis. • Infrequent updates (e.g. for survey data) can limit the relevance and usability of the data. Facilitators: • Effective collaboration and communication between Statbel and governmental agencies/academic institutions facilitate data sharing and access.
Individual-level socio-economic data integration	
Linkage capabilities with relevant population health data	Based on the national registry number (NISS)
Pseudonymization procedures	Only pseudonymized data are available to researchers. In general, the date of birth is converted to an age class, the postal code or municipality is replaced by the district, province or region, and amounts are expressed in classes or percentiles. Statbel is legally recognized (Royal Decree of 13 June 2014) as a trusted third party and can guarantee the linking of both data and their pseudonymization. Statbel keeps the key of the pseudonymization so that possibly additional data can be added at a later date.
Data integration barriers/facilitators	Need for a common identifier (the national registry number) for data linkage, a trusted third party for data pseudonymization, and a trusted research environment for data sharing and analysis.
Individual-level socio-economic data use	
Examples of national studies/projects integrating individual-level socio-economic data to answer policy-relevant research questions in the field of infectious diseases	• Linkage of registries for COVID-19 vaccine surveillance (LINK-VACC) • Unravelling the long-term and indirect health impact of the COVID-19 crisis in Belgium (HELICON)
Socio-economic indicators frequently used to answer policy-relevant research questions in the field of infectious diseases	• Civil status • Household size and type • Individual migration background: country of birth and country of descent • Education level

	- Occupational information: working status, self-employment, and sector of employment operationalized - Net taxable income on individual and household levels expressed as deciles

Table A4. Individual-level socio-economic data availability in Aragon, Spain.

General elements	
Public Health Institute	IACS
Country	Aragon, Spain
Complying to the Baseline Use Case's Common Data Model: optional variable on individual-level socio-economic status?	No. No individual-level socio-economic data is available in healthcare. 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 as they only differentiate workers or pensioners with personal income below 18k€ between 18k€ and 100k€ from those above 100k€.
Individual-level socio-economic data availability	
Data source(s)	Administrative data sources (existing data from public or private institutions) - Census - Municipal register - Mortality data with cause of death Surveys among citizens and enterprises - Labour force survey - Household budget survey - Survey on income and living conditions
Data provider(s)	Aragonese Statistics Institute (IAEST, URL: https://www.aragon.es/organismos/departamento-de-economia-empleo-e-industria/direccion-general-de-politica-economica/instituto-aragones-de-estadistica-iaest-); the Spanish Statistical Institute (INE, URL: https://ine.es/en/index.htm)
Data discoverability	Metadata of the 'Life conditions survey' (and other socio-economic data sources) included in the Health Information Portal: Life conditions survey (INE, SPAIN) https://www.healthinformationportal.eu/health-information-sources/life-conditions-survey The codebook and methodology of the survey are available in Spanish: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=metodologia&idp=1254735976608 Access to 'social indicators' at national and regional level (INE) is available in English: https://ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736176825&menu=resultados&idp=1254735576508
Data access specifications (who grants access, access procedures, access type, ...)	Public access to microdata from administrative data sources and surveys conducted by National or Regional statistical Institutes with standard anonymization is available. Microdata is available at:

	https://ine.es/prodyser/microdatos.htm. Access to confidential microdata is exclusive to institutions that properly justify the need to use confidential statistical data to conduct statistical analyses for scientific purposes in the public interest. Entities interested in accessing confidential statistical information must, in accordance with LFEP 12/1989, submit documentation justifying that the requesting entity is a recognized entity in the field of research, studies, or analysis and that the information is required for a project of public interest endorsed by a public institution. Private petitioning entities require additional endorsement by a public institution justifying the public interest of the project. The National Statistical Institute (INE) will assess each request for access to confidential data and may, for justified reasons, deny access to the data. In the event that the project is approved, the INE shall only provide the information necessary to carry out the research project. Access to the information will require the prior signing of conditions of use and a commitment to preserve Statistical Secrecy by all members of the research team. The researcher within a public institution (or endorsed by a public institution) can submit a formal application using a standardized form: https://ine.es/ss/Satellite?L=es_ES&c=Page&cid=1259953312408&p=1259953312408&pagename=ProductosYServicios%2FPYSLayout. The National Statistical Institute (INE) will only provide the minimum information necessary to carry out the research project for which the access request is made. Researchers will not be able to access information containing direct identifiers, such as first and last names, company names, identifying numbers such as DNI, Social Security, CIF, postal address or any other information considered a direct identifier.
Data discoverability and data access barriers/facilitators	Although it is possible to link individual-level socio-economic data with patient data using a common unique identifier, individual-level socio-economic information is not currently linked to patient data; thus, it is not easily available for re-use in research. In the current framework, linking socio-economic data to patient data requires the coordination of at least the Health System with the National or Regional Statistics Institute, which implies completing a data request to both and a data processing request to the National (or Regional) Statistics Institute to produce a common pseudonymization from patient data. On the other hand, all patients in the National Health System in Spain are assigned to a reference primary care setting as the smallest healthcare administrative area. Socio-economic information from the census area level (smaller than the municipality level) can be summarized to characterize the reference population assigned at each primary care setting.
Individual-level socio-economic data integration	
Linkage capabilities with relevant population health data	Based on the National Identity Document (DNI) and the health insurance card.
Pseudonymization procedures	Only pseudonymized data are available to researchers. In general, the date of birth is converted to an age class, the postal code or municipality is replaced by the district, province or region, and amounts are expressed in classes or percentiles. The National Statistics Institute (INE) is legally recognized as a trusted third party and can guarantee the linking of both data and their pseudonymization. The INE keeps the key of the pseudonymization so that possibly additional data can be added at a later date.

Data integration barriers/facilitators	Need of a INE (<i>or Regional Statistics Institute</i>) as trusted third party for data pseudonymization; thus requiring coordination between the Health System and INE to share data and produce a common standard pseudonym.
Individual-level socio-economic data use	
Examples of national studies/projects integrating individual-level socio-economic data to answer policy-relevant research questions in the field of infectious diseases	Available examples are based on ad-hoc survey consultations on socio-economic status in population affected by COVID-19, such as: - Miranda-Mendizabal, A., Recoder, S., Sebastian, E. C., Casajuana Closas, M., Leiva Ureña, D., Manolov, R., Matilla Santander, N., Forero, C. G., & Castellví, P. (2022). Socio-economic and psychological impact of COVID-19 pandemic in a Spanish cohort BIOVAL-D-COVID-19 study protocol. <i>Gaceta sanitaria</i>, 36(1), 70–73. https://doi.org/10.1016/j.gaceta.2021.10.003 Or based on socio-economic information linked at residence area level (municipality or census section), such as: - Aguilar-Palacio I, Maldonado L, Malo S, Sánchez-Recio R, Marcos-Campos I, Magallón-Botaya R, Rabanaque MJ. COVID-19 Inequalities: Individual and Area Socioeconomic Factors (Aragón, Spain). <i>International Journal of Environmental Research and Public Health</i>. 2021; 18(12):6607. https://doi.org/10.3390/ijerph18126607 - Fernández-Martínez, N.F., Ruiz-Montero, R., Gómez-Barroso, D. et al. Socioeconomic differences in COVID-19 infection, hospitalisation and mortality in urban areas in a region in the South of Europe. <i>BMC Public Health</i> 22, 2316 (2022). https://doi.org/10.1186/s12889-022-14774-6 - Glodeanu, A., Gullón, P., & Bilal, U. (2021). Social inequalities in mobility during and following the COVID-19 associated lockdown of the Madrid metropolitan area in Spain. <i>Health & place</i>, 70, 102580. https://doi.org/10.1016/j.healthplace.2021.102580
Socio-economic indicators frequently used to answer policy-relevant research questions in the field of infectious diseases	- Occupational information: working status, self-employment, and sector of employment operationalized - Education level - Individual migration background: country of birth and country of descent - Household size and type - Civil status - Net taxable income on individual and household levels expressed as deciles

Table A5. Individual-level socio-economic data availability in Finland.

General elements	
Public Health Institute	THL
Country	Finland
Complying to the Baseline Use Case's Common Data Model: optional variable on individual-level socio-economic status?	Yes
Individual-level socio-economic data availability	

Data source(s)	Administrative data sources (existing data from public or private institutions) • Census • Municipal register • Mortality data with cause of death Surveys among citizens and enterprises • Labour force survey • Household budget survey • Survey on income and living conditions
Data provider(s)	Tilastokeskus (Statistics Finland) (https://stat.fi/index_en.html)
Data discoverability	Metadata: https://stat.fi/meta/index_en.html
Data access specifications (who grants access, access procedures, access type, ...)	Researchers can request microdata if the data is necessary for their statistical or scientific research. The registry administrators provide the data once the permissions and applications have been approved, and they serve the data for research purposes encrypted and compressed, requiring a secure password to decompress the files.
Data discoverability and data access barriers/facilitators	When accessing data from Statistics Finland, some barriers exist. Firstly, obtaining a license to use statistical data is mandatory, involving a detailed application process. Data access is further restricted by stringent data protection regulations, ensuring data confidentiality and compliance with legal requirements. Additionally, all individuals involved in handling the data must sign a confidentiality agreement, ensuring data confidentiality and compliance with legal requirements.
Individual-level socio-economic data integration	
Linkage capabilities with relevant population health data	Personal identity codes and created IDs
Pseudonymization procedures	Pseudonymized data are available to researchers. In general, the date of birth is converted to an age class, the postal code or municipality is replaced by the district, province or region, and amounts are expressed in classes or percentiles. Identification data can be left intact when the applicant requests data on the cause of death using personal identification of a particular group of people. Data on age, sex, education, occupation and socio-economic group may also be released with identification data if the applicant is entitled to collect such data by virtue of the General Data Protection Regulation. An account of this must be included in the application. An additional requirement is that the release of the data in identifiable form is necessary with regard to the study. After this, however, the data needs to be pseudonymized, and the datasets are analyzed using IDs instead of personal identification codes.
Data integration barriers/facilitators	Need for a common identifier for data linkage, a trusted party for data pseudonymization, and a trusted research environment for data sharing and analysis.
Individual-level socio-economic data use	

Examples of national studies/projects integrating individual-level socio-economic data to answer policy-relevant research questions in the field of infectious diseases	• COVID-19 Studies During the COVID-19 pandemic, Statistics Finland and THL conducted extensive research to examine how socio-economic factors influenced the spread and incidence of the virus, utilizing register-based data. • National Vaccination Program Research on the National Vaccination Program uses individual-level socio-economic data to understand how different socio-economic groups respond to vaccinations and how vaccination coverage varies across populations.
Socio-economic indicators frequently used to answer policy-relevant research questions in the field of infectious diseases	• Occupational information: working status, self-employment, and sector of employment operationalized • Education level • Individual migration background: country of birth and country of descent • Household size and type • Civil status • Net taxable income on individual and household levels expressed as deciles

Table A6. Individual-level socio-economic data availability in Austria.

General elements	
Public Health Institute	GÖG
Country	Austria
Complying to the Baseline Use Case's Common Data Model: optional variable on individual-level socio-economic status?	No. No individual-level socio-economic data is available in healthcare. Only aggregate variables on taxable income (per district) and level of education (per municipality) are available.
Individual-level socio-economic data availability	
Data source(s)	Administrative data sources: • national Labor Market Service (unemployment register) • Coordinated Labor Statistics and Register: combining census, work place census, building and housing census, households, demographic indicators, incomes and taxes statistics (including some social transfers like family subsidy), education register Surveys among citizens and enterprises: • ATHIS Austrian Health Survey (~15k respondents) • KE 19/20 household budget survey (on income, equipment and expenditures of ~7k households) • EU-SILC Community Statistics on Income and Living Conditions (~ 6k households)
Data provider(s)	Statistics Austria (https://www.statistik.at/en/)
Data discoverability	AMDC Austrian Micro Data Centre, Microdata Catalogue: https://www.statistik.at/amdc-data/#/product)

Data access specifications (who grants access, access procedures, access type, ...)	Scientific institutions can file a request for accreditation with AMDC, then a request for online access to microdata for a research project can be submitted. Data access via a secured access point equipped with a Secure Processing Environment.
Data discoverability and data access barriers/facilitators	Setting up the secure access point is effortful, the complete process starting with accreditation takes time and is by law restricted to certain types of scientific research organizations.
Individual-level socio-economic data integration	
Linkage capabilities with relevant population health data	Technically, linkage is set up quite well with a national system of domain specific person identifiers. Still, linkage is not easily possible, see barriers below.
Pseudonymization procedures	Pseudonymized data for analysis (national system of domain specific, irreversible person identifiers). Grouping of age, area and other variables to prevent reidentification. Strict obligations on results, e.g. aggregation or k-anonymity.
Data integration barriers/facilitators	Often the legal grounds are not given: data would need to be taken out of one system (which is not allowed in many cases) to feed it into the other one. When the legal basis is given, data protection concerns are a hurdle, and negotiating the processes who gives data to whom, where will the linkage happen, etc. are time consuming and will not always succeed.
Individual-level socio-economic data use	
Examples of national studies/projects integrating individual-level socio-economic data to answer policy-relevant research questions in the field of infectious diseases	- Expanding the Austrian Health Information System (ÖGIS) held at GÖG with the aspect of social differences in mortality before/during COVID-19. - Do socio-economic disparities in COVID-19 vaccination uptake mediate subsequent disparities in COVID-19 mortality? A test of fundamental cause theory with nation-wide individual-level register data.
Socio-economic indicators frequently used to answer policy-relevant research questions in the field of infectious diseases	Taxable income (mean per district), level of education (distribution per municipality)

Appendix F – Workshop interactive session Miro board

BY-COVID Spring 2024 Baseline Use Case Workshop:

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

Topic: Exploring socioeconomic data sources in Europe: availability, limitations, and mobilization

Participants: Simon Saldner, Nina Van Goethem, Francisco Estupiñán-Romero, Enrique Bernal-Delgado, Marjan Meurisse, Lorenz Dolanski-Aghamanoukjan, Irena Vipavc Brvar, Cees Hof, Eduardo Antonio Bracho Montes de Oca, Tea Osmenaj, Vittorio Grieco

Spend some time reflecting on each question. Write down your answers to the prompts.
If someone's feedback resonates with you, feel free to add a thumbs-up emoji.

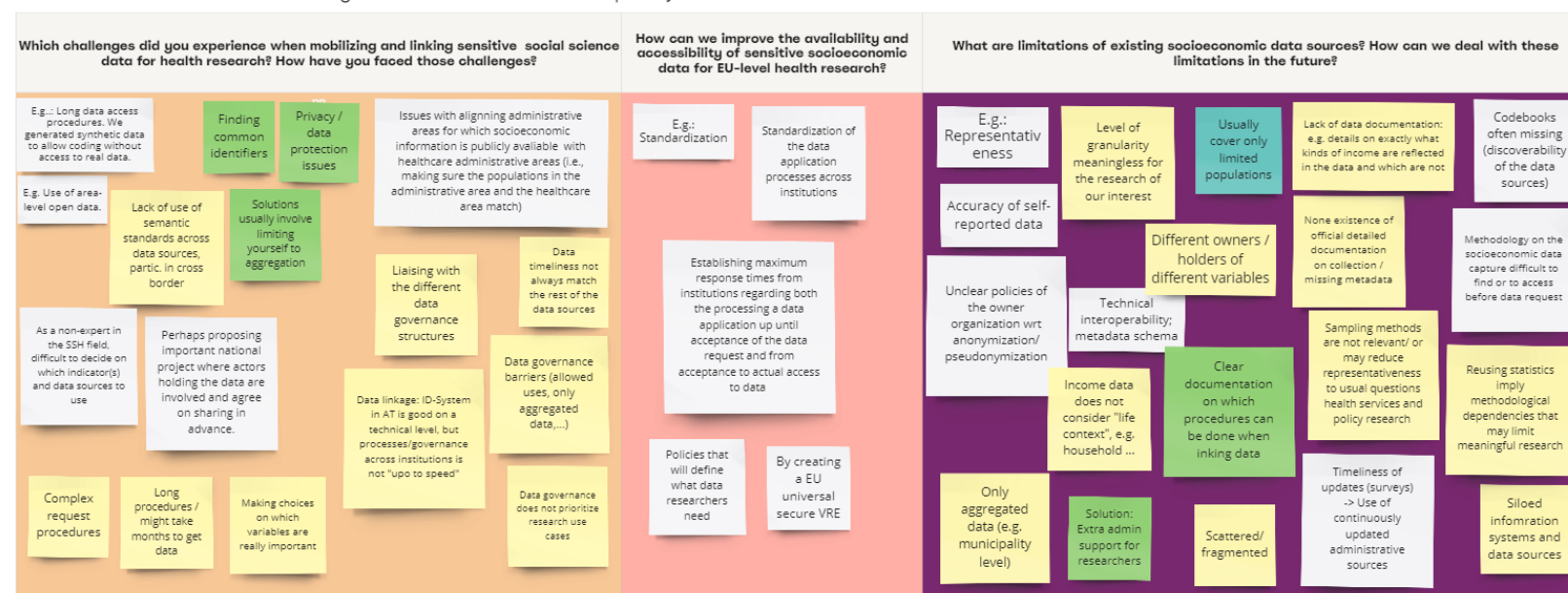


Figure A1. Workshop Miro board, including participant notes. Collected during the online workshop “BY-COVID Spring 2024 Baseline Use Case Workshop: Integration of individual-level socio-economic data for infectious diseases research and prevention in Europe” on the 6th of June 2024, 9:30-11:30 CET.

Appendix G – Thematic analysis of workshop participant notes

Table A7. Thematic analysis of participant notes, collected during the online workshop “BY-COVID Spring 2024 Baseline Use Case Workshop: Integration of individual-level socio-economic data for infectious diseases research and prevention in Europe” on the 6th of June 2024, 9:30-11:30 CET.

Comment/Note	Code	Theme
<i>Which challenges did you experience when mobilizing and linking sensitive social science data for health research? How have you faced those challenges?</i>		
Long data access procedures	Data access procedures	Data governance
Complex request procedures	Data access procedures	Data governance
Long procedures / might take months to get data	Data access procedures	Data governance
Privacy/data protection issues	Data protection	Data governance
Liaising with the different data governance structures	Data governance	Data governance
Data governance barriers (allowed uses, only aggregated data, ...)	Data governance	Data governance
Data governance does not prioritize research use cases	Data governance	Data governance
Data linkage: ID-System in AT is good on a technical level, but processes/governance across institutions is not "up to speed"	Data governance	Data governance
Issues with aligning administrative areas for which socio-economic information is publicly available with healthcare administrative areas (i.e., making sure the populations in the administrative area and the healthcare area match)	Aggregated data	Data governance
Perhaps proposing important national project where actors holding the data are involved and agree on sharing in advance.	Data access procedures	Data governance: Proposed solutions
Solutions usually involve limiting yourself to aggregation	Aggregated data	Data governance: Proposed solutions
We generated synthetic data to allow coding without access to real data.	Synthetic data	Data governance: Proposed solutions
Use of area-level open data	Aggregated data	Data governance: Proposed solutions
Data timeliness not always match the rest of the data sources	Timeliness	Data quality
Finding common identifiers	Identifiers	Data governance
As a non-expert in the SSH field, difficult to decide on which indicator(s) and data sources to use	Indicators	Data discovery
Making choices on which variables are really important	Indicators	Data discovery
<i>How can we improve the availability and accessibility of sensitive socio-economic data for EU-level health research?</i>		

Establishing maximum response times from institutions regarding both the processing a data application up until acceptance of the data request and from acceptance to actual access to data	Data access procedures, Timeliness	Data governance: Proposed solutions
Policies that will define what data researchers need	Policies	Data governance: Proposed solutions
Standardization of the data application processes across institutions	Standardization, Data access procedures	Standardization, Data governance: Proposed solutions
By creating a EU universal secure VR	Data protection, Universal	Data governance: Proposed solutions, Centralization/ Fragmentation: Proposed solutions
Standardization	Standardization	Standardization
<i>What are limitations of existing socio-economic data sources? How can we deal with these limitations in the future?</i>		
Unclear policies of the owner organization with regards to anonymization/ pseudonymization	Policies	Data governance
Only aggregated data (e.g. municipality level)	Aggregated data	Data governance
Solution: Extra admin support for researchers	Admin support	Data governance: Potential solutions
Accuracy of self-reported data	Accuracy	Data quality
Level of granularity meaningless for the research of our interest	Granularity	Data quality
Reusing statistics imply methodological dependencies that may limit meaningful research	Reusing statistics	Data quality
Timeliness of updates (surveys) → Use of continuously updated administrative sources	Timeliness	Data quality
Income data does not consider "life context", e.g. household	Not considering	Data quality
Representativeness	Representativeness	Representativeness
Usually cover only limited populations	Coverage	Representativeness
Sampling methods are not relevant/or may reduce representativeness to usual questions health services and policy research	Representativeness	Representativeness
Lack of data documentation: e.g., details on exactly what kinds of income are reflected in the data and which are not	Lack of documentation	Data discovery
Codebooks often missing (discoverability of the data sources)	Discoverability	Data discovery

Non-existence of variables official detailed documentation on collection / missing metadata	Lack of documentation	Data discovery
Clear documentation on which procedures can be done when linking data	Documentation	Data discovery
Methodology on the socio-economic data capture difficult to find or to access before data request	Data access procedures, Lack of documentation	Data discovery
Technical interoperability; metadata schema	Documentation	Data discovery, Standardization
Scattered/fragmented	Fragmentation	Centralization/ Fragmentation
Siloed information systems and data sources	Siloed information	Centralization/ Fragmentation
Different owners / holders of different variables	Fragmentation	Centralization/ Fragmentation

Appendix H – Theoretical background

Table A8. Non-exhaustive overview of pathogen genomic data repositories and their characteristics

Repository	Year of launch	Maintained by	Databases	Access	Conditions for data access	Hosted data	Hosted genomic data formats
Global initiative on sharing avian flu data (GISAID)	2008	The board of 'Friends of GISAID'	- EpiFlu (2008) - EpiCoV (2020) - EpiRSV (2021) - EpiPox (2022) - EpiArbo (2022)	Controlled access	Access agreement (agree to authenticate academic identity and not to republish or link GISAID genomes without permission from the data producer)	- Raw sequences - Assembles sequences - Annotations	- FASTA - VCF
European Nucleotide Archive (ENA)	1980	EMBL-EBI*	- Sequence Read Archive (SRA) - Trace Archive - EMBL-Bank	Open access (pathogen genomic data)	Free, unrestricted access	- Raw sequences (SRA if produced by NGS techniques, Trace Archive if produced by older technologies like Sanger sequencing) - Assembled sequences (EMBL-Bank) - Annotations (EMBL-Bank)	- FASTA - FASTQ - BAM/CRAM - AGP - SFF - VCF - EMBL flat file format (annotations)
DNA Data Bank of Japan (DDBJ)	1987	DDBJ Center of the NIG**	- DDBJ Sequence Read Archive (DRA) - DDBJ (narrow) - BioProject - BioSample	Open access (pathogen genomic data)	Free, unrestricted access	- Raw sequences (DRA) - Aligned sequences (DRA) - Annotated sequences (DDBJ) - Assembled sequences (DDBJ) - Study information (BioProject) - Sample information (BioSample)	- FASTQ - BAM - DDBJ format (annotations)
NCBI databases	1982	NCBI***	- Sequence Read Archive (SRA) - GenBank	Open access	Free, unrestricted access	- Raw sequences (SRA) - Aligned sequences (SRA) - Annotated sequences (GenBank)	- BAM/CRAM - SFF - HDF5 - FASTQ - FASTA/QUAL pairs

							- GenBank sequence format (annotations)

* EMBL-EBI: The European Molecular Biology Laboratory's European Bioinformatics Institute

** NIG: National Institute of Genetics

*** NCBI: The NIH-NLM National Center for Biotechnology

Appendix I – Thematic analysis interview transcripts

Table A9. Selected statements or quotations, and keywords, together with identified codes from the thematic analysis of seven in-depth interviews on the integration of pathogen genomic data with epidemiological and clinical data sources for public health research.

Statements/ Quotations	Keywords	Codes
“Only recently, when the pandemic started, our institute started to build capacities in the field of bioinformatics.”	Build capacities in the field of bioinformatics	Bioinformatics and computational resources (Feasibility: Resources – Infrastructure)
		Workforce capacity (Feasibility: Resources)
		Capacity building (Feasibility: Resources)
	Only recently	Delayed response (Feasibility: Barriers)
“Essential equipment was procured several years ago for processing of genome sequencing of the coronavirus.”	Equipment for processing genome sequences	Bioinformatics and computational resources (Feasibility: Resources – Infrastructure)
		Laboratory capacity (Feasibility: Resources, Data lifecycle governance: Data generation)
“in order to complement the genetic epidemiology and following of emerging types of coronaviruses”	Complement genetic epidemiology	Genetic epidemiology (Utility: Genomics-driven infectious disease management)
	Following of emerging types of coronaviruses	Monitoring emerging viral variants (Utility: Analysis of viral variants)
“We are still very young as an institute, as a team, dealing with genomic data.”	Institute/team dealing with genomic data	Workforce capacity (Feasibility: Resources)
		Bioinformatics and computational resources (Feasibility: Resources – Infrastructure)
	Very young	Delayed response (Feasibility: Barriers)

“There is not yet any strategic planning in the field of developing a bioinformatics department”	Not yet any strategic planning	Lack of strategic planning (Feasibility: Barriers), Strategic preparedness (Feasibility: Preparedness)
		Delayed response (Feasibility: Barriers)
	Developing a bioinformatics department	Bioinformatics and computational resources (Feasibility: Resources – Infrastructure)
		Workforce capacity (Feasibility: Resources)
		Capacity building (Feasibility: Resources)
“Some of those are using human samples, some of those are using waste water sampling, and some of those are using isolates of microorganisms, in order to sequence it for various purposes.”	Human sample	Host sample (Data lifecycle governance: Sampling – Source)
	Waste water sample	Wastewater sample (Data lifecycle governance: Sampling – Source)
	Isolates of microorganisms	Pathogen sample (Data lifecycle governance: Sampling – Source)
“My department didn't do the linkage of sequencing data with the population health data sources so far. We didn't hit any requests in order to conduct such research.”	Linkage of genomic data with population health data sources	Linking heterogeneous data sources (Data lifecycle governance: Data re-use – Data linkage)
	No request	Perception (Feasibility: Barriers)
		Delayed response (Feasibility: Barriers)
“We have access to health care data from ehealth records, from hospital information systems, patient registries and others.”	Access	Data access permission (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
	Health care data from ehealth records, hospital information systems, patient registries	Heterogeneous data sources (Data lifecycle governance: Data collection/Data re-use)
“There was no need so far to link those data for any particular questions. There is potential, but no requests were received.”	No need, no requests received	Perception (Feasibility: Barriers – Human and political interests)

		Delayed response (Feasibility: Barriers)
	Link those data	Linking heterogeneous data (Data lifecycle governance: Data re-use – Data linkage)
“following of the coronavirus subtypes in the population”	Following coronavirus subtypes	Monitoring emerging viral variants (Utility: Analysis of viral variants)
“The complete material is being created in English, because it's being defined by the European Commission as well.”	English	Language (Feasibility: Facilitators)
	Defined by the European Commission	European-level recommendations (Feasibility: European-level harmonization)
“It is never too late to establish genetic epidemiology teams that are in close collaboration with other teams.”	Never too late	Aspiration (Feasibility: Human and political interests)
	Establish genetic epidemiology teams	Workforce capacity (Feasibility: Resources)
		Capacity building (Feasibility: Resources)
	Collaboration with other teams	Collaboration with other teams (Feasibility: Collaboration)
“Integration with HERA and establishing integration with other laboratories within European projects.”	HERA	European projects (Feasibility: European-level harmonization)
	European projects	
	Integration with other laboratories	Collaboration with laboratories (Feasibility: Collaboration)
“A strategic letter was sent to our director in order to establish and develop a bioinformatics team.”	Strategical letter	Strategic preparedness (Feasibility: Preparedness)
	Develop a bioinformatics team	Workforce capacity (Feasibility: Resources)
		Bioinformatics and computational resources (Feasibility: Resources – Infrastructure)

		Capacity building (Feasibility: Resources)
“Everything stops at the administration level, because politically it's not that well recognized and important work.”	Everything stops at the administration level	Administrative barriers (Feasibility: Barriers)
	Politically it's not that well recognized	Perception (Feasibility: Barriers)
		Lack of political recognition (Feasibility: Barriers – Human and political interests)
“We are missing the critical mass in order to start this strategic approach.”	Missing critical mass	Lack of critical mass (Feasibility: Barriers – Human and political interests)
	Strategic approach	Strategic preparedness (Feasibility: Preparedness)
“a direct grant, that has been assigned by the European Commission”	Direct grant assigned by the European Commission	Funding (Feasibility: European-level harmonization)
“We have already started to establish services for opening the data, where particular elements need to be built, for example a data catalogue, safety processing, secure processing environments, gateway to the European Health Data Space and others.”	Services for opening the data, where particular elements need to be built	Infrastructure (Data lifecycle governance: Data sharing), Infrastructure development (Feasibility: Resources – Infrastructure)
“We didn't recognize the potential from bioinformatics and therefore the coordination of the team, the responsibility, and the mandate of the institution is not clear yet.”	Didn't recognize the potential	Perception (Feasibility: Barriers)
		Delayed response (Feasibility: Barriers)
	Coordination of the team, the responsibility, and the mandate of the institution is not clear	Lack of organizational preparedness (Feasibility: Barriers – Lack of preparedness)
“if there are recommendations at the European level: how each member state should tackle genetic epidemiology, what are the expected skills, what are the expected themes and their responsibilities.”	Recommendations at the European level	European-level recommendations (Feasibility: European-level harmonization)
		Aspiration (Feasibility: Human and political interests)
	Expected skills	Capacity building (Feasibility: Resources)

	Responsibilities	Organizational preparedness (Feasibility: Preparedness)
“As a team, we are not ambitious to take over the strategic activities, but just to help and stress to our government on the importance of the field.”	Strategic activities	Strategic preparedness (Feasibility: Preparedness)
	Stress to our government the importance of the field	Political willingness (Feasibility: Barriers – Human and political interests), Perception (Feasibility: Barriers – Human and political interests)
“We are working to implement a new lab-based surveillance system, so that we get our data in the wake of the pandemic a lot more timely and complete”	New system	Innovation (Feasibility: Facilitators)
	Surveillance system	Surveillance system (Data lifecycle governance: Data collection – Infectious disease surveillance)
	More timely	Timeliness (Data lifecycle governance: Data collection – Data quality)
		Efficiency (Feasibility: Facilitators)
	More complete	Data completeness (Data lifecycle governance: Data collection – Data quality), Data quality (Data lifecycle governance: Data collection – Data quality)
“It’s actual continuous surveillance of infectious diseases.”	Continuous surveillance of infectious diseases	Continuous surveillance (Data lifecycle governance: Data collection – Infectious disease surveillance)
“The clinical laboratories send in data, also physical samples”	Clinical laboratories	Data provider (Data lifecycle governance: Data collection)
	Send in data	Data reporting (Data lifecycle governance: Data collection)
	Physical samples	Pathogen sample (Data lifecycle governance: Sampling – Source)

“There's the legal obligation to notify cases of infectious diseases from the labs to the municipal health services.”	Legal obligation to notify cases of infectious disease	Legal obligation for case notification (Feasibility: Facilitators)
	Notify cases	Case notification (Data lifecycle governance: Data collection – Infectious disease surveillance)
“exhaustive for those cases, for those diseases, that are notifiable”	Exhaustive	Exhaustive sampling (Data lifecycle governance: Sampling – Sampling strategy)
	Notifiable	Legal obligation for case notification (Feasibility: Facilitators – Legal and regulatory aspects)
“The stream through the clinical laboratories, it's voluntary to start with. They are not paid to do that. It is purely voluntary that they send data through.”	Stream / Send data	Data reporting (Data lifecycle governance: Data collection)
	Clinical laboratories	Data provider (Data lifecycle governance: Data collection)
	Voluntary that they send data through	Voluntary reporting (Feasibility: Barriers – Legal challenges)
	Not paid / No budget	No financing (Feasibility: Barriers)
“Accompanied with that, there is quite extensive additional information, clinical information: use of antivirals, comorbidities, substance use, symptoms, etc.”	Extensive clinical information	Data completeness (Data lifecycle: Data collection – Data quality)
“During the pandemic, of course, all the sequencing capacity that you could get, was used, including in the academic centers.”	All sequencing capacity used	Laboratory capacity (Feasibility: Resources, Data lifecycle governance: Data generation)
“For SARS-CoV-2 there were some quick legal additions made so we were able to link the vaccination status with the lab data in general, including the sequence.”	Quick legal additions	Timeliness (Feasibility: Facilitators)
		Legal amendments (Feasibility: Facilitators – Legal and regulatory aspects)
	Link the vaccination status with lab data	Linking heterogeneous data sources (Data lifecycle governance: Data re-use – Data linkage)

“allowed us to analyze if there was vaccine escape”	Analyze if there was vaccine escape	Impact of viral variants on vaccine effectiveness (Utility: Analysis of viral variants)
“continuous monitoring, of course, surveillance, to see whether your vaccines are still working in the general population”	Continuous monitoring	Continuous surveillance (Data lifecycle governance: data collection – Infectious disease surveillance)
	Whether your vaccines are still working in the general population	Vaccine effectiveness (Utility)
“The sample ID was coming through in both streams.”	Sample ID	Unique sample identifier (Data lifecycle governance: Data collection – Identifiers)
“Just as for SARS-CoV-2, we're trying to get a change in the law, that you get that ID that is allowed to be sent through and that reporting by labs to the national institute, becomes mandatory for selective diseases again. Those 2 changes implemented, that would be a fundamental improvement.”	Just as for SARS-CoV-2, change in the law	Legal amendments (Feasibility: Facilitators – Legal and regulatory aspects)
	ID that is allowed to be sent through	Transmittable unique identifiers (Feasibility: Facilitators – Data management aspects)
	Reporting becomes mandatory	Legal obligation for reporting (Feasibility: Facilitators – Legal and regulatory aspects)
“The sequences that we generated, we deposit them on GISAID, that's excluding most of the clinical information.”	Sequences deposited on GISAID	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	Excluding most of the clinical information	Limited metadata (Data lifecycle governance: Data sharing)
“The system we had for collecting sequences basically broke down because of the number of data in it.”	System broke down	Limited system robustness (Feasibility: Barriers – Lack of preparedness)
	Because of the number of data	Limited data storage capacity (Feasibility: Barriers – Technical challenges)
“The project we are working on is to have a single database for all pathogens, where this information is stored in a standardized form.”	A single database	Centralized data storage (Feasibility: Resources – Infrastructure), Centralized data storage (Data lifecycle governance: Data storage),

		Centralization (Feasibility: Facilitators)
	Stored in a standardized form	Standardized data storage (Feasibility: Resources – Infrastructure), Standardized data storage (Data lifecycle governance: Data storage), Standardization (Feasibility: Facilitators)
“You have academic centers, academic hospitals, they have clinical labs, but one of their primary interests is research. If they need to send through data that they could do research on eventually voluntarily to the national institute, and the national institute also is a research institute, you can imagine that you have to start discussing.”	Primary interest is research	Competing research interests (Feasibility: Barriers – Human and political interests)
“Commercial labs are more easy to collaborate with” “It is quite hammered on that they have to contribute to public health.” “If you don't have this research aim, then it's actually easier to do that.”	Collaborate	Collaboration with commercial labs (Feasibility: Collaboration)
	Contribute to public health	Sense of purpose (Feasibility: Facilitators – Human and political interests)
	Research aim	No competing research interests (Feasibility: Facilitators – Human and political interests)
“The sequence data are not normally stored in the labs on LIMS, the laboratory information system. That already creates 2 streams within the laboratories.”	Sequence data stored	Storing sequence data (Data lifecycle governance: Data storage)
	2 streams within the laboratories	Separate streams (Feasibility: Barriers – Technical challenges)
“If each of them implements the machine-to-machine connection, then it saves a lot of time for the labs.”	Machine-to-machine connection	Technical solutions (Feasibility: Facilitators – Data management aspects)
	Saves a lot of time	Timeliness (Feasibility: Facilitators)
“There's no real financing for that, unfortunately.”	No financing	No financing (Feasibility: Barriers)

“Everybody sends through their data in their own format, with their own frequency.”	Own format, own frequency	Lack of standardized data reporting (Feasibility: Barriers)
“We need to store that data in one place and in a standardized format.” “You need to have a standardized internal storage of your data, that is also pathogen or disease independent”	Store data in one place	Centralized data storage (Feasibility: Resources - Infrastructure), Centralized data storage (Data lifecycle governance: Data storage), Centralization (Feasibility: Facilitator)
	Standardized format, Standardized internal storage	Standardized data storage (Feasibility: Resources – Infrastructure), Standardized data storage (Data lifecycle governance: Data storage), Standardization (Feasibility: Facilitators)
	Pathogen or disease independent	Flexibility (Feasibility: Facilitators), Scalable infrastructure (Feasibility: Resources – Infrastructure)
“If you have a new pathogen that comes up, then you need to be able to add that quite quickly.”	New pathogen	Pandemic preparedness (Feasibility: Preparedness)
	Add quickly	Infrastructure preparedness (Feasibility: Preparedness), Capacity: Scalable infrastructure (Feasibility: Resources – Infrastructure)
“For most of the diseases the amount of cases is low enough that you can actually just sequence all of them.”	Sequence all cases	Laboratory capacity (Feasibility: Resources, Data lifecycle governance: Data generation)
“If you don't have the denominator, in terms of the negative tests, you actually don't know if your increase is due to increased testing rather than increase of the incidence. So a fundamental improvement, apart from legal requirements, is to send through the negative test results as well.”	Send through the negative test results	Importance of reporting negative test results (Data lifecycle governance: Data collection – Infectious disease surveillance)
“Every single IT Department in every single site has to jump through hoops.”	IT Department has to jump through hoops	IT challenges (Feasibility: Barriers – Technical challenges)
		Collaboration with IT departments (Feasibility: Collaboration)

“I would think the actual databases that should be used for these purposes are public, they are pathogen aspecific, have huge structural funding, and huge databases as well, including the reads.”	Public databases	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	Pathogen aspecific	Flexibility (Feasibility: Facilitators), Scalable infrastructure (Feasibility: Resources – Infrastructure)
	Structural funding	Sustainable funding (Feasibility: Facilitators – Funding)
	Huge databases	Data storage capacity (Feasibility: Resources – Infrastructure)
“During the pandemic we were asked to build a data platform to provide infection data on COVID-19 for researchers.”	Build a data platform	Infrastructure development (Feasibility: Resources – Infrastructure)
	Provide data for researchers	Data access permission (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
“Before the COVID-19 crisis, the genomic data was under the radar. Genomic data came into our focus only because of the COVID-19 pandemic.”	Under the radar	Perception (Feasibility: Barriers – Human and political interests)
	Came into our focus because of the COVID-19 pandemic	Delayed response (Feasibility: Barriers)
“Right now we are trying to establish a research project on long or post COVID, where we try to link genomic data with outcome data.”	Research project on long- or post-COVID	Research on long COVID (Utility: Clinical impact)
	Link genomic data with outcome data	Linking heterogeneous data sources (Data lifecycle governance: Data re-use – Data linkage)
“We try to establish a correlation between variants, and vaccine effectiveness and outcomes.”	Vaccine effectiveness	Impact of viral variants on vaccine effectiveness (Utility: Analysis of viral variants)
		Clinical impact of viral variants (Utility: Analysis of viral variants)
“There was an obligation to report COVID-19. It's an epidemiological reporting system. In this	An obligation	Legal obligation for case notification (Facilitators)

dataset you had a unique identifier for each person, pseudonymized data. This was until June of 2023 and then this obligation was gone.”	Report COVID-19	Data reporting (Data lifecycle governance: Data collection)
	Epidemiological reporting system	Surveillance system (Data lifecycle governance: Data collection – Infectious disease surveillance)
	Unique identifier for each person	Unique person identifier (Data lifecycle governance: Data collection – Identifiers)
	Pseudonymized data	Pseudonymization (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
“You have a law that all data on persons, on people, are provided with an identifier. This identifier is linkable for each area of governance.”	A law	Legal obligation to provide unique identifiers (Feasibility: Facilitators – Legal and regulatory aspects)
	Linkable	Linking heterogeneous data sources (Data lifecycle governance: Data re-use – Data linkage)
“We could theoretically link this data with other data sources, because it has the identifier. But it is, or it wasn't permissible to link this data together.” “So in theory we would have been able to, but there was no legal basis to do it.”	It has the identifier	Unique identifier (Data lifecycle governance: Data collection – Identifiers)
	It wasn't permissible, no legal basis	Legal restrictions (Feasibility: Barriers – Legal challenges)
“could only be linked together via the regional correlation”	Linked together via regional correlation	Linkage via regional correlation (Data lifecycle governance: Data re-use – Data linkage)
“Laboratories in each federal state sent the samples to these two institutes who did the sequencing.” “random samples from each federal state” “Until June 2023, it was representative for federal states”	Two institutes who did the sequencing	Centralized sequencing (Feasibility: Resources, Data lifecycle governance: Data generation), Centralization (Feasibility: Facilitators)
	Random samples	Random sampling (Data lifecycle governance: Sampling – Sampling strategy)

	Representative	Representativeness (Data lifecycle governance: Data collection)
"There were weekly reports for the government on the variants."	Weekly reports on the variants	Variant updates (Utility: Analysis of viral variants)
"They switched to wastewater monitoring and they switched to a sentinel system right now. Different sentinel doctors and hospitals then sent the sentinel samples for sequencing."	Wastewater monitoring	Wastewater surveillance (Data lifecycle governance: Data collection – Infectious disease surveillance)
	Sentinel system	Sentinel surveillance (Data lifecycle governance: Data collection – Infectious disease surveillance)
"Research institutions could apply for access to the data. They have to state the purpose and the research questions and then a scientific advisory board would decide if it's okay to give them access to the data and the form, anonymized or pseudonymized. They would then get access to a secure environment."	Apply for access to the data	Data access application (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
	State the purpose and the research question	
	Scientific advisory board	
	Secure environment	Secure research environment (Data lifecycle governance: Data re-use – Privacy and data protection regulations), Capacity: Secure infrastructure (Feasibility: Resources – Infrastructure)
"A lot of people just wanted to know vaccine effectiveness in dependence of variants or reinfections based on variants."	Vaccine effectiveness in dependence of variants	Impact of viral variants on vaccine effectiveness (Utility: Analysis of viral variants)
	Reinfections based on variants	Impact of viral variants on reinfections (Utility: Analysis of viral variants)
"We don't have the legal basis to link individual datasets on health, because health data is very sensitive, of course, and we have to follow a lot of rules. And the GDPR is like a sword hanging over the heads of everyone and everyone is very cautious."	Legal basis	Legal restrictions (Feasibility: Barriers – Legal challenges)
		Regulatory framework (Feasibility: Facilitators – Legal and regulatory aspects)
	Health data is very sensitive	Data privacy concerns (Feasibility: Barriers)

	Follow a lot of rules, GDPR	Regulatory compliance (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
	A sword hanging over our heads, cautious	Compliance anxiety (Feasibility: Barriers – Legal challenges)
“What was facilitating the process was, in the state of emergency, or in the pandemic state, people were willing to be more generous with the interpretation of the GDPR.”	In state of emergency	Emergency provisions (Feasibility: Facilitators – Legal and regulatory aspects)
	Generous with the interpretation of the GDPR	Flexible GDPR interpretation (Feasibility: Facilitators – Legal and regulatory aspects)
“The cooperation between researchers was really good and the administration was kept to a minimum.”	Cooperation between researchers	Collaboration between researchers (Feasibility: Collaboration)
	Administration was kept to a minimum	Minimal administration (Feasibility: Facilitators – Data management aspects)
“What we're trying to implement now is a legal basis to link health data on an individual level.”	Legal basis to link health data on an individual level	Regulatory framework (Feasibility: Facilitators – Legal and regulatory aspects), Legal basis for data integration (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
		Individual Health Data Linkage (Data lifecycle governance: Data re-use – Data linkage)
“Things that facilitated the linkage of data was the cooperation between researchers, the cooperation between agencies.”	Cooperation between researchers	Collaboration between researchers (Feasibility: Collaboration),
	Cooperation between agencies	Collaboration between agencies (Feasibility: Collaboration)
“We weren't really prepared for these kinds of questions and to link data together.”	We weren't really prepared	Lack of preparedness (Feasibility: Barriers – Lack of preparedness)
“We weren't prepared as an organization to handle such big datasets.”	We weren't prepared	Lack of preparedness (Feasibility: Barriers – Lack of preparedness)

	Handle such big datasets	Data storage capacity (Feasibility: Resources – Infrastructure)
“Linking of the data, providing access to the data, had to be secure. So we had to establish a 2-factor authentication for access to the research environment and then we had to draft up contracts, where the liability is regulated etc.”	Linking of the data had to be secure	Secure data linkage (Data lifecycle governance: Data re-use – Data linkage)
	2-factor authentication for access to the research environment	Secure infrastructure (Feasibility: Resources – Infrastructure), Secure research environment (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
	Draft up contracts	Legal agreements (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
“It was learning for us.”	It was learning	Learning from experience (Feasibility: Resources – Expertise)
“The biggest unmet need is of course the legal basis.”	Biggest unmet need is the legal basis	Legal restrictions (Feasibility: Barriers – Legal challenges)
“People don't really trust the government here and the government agencies. So everything data related, and the data linkage, and health data, is viewed very skeptical here”	People don't really trust the government	Distrust (Feasibility: Barriers – Human and political interests)
	Viewed very skeptical	
“A big unmet need would be to establish trust with the population, to establish the narrative that research with secondary health data can benefit the society, can benefit the individual, and is safe under certain preconditions.”	Establish trust	Establishing trust (Feasibility: Perception)
	Establish a narrative	Establish a positive narrative (Feasibility: Perception)
“The European Health Data Space is in the pipeline, so we hope that this will help us in the long run.”	The European Health Data Space	European-level regulations (Feasibility: European-level harmonization)
	We hope	Aspiration (Feasibility: Facilitators – Human and political interests)
“It is really difficult to find common ground, common standards on a national level. We have	Find common ground	Finding a consensus (Feasibility: Collaboration)

9 federal states, they're very independent in their decision making, in their regional laws."	Common standards	Standardization (Feasibility: Facilitators)
	Federal states independent in their decision making, in their regional laws	Decentralization (Feasibility: Barriers)
"On a European level to find common data standards, common metadata standards, common ways to exchange data, I think it's difficult."	European level common data standards	Standardization (Feasibility: Facilitators), European level data standards (Feasibility: European-level harmonization)
	European level common metadata standards	Standardization (Feasibility: Facilitators), European level metadata standards (Feasibility: European-level harmonization)
	European level common ways to exchange data	Standardization (Feasibility: Facilitators), European level standards for data exchange (Feasibility: European-level harmonization)
"Genomic surveillance is centralized."	Genomic surveillance	Centralized genomic surveillance (Feasibility: Resources), Centralization (Feasibility: Facilitators)
	Centralized	
"We have a long-term experience in the application of genomics for research purposes for several pathogens, mostly focusing on exploring genetic diversity and evolution of pathogens."	Long-term experience	Established expertise (Feasibility: Resources – Expertise)
	Exploring genetic diversity and evolution of pathogens	Exploring genetic diversity and evolution of pathogens (Utility: Analysis of viral variants)
"the application of genomics for prevention and control of infectious diseases, so that we can provide more evidence- and genomic-informed public health clinical decision making"	Application of genomics for prevention and control of infectious diseases	Infectious disease control (Utility: Genomics-driven infectious disease management)
	Evidence- and genomic-informed public health and clinical decision making	Genomic-informed public health practice (Utility: Genomics-driven infectious disease management)

“We collaborate with the multiple geographically spread laboratories and health centers. We collaborate with the local, regional, and national health authorities, and also with the epidemiology team here, which was responsible for all the modelling and forecasting during the pandemic.”	Collaborate with geographically spread laboratories and health centers	Collaboration with laboratories (Feasibility: Collaboration)
	Collaborate with the local, regional, and national health authorities	Collaboration with public health authorities (Feasibility: Collaboration)
	Collaborate with the epidemiology team here	Collaboration: Collaboration with other teams (Feasibility: Collaboration)
“Our role there was to inform the ministry of health about the emergence and circulation of the SARS-CoV-2 variants. That had a huge impact in terms of public health measures.”	Inform about the emergence and circulation of SARS-CoV-2 variants	Monitoring emerging viral variants (Utility)
	Impact in terms of public health measures	Genomic-informed public health practice (Utility: Analysis of viral variants)
“For the genomics, we received the samples. So almost all the sequencing was conducted in our team.”	Almost all sequencing conducted in our team	Centralized sequencing (Feasibility: Resources, Data lifecycle governance: Data generation)
“We work with the sequence data that we generate. The sequence data are available in multiple databases (namely the GISAID, GenBank, ENA), and epidemiological and clinical data are coming from different sources. But our common sources are laboratories and hospitals from our network, local, regional, and national health authorities, and the national epidemiological surveillance system that also gathers all the demographic, clinical and epidemiological data for notifiable diseases. These sources are routinely available.”	Sequence data	Heterogeneous data sources (Data lifecycle governance: Data re-use)
	Epidemiological and clinical data coming from different sources	
	These sources are routinely available	Routine data (Data lifecycle governance: Data re-use)
“We have publications regarding the introduction and transmission dynamics of the virus in the country.”	Introduction and transmission dynamics of the virus	Virus transmission dynamics (Utility: Analysis of viral transmission)
“We had publications regarding typical clinical cases, such as immunocompromised patients, a congenital infection, and some studies with health care professionals.”	Typical clinical cases	Clinical case studies (Utility: Clinical impact)

“We link clinical and epidemiological data in the context of investigation of nosocomial outbreaks or local outbreaks in some particular regions, always in connection with public health authorities.”	Link clinical and epidemiological data	Linking heterogeneous data sources (Data lifecycle governance: Data re-use – Data linkage)
	Investigation of nosocomial outbreaks or local outbreaks	Outbreak investigation (Utility: Analysis of viral transmission)
	Connection with public health authorities	Collaboration with public health authorities (Feasibility: Collaboration)
“The team from epidemiology that was responsible for vaccine effectiveness assessments is also hosted here, so we collaborated a lot and we have publications regarding the vaccine effectiveness studies during the variant wave change.”	The team from epidemiology, we collaborated a lot	Collaboration with other teams (Feasibility: Collaboration)
	Vaccine effectiveness studies during the variant wave change	Impact of viral variants on vaccine effectiveness (Utility: Analysis of viral variants)
“We tried to link the data to investigate where these new variants were circulating and which kind of cases they were causing.”	Link the data	Data linkage (Data lifecycle governance: Data re-use – Data linkage)
	Investigate where these new variants were circulating	Virus transmission dynamics (Utility: Analysis of viral transmission)
		Monitoring emerging viral variants (Utility: Analysis of viral variants)
	Which kind of cases they were causing	Clinical impact of viral variants (Utility: Analysis of viral variants/Clinical impact)
“We create a report that we send to the ministry of health. And we also insert it in the public, bilingual site that is available to everyone that wants to see it.”	Insert it in a public, bilingual site	Language (Feasibility: Facilitators)
	Available to everyone	Public data (Data lifecycle governance: Data sharing)
“If you take a look at the website and if you click on each of the infections, the dots, you see the map and the locations of all the infections and circulation of the variants.”	The website	Public data (Data lifecycle governance: Data sharing)

"The only data that is linked systematically is the date of collection, the age group, and the location."	Linked systematically	Systematic data linkage (Data lifecycle governance: Data use/analysis)
"We need to have permission to access this data."	Permission to access this data	Data access permission (Data lifecycle governance: Data re-use – Privacy and data access regulations)
"Normally we use the national health system patient ID and less frequently we also use the notification number of that platform."	Patient ID	Unique person identifier (Data lifecycle governance: Data collection – Identifiers)
	Notification number of that platform	
"When the need arose, the national health authority collaborated with us, and they provided the epidemiological and clinical data that was inserted in the platform, and we provided the genomic data, and we worked together."	When the need arose	Ad-hoc data linkage (Data lifecycle governance: Data re-use – Data linkage)
	The national health authority collaborated with us	Collaboration with public health authorities (Feasibility: Collaboration)
"The system is already built and adapted, but it's still difficult to operationalize the submission of batches of sequences."	Difficult to operationalize the submission of batches of sequences	Technical difficulties (Feasibility: Barriers – Technical challenges)
		Data reporting (Data lifecycle governance: Data collection)
"The sampling strategy mostly relied on random nationwide surveillance conducted throughout the time, to capture spatiotemporal dissemination of the virus. So we build a laboratory network with laboratories from all the regions, and we systematically receive samples."	Random nationwide surveillance conducted throughout the time	Random sampling (Data lifecycle governance: Sampling – Sampling strategy)
		Continuous surveillance (Data lifecycle governance: Data collection – Infectious disease surveillance)
	Capture spatiotemporal dissemination of the virus	Virus transmission dynamics (Utility: Analysis of viral transmission)
"The coverage and the amount of samples that we collected changed throughout the pandemic. We tried to shape it with the epidemic curve, the number of samples that the labs were able to provide us, and of course with the laboratory capacity that we had installed at a given point. Sometimes it was possible to accomplish with the"	The coverage and the amount of samples that we collected changed through the pandemic	Sampling strategy (Data lifecycle governance: Sampling – Sampling strategy)
		Sample coverage (Data lifecycle governance: Sampling – Sampling strategy)

cutoffs suggested by the ECDC and sometimes it was not possible.”	The number of samples that the labs were able to provide us	Laboratory capacity (Feasibility: Resources, Data lifecycle governance: Data generation)
	The laboratory capacity that we had installed at a given point	Laboratory capacity (Feasibility: Resources, Data lifecycle governance: Data generation)
“All sequence data, consensus sequences are released in GISAID in a timely manner. We also release all the raw data, the reads, to the European Nucleotide Archive. We also report everything on the website. We wanted since the beginning to be completely transparent and share all that.”	GISAID, European Nucleotide Archive	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	Release data in a timely manner, report everything on the website	Timely data sharing (Data lifecycle governance: Data sharing)
	We wanted to be completely transparent	Transparency (Feasibility: Facilitators)
“Anyone can use our data because genomic data is completely released in ENA, in GISAID, and also on our website. However, if they want more information, more metadata, we don't have that. We always need the national health authority. So any university would also need the cooperation of the national health authority.”	Genomic data completely released in ENA, GISAID	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	If they want more metadata, we need the cooperation of the national health authority	Data access restrictions (Feasibility: Barriers – Legal challenges)
	Need the cooperation of the national health authority	Collaboration with public health authorities (Feasibility: Collaboration)
“Just to summarize the website, you'll find multiple sorts of data. You'll find our situation reports and we had more than 100 throughout the pandemic where we do this analysis of frequency change throughout the time. We provide a dynamic, interactive table, where you can search for any specific lineage or variant and see the geographic spread and also the frequency throughout the time. And we also provide phylogenetic trees, that are grouped by time period, but where people can search for the profile of the sequence, the location, the laboratory that sent the sample. Everything is open.”	The website, everything is open	Public data (Data lifecycle governance: Data sharing)
“The linkage was mostly the lineage, our variant classification, which has been the key information for routine surveillance and for vaccine effectiveness studies. For specific studies	The linkage was mostly the lineage, our variant classification	Linking variant classification information (Data lifecycle governance: Data re-use – Data linkage)

we go through the in depth phylogenetic analysis to identify clusters of interest, to identify local outbreaks. We also do specific mutation analysis to check the evolution of the virus in immunocompromised patients.”	Variant classification, key information for vaccine effectiveness studies	Impact of viral variants on vaccine effectiveness (Utility: Analysis of viral variants)
	Phylogenetic analysis to identify clusters of interest, local outbreaks	Outbreak investigation (Utility: Analysis of viral transmission)
	Evolution of the virus in immunocompromised patients	Clinical case studies (Utility: Clinical impact)
		Exploring genetic diversity and evolution of pathogens (Utility: Analysis of viral variants)
“This surveillance system was adapted to collect information about whole genome sequencing, the lineage and also the code for the sequence, the GISAID code.”	Surveillance system was adapted	Flexibility (Feasibility: Facilitators), Scalable infrastructure (Feasibility: Resources – Infrastructure)
“One thing that was facilitating was the centralization of the sequencing and analysis in a single institute.”	Centralization of the sequencing	Centralized sequencing (Feasibility: Resources, Data lifecycle governance: Data generation), Centralization (Feasibility: Facilitators)
“The funding that we received, namely from the European Commission, was critical because we had the opportunity to renew and improve the infrastructures, to acquire reagents, and to hire more people to the team.”	Funding that we received, mainly from the European Commission	Funding (Feasibility: Facilitators), Funding (Feasibility: European-level harmonization)
	Renew and improve the infrastructures	Infrastructure preparedness (Feasibility: Preparedness), Infrastructure development (Feasibility: Resources – Infrastructure), Infrastructure improvement (Feasibility: Resources – Infrastructure)
	Acquire reagents	Laboratory capacity (Feasibility: Resources, Data lifecycle governance: Data generation)
	Hire more people	Workforce capacity (Feasibility: Resources)

“One thing that facilitated a lot and helped us to start doing sequencing since the first case was our previous capacity building in genomics and bioinformatics. We had developed a bioinformatics platform that was rapidly adapted to analyze SARS-CoV-2 data. It's very user friendly, and even when we recruited new people, it was quite easy to train them to use the platform.”	Previous capacity building in genomics and bioinformatics	Capacity building (Feasibility: Resources), Bioinformatics and computational resources (Feasibility: Resources – Infrastructure), Established expertise (Feasibility: Resources – Expertise)
	Bioinformatics platform	Infrastructure development (Feasibility: Resources – Infrastructure)
	Rapidly adapted	Scalable infrastructure (Feasibility: Resources – Infrastructure), Flexibility (Feasibility: Facilitators)
	User-friendly platform, easy to train	User-friendly infrastructure (Feasibility: Resources – Infrastructure)
“The fact that our national institute of health doesn't have access to this data that is collected systematically is something that needs to be solved because otherwise the research studies will depend on the availability, on the willingness, of the national health authorities to release or to share the data.”	Access to data that is collected systematically	Continuous surveillance (Data lifecycle governance: Data collection – Infectious disease surveillance), Data access permission (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
	Availability, willingness	Workforce capacity (Feasibility: Resources), Willingness to share data (Feasibility: Facilitators – Human and political interests)
“In such a small country, we have the excellent conditions to uniformize all of this.”	Excellent conditions to uniformize	Standardization (Feasibility: Facilitators), Harmonization (Feasibility: Facilitators)
“I would like to have a legal framework with everything well defined.”	Legal framework	Regulatory framework (Feasibility: Facilitators)
“We also need to develop more efficient systems for data sharing and communication between our institute and the data owners, the laboratories.”	Efficient systems for data sharing	Efficient systems for data sharing (Feasibility: Resources)
		Efficiency (Feasibility: Facilitators)

	Efficient communication with the data owners	Collaboration with data owners (Feasibility: Collaboration)
“Usually most of the questionnaires for each individual patient are not filled in a complete manner and that's a problem.”	Questionnaires not filled in a complete manner	Lack of complete data (Feasibility: Barriers), Data completeness (Data lifecycle governance: Data collection – Data quality)
“We really need to work towards a better harmonization of inquiries at national and international level and improve, and harmonize the articulation between the member states and the ECDC. This should be something that is agreed at the top level. I think it should come from the top to the bottom.”	Harmonization	Harmonization (Feasibility: Facilitators)
	Agreed at the top, should come from the top to the bottom	Top-down decision-making (Feasibility: European-level harmonization)
“We want the data to be useful. And, this was the main reason why we are submitting to GISAID, because you know that lots of platforms that were critical for variant monitoring, covSPECTRUM, Nextstrain, CoVariants, outbreak.Info, and so on, systematically collective data from GISAID. So we found it important to have our sequence data there.”	Submitting to GISAID	Data sharing: Depositing in public data repositories (Data lifecycle)
	Variant monitoring	Monitoring emerging viral variants (Utility: Analysis of viral variants)
(About GISAID) “We know problems with transparency. We know the problems with conditional data access. We decided that everything will also move to GenBank and we'll have all data in both databases, with this goal that the data is as useful as possible.”	Problems with transparency	Lack of transparency (Feasibility: Barriers)
	Problems with conditional data access	Data access restrictions (Feasibility: Barriers – Legal challenges)
	GenBank	Depositing in public data repositories (Data lifecycle governance: Data sharing)
“I think the international authorities should provide not only guidance, but also should provide more strong recommendations about this.”	International authorities should provide strong recommendations	European-level recommendations (Feasibility: European-level harmonization)
“It is a European funded project that aims at setting up and consolidating an overarching infrastructure for the collection of genomic, clinical and epidemiological data for studying infectious diseases”	European funded project	Funding (Feasibility: Facilitators), European projects (Feasibility: European-level harmonization)
	Setting up an overarching infrastructure	Infrastructure development (Feasibility: Resources)
“The intention is that if we have this database and the sequencing data that it might be pushed	Pushed to public databases, like GISAID or ENA	Depositing in public data repositories (Data lifecycle governance: Data sharing)

to public databases, like GISAID or ENA, later on to make it more available.”	Make it more available	Public data (Data lifecycle governance: Data sharing)
“It's mostly a technological investment. So what is facilitating this is having funding. Sustainable funding is needed. If there's no sustainable funding, then we can't maintain all the people working on the project now, and the infrastructure and keep it up and running.”	Technological investment	Technological investment (Feasibility: Resources – Infrastructure)
	Sustainable funding	Sustainable funding (Feasibility: Facilitators – Funding)
	Maintain people working on the project	Workforce capacity (Feasibility: Resources)
	Keep the infrastructure up and running	Infrastructure preparedness (Feasibility: Preparedness)
(Facilitating factor) “Of course, the people, the expertise that they have in setting this up”	The people, the expertise	Established expertise (Feasibility: Resources – Expertise)
“There was a COVID crisis, but everyone has moved on from that. But it's still very important to take the lessons learned and to put that in sustainable infrastructure for further use and for the future. So everyone has moved on, but at the same time we should still invest in what was learned or what was built then.”	Take the lessons learned	Learning from experience (Feasibility: Resources – Expertise)
	Sustainable infrastructure	Infrastructure preparedness (Feasibility: Preparedness)
“It's a bit of a balancing act of what is the cost versus what do you want in the end. It's more an economic point of view, I think.”	Balancing act of the cost versus what you want, economic point of view	Cost-benefit analysis (Feasibility: Facilitators)
“What is needed in order to link the data more easily is that there would be some kind of standard for data collection, standard terminology and so on, really at each and every level. You need digitalization and standardization in terms of data collection. So I think that would be an ideal situation when everyone would be aligned.”	Standard for data collection	Standard for data collection (Data lifecycle governance: Data collection)
	Standard terminology	Standardization (Feasibility: Facilitator)
	Standardization	
	Digitalization	Digitalization (Feasibility: Resources – Infrastructure)

“we put together a European community of practice of people doing similar things”	European community	European community (Feasibility: European-level harmonization/Collaboration)
“We basically provide data to the federal office of public health, and they can link the genomic data that they receive from us with the notifications that they receive.”	Provide data	Data provider (Data lifecycle governance: Data collection)
	Link the genomic data	Data linkage (Data lifecycle governance: Data use/analysis)
“We just need a common identifier. Sure. But what is this common identifier? How do we ensure that every hospital can generate it? It also needs to be a pseudonymized identifier and so on. So it was more challenging than I thought.”	Common identifier	Common identifier (Data lifecycle governance: Data collection – Identifiers)
	Pseudonymized identifier	Pseudonymized identifier (Data lifecycle governance: Data re-use – Data linkage)
“The data that we receive is not anonymized. So, it gives those possibilities and that's something very important to us. That's why we host the database on a very highly secure infrastructure to make sure that we can have identifiers that can allow others, provided of course they have ethical clearance, to actually link data sets.”	Not anonymized	Non-anonymized data (Data lifecycle governance: Data collection)
	Highly secure infrastructure	Secure infrastructure (Feasibility: Resources – Infrastructure)
	Ethical clearance to link data sets	Data linkage restrictions (Feasibility: Barriers – Legal challenges)
	Link data sets	Data linkage
“The tools have been updated very often during the pandemic.”	Updated very often	Flexibility (Feasibility: Facilitators)
“We work very closely and really tailored to their needs.”	Work very closely	Collaboration (Feasibility: Collaboration)
“The diagnostic labs are in charge of the sampling. At the beginning of the pandemic, the idea was to do a random sampling.” “This lasted until the end of 2022, more or less, and then they decided to focus more the sampling strategy on sequencing hospitalized patients. So the focus is now more on that and to	Random sampling	Random sampling (Data lifecycle governance: Sampling – Sampling strategy)
	Sampling strategy	Sampling strategy (Data lifecycle governance: Sampling – Sampling strategy)

also have the wastewater surveillance to know a bit more about the general contextual situation”	Sequence hospitalized patients	Targeted sampling (Data lifecycle governance: Sampling – Sampling strategy)
	Wastewater surveillance	Wastewater surveillance (Data lifecycle governance: Sampling)
“They are interested in understanding better what variants make people go to the hospital.”	What variants make people go to the hospital	Clinical impact of viral variants
“We publish all the data to the ENA. The consensus genome is published right away and the raw genomic data sometimes take a bit more time. If we have a lot of big data sets, it just takes time to forward them to the ENA. But everything is openly accessible very quickly.”	ENA	Public data repositories
	Big data sets, takes more time to transfer	Data transfer challenges
	Openly accessible very quickly	Timely data sharing
“We have very clear rules on how to access the sensitive data. That might also be information about the age or the sex of the patient and things like that, that we don't share openly.”	Clear rules on how to access the data	Data access restrictions (Feasibility: Barriers – Legal challenges), Data access permission (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
“People can ask basically to make a project on the data.” “They should ask for ethical clearance to use the data. If they have that, and that the data providers agree that their data is used for that research project, then they can have access to the data.”	Ask for ethical clearance to use the data	Data access application (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
	Data providers agree that their data is used	
“To facilitate discoverability of the data, we now also have a public platform where people can go and they can search.”	Discoverability	Discoverability (Data lifecycle governance: Data re-use)
	Public platform	Technical solutions (Feasibility: Resources – Infrastructure)
“We needed the technical IT infrastructure to be able to host sensitive data.”	Technical IT infrastructure	Infrastructure development (Feasibility: Resources – infrastructure)
	Host sensitive data	Secure infrastructure (Feasibility: Resources – infrastructure)

“We also have ethical clearance to actually host sensitive data. To host those identifiers, we also asked at the beginning for ethical clearance, because it was not clear if it was always the law of epidemics that was applying.”	Ethical clearance to host sensitive data	Ethical clearance to host sensitive data (Data lifecycle governance: Data storage)
	also asked at the beginning for ethical clearance, because it was not clear if was the law of epidemics applying	Compliance anxiety (Feasibility: Barriers – Legal challenges)
“From a more legal perspective, we have all the contracts. It took almost two years to get signed. So that's something that was very important. But to make sure that everyone agrees on data access rules, data re-use rules, data transfer rules.”	It took almost two years to get signed	Delayed response (Feasibility: Barriers)
	Make sure that everyone agrees on data access rules, data re-use rules, data transfer rules	Finding a consensus (Feasibility: Collaboration)
“We were also lucky that people are okay with sharing their data openly. At least we didn't have to fight too much against that. So as soon as we could provide technical means of sharing data openly without giving more work to the data providers, they were fine with it. So what we agreed is that they would submit to us and that we would submit on their behalf to GISAID and the ENA. And then the data goes also to public health authorities, to many other places, and they don't need to worry about that.”	People are okay with sharing their data	Willingness to share data (Feasibility: Facilitators – Human and political interests)
	Technical means	Technical solutions (Feasibility: Facilitators – Data management aspects)
	Without giving more work to the providers	Workforce capacity (Feasibility: Resources)
“We also established a lot of working procedures during the pandemic by learning lessons of what works and doesn't work.”	Establish working procedures	Establishing working procedures (Feasibility: Resources – Expertise)
	Learning lessons	Learning from experience (Feasibility: Resources – Expertise)
“We will soon be covering several viruses and several bacteria, and this means that we would already have bioinformatics pipelines that work a bit for both worlds. And so, even if now the new pathogens would be something yet different, everything is already there, the infrastructure, all the models to share the data with the public health authorities are already there. So now adapting things is much quicker. So we really believe that it makes us much more reactive now, in case of something else happening.”	Everything is already there, the infrastructure, the models to share data	Infrastructure preparedness (Feasibility: Preparedness)
	Adapting is much quicker	Scalable infrastructure (Feasibility: Resources – Infrastructure)
		Flexibility (Feasibility: Facilitators)

	It makes us reactive	
		Pandemic preparedness (Feasibility: Preparedness)
“You need to have an identifier that uniquely identifies an isolate, and this is not always that easy to get in the end.” “This identifier had to go from the lab that was doing the sequencing up to us. And in some labs, it was tricky to get this identifier and they would send it afterwards, after several weeks. So then they had to send data updates to us to give us this information. Then there is confusion.”	Identifier that uniquely identifies an isolate	Unique isolate identifier (Data lifecycle governance: Data collection – Identifiers)
	Confusion, send data updates	Confusion with retrospective data updates (Data lifecycle governance: Data collection)
“For our public health authorities at the beginning, it took them a couple of months to really make sure that they were allowed to link data coming from our platform.” “For a public health authority, they always need to be very careful and make sure that they're not accessing or linking data that they should not be. And they were very careful about that.”	It took them a couple of months	Delayed response (Feasibility: Barriers)
	Need to be careful they're not accessing or linking data that they should not	Data privacy concerns (Feasibility: Barriers – Legal challenges)
		Regulatory compliance (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
		Compliance anxiety (Feasibility: Barriers – Legal challenges)
“But in the end, it's also about how you interpret the law.”	Interpret the law	Flexible GDPR interpretation (Feasibility: Facilitators – Legal and regulatory aspects)
“We had cases where some labs just decided that there was no reason why they should be sending those identifiers to us and we really had to prove to them that it was public health authority that was the one requesting that. It's not that we asked that just for fun, and people had the feeling at some point that we were asking it just in case. And for them, it was not justified then.” “It was really their consideration for data privacy, and that they were afraid that they were doing something wrong in sharing that and that they	Consideration for data privacy	Data privacy concerns (Feasibility: Barriers – Legal challenges)
	They were afraid that they were doing something wrong, that they could be accused of doing something wrong	Compliance anxiety (Feasibility: Barriers - Legal challenges)

could be then accused of doing something wrong.”		
“You need to collect just the minimal things that ensure that your data is not a silo.”	Collect just the minimal things	Data completeness (Data lifecycle governance: Data collection – Data quality)
		Data minimization (Data lifecycle governance: Data collection)
	Ensure that your data is not a silo	Avoiding data silos (Data lifecycle governance: Data collection)
“We had a list of metadata that was super long, but people were not providing that because there was no immediate use.” “So now, I just insist on the fact that they send identifiers that allow others for the research projects to potentially link data if they needed to. And this is the minimal thing that we should have and that we are allowed to have from the review board.”	Metadata not provided, because there was no immediate use	Data privacy concerns (Feasibility: Barriers – Legal challenges)
		Data minimization (Data lifecycle governance: Data collection)
	Insist that they send identifiers that allow others to potentially link data	Common identifier (Data lifecycle governance: Data collection – Identifiers)
		Avoiding data silos (Data lifecycle governance: Data collection)
	Minimal thing that we should have and that we are allowed to have	Data minimization (Data lifecycle governance: Data collection)
“There was a lot of sequencing. It was really massive. And maybe it was a bit too much. Because it's also very costly. So, in the sampling strategy, for sure, there are some improvements possible.”	A lot of sequencing, very costly	Cost-benefit analysis (Feasibility: Facilitators)
	In the sampling strategy there are some improvements possible	Sampling strategy (Data lifecycle governance: Sampling – Sampling strategy)
“There is the wastewater surveillance that is gaining a lot of attention, because, of course, you need to sample much less to get this overview.”	Wastewater surveillance	Wastewater surveillance (Data lifecycle governance: Sampling)
	Sample much less to get this overview	Sampling strategy (Data lifecycle governance: Sampling – Sampling strategy)

“We optimized by centralizing in a few big sequencing facilities, the samples where they were sequenced. This also allowed us to reduce the costs.”	Centralizing in a few big sequencing facilities	Centralized sequencing (Feasibility: Resources)
	Allowed us to reduce the costs	Cost reduction (Feasibility: Facilitators - Funding)
“In terms of data re-use we are publishing very rapidly, openly non sensitive data. Of course, the sensitive data could be more discoverable.” “We would like to be able to make sparkle queries on our database and to be able to really discover the data and know what is in there.” “Ideally be able to do that from a central portal. That you could go there and know in every country, okay, if I put together all the European countries and I need this metadata information, this is the amount of data and I know whom to contact to get access to the data. And this is what we would like to see 1 day happening.”	Publishing very rapidly	Timely data sharing (Data lifecycle governance: Data sharing)
	Could be more discoverable, to be able to really discover the data	Discoverability (Data lifecycle governance: Data re-use)
“On GISAID, you share only consensus genomes, so of course it's much lighter. So it was pretty easy to put in place.” “There was overhead afterwards, because they were coming back a lot if they for instance observed frameshifts or a premature stop codon and they were asking if we were sure about it. It was just ridiculous manual work that we had to do.” “Luckily afterwards we could automate some of these things, because they allowed us to upfront say if you had validated the sequence or not. So, this was greatly reduced afterwards.”	GISAID	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	Overhead	Overhead (Feasibility: Barriers)
	Ridiculous manual work	Lack of automation (Feasibility: Barriers – Technological challenges)
	Afterwards we could automate some of these things	Technological solutions (Feasibility: Resources - Infrastructure)
	Validated the sequence	Data validation (Data lifecycle governance: Data collection - Data quality)
“For the ENA putting in place the system is much more cumbersome. Then afterwards the overhead on the opposite side is very low.” “We do sometimes have some issues with the raw data because there are large files. You can have more issues when you're transferring them and things like that.”	ENA	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	Overhead	Overhead (Feasibility: Barriers)
	Issues with the raw data because there are large files, issues when you're transferring them	Challenges with transferring large data files (Data lifecycle governance: Data sharing)

"It seems to be very tricky, but that I think is quite a big problem, because of how much you can trust the data."	Trust the data	Data trustworthiness (Data lifecycle governance: Data re-use)
"The timeliness. So the consensus genomes, they usually are published the same day we receive the data and the raw data, it can take a few more days."	Timeliness, published the same day	Timely data sharing (Data lifecycle governance: Data sharing)
"Some epidemiologist asked us whether we could deploy a server because he developed tools for epidemiology in R."	Deploy a server	Infrastructure development (Feasibility: Resources - Infrastructure)
		Technical solutions (Feasibility: Resources - Infrastructure)
"Elixir started organizing a network with all the countries, to know who was doing what and sharing experience and so on." "To share experience and also to build projects together."	Network with all the countries	International collaboration (Feasibility: Collaboration)
	Share experience	Sharing experience (Feasibility: Resources - Expertise)
	Build projects together	Collaboratively building projects (Feasibility: Collaboration)
"The government developed a strong program for funding national research infrastructures." "For instance, for bioinformatics, for sequencing" "So they bought all the equipment and funded the functioning costs and so on."	Strong funding program	Funding (Feasibility: Facilitators - Funding)
	Research infrastructures	Infrastructure development (Feasibility: Resources - Infrastructure)
"They funded infrastructure so that all the labs could use the infrastructure instead of spreading the money and that each lab would have to build their own, to buy their own computer, install the software and so on."	Funded infrastructure	Funding development (Feasibility: Facilitators - Funding)
		Infrastructure development (Feasibility: Resources - Infrastructure)
	Instead of spreading the money	Centralization (Feasibility: Facilitators)
"And so for this reason, we were equipped, trained."	Equipped, trained	Established expertise (Feasibility: Resources - Expertise)

“In the federation, what we did is to build cross-team platforms so that the people collaborate in national projects.” “So it's a very strong mutualization not only of the funding, but also a mutualization of the human resources.”	Build cross-team platforms	Infrastructure development (Feasibility: Resources - Infrastructure)
	Mutualization of the funding, human resources	Collaboration (Feasibility: Collaboration)
“It's very stimulating for the people to feel that they're part of a network, part of the community, to build together and to reuse the material and so on.”	Part of a network, community	Community engagement (Feasibility: Collaboration)
“They didn't know where to centralize the sequences. Because each of these labs had their own computing facility. But it was completely secured, so that you couldn't transfer sequences from one hospital to another one, for instance.” “So they asked us whether we would be able to deploy a bioinformatics platform that would allow the gathering of all the sequences, gather all the metadata, and provide access to them, so that they could do the weekly report and so on.”	Centralize the sequences, gathering all the sequences	Centralized genomic surveillance (Feasibility: Resources)
		Centralized data storage (Feasibility: Resources - Infrastructure)
	Deploy a bioinformatics platform	Infrastructure development (Feasibility: Resources - Infrastructure)
“There were several sampling strategies.” “One of them was the use of indications.” “If the patient who comes for a COVID test has also, for instance, an immunodeficiency, long COVID, a comorbidity, or recent travel history, then automatically this sample is sent to perform full genome sequencing of the virus.” “Then another part of it was random sequencing. They called it flash studies. Every Tuesday the labs were asked to send all the samples for genomic sequencing.” “At one moment it became too much, when we had the waves of COVID infections” “the sequencing labs were not able to follow anymore, because there were too many cases for the flash survey”	Sampling strategies	Sampling strategy (Data lifecycle governance: Sampling - Sampling strategy)
	Use of indications	Targeted sampling (Data lifecycle governance: Sampling - Sampling strategy)
	Random sequencing	Random sampling (Data lifecycle governance: Sampling – Sampling strategy)
	At one moment it became too much	Laboratory capacity (Feasibility: Resources)
“But there was not a very strong sampling strategy because we had to act very rapidly.” “It was sort of tradeoff between what is feasible and what would be ideal”	Sampling strategy	Sampling strategy (Data lifecycle governance: Sampling – Sampling strategy)
“Currently it's really a challenge to have an academic server hosting his data. Of course, the hospitals are doing this. They have the right to	Don't have the right to centralize, give access to external people	Legal restrictions (Feasibility: Barriers - Legal challenges)

host the data that they produce, but they don't have the right to centralize, to collect, and to give access to external people.”		Data access restrictions (Feasibility: Barriers - Legal challenges)
		Decentralization (Feasibility: Barriers)
“We worked a lot on the specification of the metadata on the defining the granularity, and the tradeoff between having as much metadata as possible to be able to exploit it, but not opening the risk of identification of persons, even indirect”	Metadata not opening the risk of identification of persons	Limited metadata (Data lifecycle governance: Data sharing)
		Data privacy concerns (Feasibility: Barriers - Legal challenges)
“A lot of discussions, because what we said is: okay, but someone can have several sequencing indications. But the forms that they built in the beginning were only for one sequencing indication.”		Technical difficulties (Feasibility: Barriers - Technical challenges)
		Administrative barriers (Feasibility: Barriers)
“So the sequencing lab needs to put all the data that would enable pairing, and send it to the place where the pairing would be done” “And for this, we used a certified third party. It's a company who is certified in order to do pseudonymization. And so they hold the original data, they generate the pseudonyms. And then they push the pseudonyms in the database.” “This pseudonym should have enabled it to pair the data from the four databases” “And actually, this is what never happened, because this part that is still dotted here, what was conceived was not implemented.”	Pseudonymization	Pseudonymization (Data lifecycle governance: Data re-use - Privacy and data protection regulations)
	Pair the data	Data linkage (Data lifecycle governance: Data re-use - Data linkage)
“You have many different agencies for research” “There is a fragmentation of the research organization, with collaborations, but also competition between them.” “And the administration rules of these are different.”	Fragmentation	Decentralization (Feasibility: Barriers)
	Collaborations	Collaboration (Feasibility: Collaboration)
	Competition between them	Competing research interests (Feasibility: Barriers – Human and political interests)
	Different administration rules	Administrative barriers (Feasibility: Barriers)

“All the scientists, they want to work together, but the administration does not allow some things to be done. So it's something that still needs to progress”	Administration does not allow some things to be done	Administrative barriers (Feasibility: Barriers)
“There were also all the institutional complications and the fact that people were a bit camping on their positions and protecting their own green square. There are also human difficulties.”	Institutional complications	Institutional complications (Feasibility: Barriers)
	Challenges in human interactions	Challenges in human interactions (Feasibility: Barriers – Human and political interests)
“So the idea is to build a sustainable infrastructure that will be everlasting, that will be at least funded structurally, but that will work at a reduced rhythm between pandemics, and in case of pandemic, it can be scaled up.”	Build a sustainable infrastructure	Sustainable infrastructure (Feasibility: Resources – Infrastructure)
		Infrastructure development (Feasibility: Resources – Infrastructure)
	Funded structurally	Sustainable funding (Feasibility: Facilitators – Funding)
	Can be scaled up	Scalable infrastructure (Feasibility: Resources – Infrastructure)
“So we are really actively involved in open science, in sharing the data, the bioinformatics solutions, the software and so on. We are really working in an open science spirit.”	Open science	Open science (Data lifecycle governance: Data sharing)
“Since the beginning, the sequencing labs didn't want to open the data.” “They wanted to submit only to GISAID. It's data sharing under a protective licence. And GISAID had in addition very strong governance issues that were revealed in June 2023.” “So we had problems with GISAID, but the sequencing labs only wanted to upload the data to GISAID and they really refused to publish the data on ENA.” “GISAID only takes the assembled genomes, but not the raw data. The raw data is sometimes important for co-infections and so on.” “And then the ministry said that the data must go to ENA because this is legally required here. Any data that is produced with public funding must go open unless there is a personal data	GISAID, ENA	Depositing in public data repositories (Data lifecycle governance: Data sharing)
	Didn't want to open the data	Lack of recognition (Feasibility: Barriers – Human and political interests)
		Competing research interests (Feasibility: Barriers – Human and political interests)
	The data must go to ENA because this is legally required	Legal obligation for opening data (Feasibility: Facilitators – Legal and regulatory aspects)

protection issue. But for viral genomes, there was no reason to prevent the opening.” “The sequencing labs were in this position because they have a dual mission. They are the National Center of Reference, they are involved in public health surveillance and so on. But at the same time, they are research laboratories in universities or hospitals. And so they have contradictory goals or indicators, because they need to publish.”	Dual mission	Competing research interests (Feasibility: Barriers - Human and political interests)
“They said, okay, we will transfer the raw data only once the consortium agreement is signed.” “The fact is that this consortium agreement took two years to be signed. And so they never transferred the raw data.”	Took two years to be signed	Administrative barriers (Feasibility: Barriers)
		Delayed response (Feasibility: Barriers)
“If you overprotect, nothing happens and then you don't protect because you cannot set up the health system that the people need.”	Overprotect	Compliance anxiety (Feasibility: Barriers – Legal challenges)
“For political reasons we were never granted to start doing the pseudonymization.” “So the sequencing data could never be merged”	Political reasons	Political barriers (Feasibility: Barriers – Human and political interests)
	Pseudonymization	Pseudonymization (Data lifecycle governance: Data re-use – Privacy and data protection regulations)
“They asked us to build an infrastructure that would enable us to collect and provide access to all the data and metadata. And actually in one week we did it, because actually they had funded a national infrastructure of bioinformatics already for several years. And this national infrastructure was a federation of different platforms with a mutualized task force and people with high skills in computing storage, computing databases, software deployment, and developers.” “Due to that, we already had the national server, a cluster that had a galaxy server, in order to deploy bioinformatics tools for sequence analysis. We already had all this infrastructure.”	Build an infrastructure	Infrastructure development (Feasibility: Resources – Infrastructure)
	Funded a national infrastructure for several years	Sustainable funding (Feasibility: Facilitators – Funding)
	Mutualized task force	Organizational preparedness (Feasibility: Preparedness)
	We already had all this infrastructure	Infrastructure preparedness (Feasibility: Preparedness)

“The fact that we had a national bioinformatics infrastructure enabled us to react very rapidly.”	bioinformatics infrastructure, react rapidly	Infrastructure preparedness (Feasibility: Preparedness)
“At the international level, actually, the fact that ELIXIR exists. It is a federation of 23 European countries.” “We are working together on software tools, data management, ...”	Federation of 23 European countries	International collaboration (Feasibility: Collaboration)
“I think this is a positive lesson. Given the investments both at the national and European level on bioinformatics infrastructures, we were able to very rapidly set up a solution.”	Bioinformatics infrastructures, we were able to very rapidly	Infrastructure preparedness (Feasibility: Preparedness)
	Investments	Funding (Feasibility: Facilitators - Funding)

Table A10. Identified codes, themes, sub-themes and sub-sub-themes from the thematic analysis of seven in-depth interviews on the integration of pathogen genomic data with epidemiological and clinical data sources for public health research.

Codes	Sub-sub-theme	Sub-theme	Theme	
Capacity building		Resources	Feasibility	
Workforce capacity				
Laboratory capacity				
Centralized sequencing				
Centralized genomic surveillance				
Efficient systems for data sharing				
Infrastructure development	Infrastructure			
Infrastructure improvement				
Scalable infrastructure				
Sustainable infrastructure				
Secure infrastructure				
User-friendly infrastructure				

Bioinformatics and computational resources		
Standardized data storage		
Centralized data storage		
Data storage capacity		
Technological investment		
Digitalization		
Technical solutions		
Learning from experience	Expertise	
Sharing experience		
Established expertise		
Establishing working procedures		
Strategic preparedness		Preparedness
Organizational preparedness		
Pandemic preparedness		
Infrastructure preparedness		
European projects		European-level harmonization
European-level funding		
European-level recommendations		
European-level data standards		
European-level metadata standards		
European-level standards for data exchange		

European-level regulations		Collaboration	
Top-down decision-making			
European community			
Collaboration with other teams			
Collaboration with laboratories			
Collaboration with commercial labs			
Collaboration with IT departments			
Collaboration between researchers			
Collaboration between agencies			
Collaboration with public health authorities			
Collaboration with clinicians			
Collaboration with data owners			
International collaboration			
Collaboration			
European community			
Collaboratively building projects			
Community engagement			
Finding a consensus			
Establishing trust		Perception	
Establishing a positive narrative			

Legal obligation for case notification	Legal and regulatory aspects	Facilitators	
Regulatory framework			
Legal amendments			
Legal obligation to provide unique identifiers			
Flexible GDPR interpretation			
Legal obligation for reporting			
Legal obligation for opening data			
Emergency provisions			
Transmittable unique identifiers	Data management aspects		
Minimal administration			
Technical solutions			
Lack of data			
Sense of purpose	Human and political interests		
No competing research interests			
Aspiration			
Willingness to share data			
Political willingness			
Sustainable funding	Funding		
Funding			
Cost reduction			
Standardization			

Harmonization			
Centralization			
Flexibility			
Efficiency			
Transparency			
Timeliness			
Language			
Innovation			
Cost-benefit analysis			
Perception	Human and political interests	Barriers	
Lack of political recognition			
Lack of recognition			
Competing research interests			
Political barriers			
Distrust			
Lack of critical mass			
Challenges in human interactions			
Limited system robustness	Lack of preparedness		
Lack of organizational preparedness			
Lack of preparedness			
Lack of strategic planning			

Lack of standardized data reporting			
IT challenges	Technical challenges		
Technical difficulties			
Separate streams			
Lack of automation			
Limited data storage capacity			
Legal restrictions	Legal challenges		
Data access restrictions			
Data linkage restrictions			
Data privacy concerns			
Voluntary reporting			
Compliance anxiety			
Decentralization			
Lack of complete data			
Lack of transparency			
Delayed response			
Administrative barriers			
Institutional complications			
Overhead			
No financing			
Host sample	Source	Sampling	Data lifecycle governance
Wastewater sample			

Pathogen sample			
Exhaustive sampling	Sampling strategy		
Random sampling			
Targeted sampling			
Sampling strategy			
Sample coverage			
Laboratory capacity		Data generation	
Centralized sequencing			
Heterogeneous data sources		Data collection	
Surveillance system	Infectious disease surveillance		
Wastewater surveillance			
Sentinel surveillance			
Continuous surveillance			
Case notification			
Importance of reporting negative test results			
Timeliness	Data quality		
Data quality			
Data completeness			
Data validation			
Representativeness	Representativeness		
Data provider			
Data reporting			

Non-anonymized data			
Confusion with retrospective data updates			
Avoiding data silos			
Data minimization			
Standard for data collection			
Unique person identifier	Identifiers		
Unique sample identifier			
Unique isolate identifier			
Unique identifier			
Common identifier			
Centralized data storage		Data storage	
Standardized data storage			
Storing sequence data			
Ethical clearance to host sensitive data			
Data quality		Data use/analysis	
Data linkage			
Systematic data linkage			
Infrastructure		Data sharing	
Depositing in public data repositories			
Public data			
Open science			

Limited metadata				
Challenges with transferring large data files				
Timely data sharing				
Heterogeneous data sources		Data re-use		
Routine data				
Discoverability				
Data trustworthiness				
Data linkage	Data linkage			
Linking heterogeneous data sources				
Individual Health Data Linkage				
Linkage via regional correlation				
Linking variant classification information				
Ad-hoc data linkage				
Secure data linkage				
Pseudonymized identifier				
Data access permission				Privacy and data protection regulations
Data access application				
Pseudonymization				
Secure research environment				
Regulatory compliance				
Legal agreements				
Legal basis for data integration				

Vaccine effectiveness			Utility
Monitoring emerging viral variants		Analysis of viral variants	
Impact of viral variants on vaccine effectiveness			
Clinical impact of viral variants			
Impact of viral variants on reinfections			
Variant updates			
Exploring genetic diversity and evolution of pathogens			
Outbreak investigation		Analysis of viral transmission	
Virus transmission dynamics			
Clinical case studies		Clinical impact	
Research on long COVID			
Infectious disease control		Genomics-driven infectious disease management	
Genetic epidemiology			
Monitoring infectious diseases			
Genomic-informed public health practice			

SECTION IV

General discussion and recommendations

The primary objective of this thesis was to **develop, describe, demonstrate, and evaluate methodological approaches to integrate heterogeneous real-world data (RWD) sources at the Belgian and cross-jurisdictional border level**, enabling researchers to respond to emerging (causal) public health research questions in a more timely and accurate manner, thereby pursuing to provide insights that can enhance elements of pandemic preparedness and response strategies for potential future public health crises.

To achieve this objective, we demonstrated the effective integration of RWD to conduct public health research both within and across jurisdictional borders. More specifically, we successfully showcased the integration of heterogeneous data for policy-relevant causal research at the Belgian level ([Chapter 1](#), [Chapter 2](#)). During this process and by reviewing literature, we identified important challenges and potential biases that could impact the accuracy, reliability, and/or generalizability of the results. We formulated best practices to identify and mitigate potential biases arising from the study design, data analysis methods, underlying surveillance strategies and/or the data infrastructure ([Chapter 1](#)). Going beyond the jurisdictional borders of Belgium, we proposed a methodological framework to approach causal interference in a federated research infrastructure for cross-border research ([Chapter 3](#)) and we demonstrated this framework to address a policy-relevant research question by leveraging RWD from three countries ([Chapter 4](#)). Given the difficulties in integrating data from other fields, such as pathogen genomic and socio-economic data, we conducted a qualitative assessment of experts' experiences and perceptions towards integrating these data sources for public health research ([Chapter 5](#)).

Throughout this work, we identified challenges and requirements for successfully integrating heterogeneous RWD sources for policy-relevant public health research. We proposed approaches and formulated recommendations to mitigate potential biases and optimize the integration of RWD for observational public health research. This thesis's focus is facilitating more complex causal research; however, the approaches and recommendations are equally applicable to other types of observational research, such as studies examining correlations and trends. These methodologies and insights can thus enhance various observational research designs, making them applicable across a broad spectrum of public health inquiries. Finally, here we outline a way to enhance further the quality and efficiency of (causal) observational public health research, ensuring more robust and actionable insights for public health policy.

The results and insights from these different chapters are incorporated hereafter to present a broad approach and shared vision, and provide recommendations for future pandemic preparedness.

Table 1. Overview of the thesis objectives, expected outcomes and potential impact.

Objective:	Develop and describe methodological approaches to leverage, integrate and analyze heterogeneous RWD sources for public health research within and across jurisdictional borders	
Expected outcomes	Working methodologies	Prototyped workflows
Chapter(s)	Chapter 1, 3	Chapter 3
Expected impact	Provide insights and recommendations that can enhance elements of pandemic preparedness and response strategies for potential future public health crisis	
Obtained outcomes	<u>Chapter 1</u> : Insights and best practices to identify and mitigate potential biases when assessing the severity of infection with alternative viral variants	<u>Chapter 3</u> : A reproducible workflow, FAIR research objects
	<u>Chapter 3</u> : A methodological framework to develop federated cross-border (causal) observational studies in a privacy-preserving and interoperable way (step-by-step guidance)	
Objective:	Demonstrate and evaluate the methodological approaches to leverage, integrate and analyze heterogeneous RWD sources for observational public health research at the Belgian and cross-jurisdictional border level	
Expected outcomes	Scientific evidence	Preparedness
Chapter(s)	Chapter 1, 2, 4	Chapter 1-5
Expected impact	Expanding the evidence pool which can be consulted for decisive action	Identify targets for future action
Obtained outcomes	<u>Chapter 1</u> : The impact of different methodological choices when studying the severity of infection with SARS-CoV-2 variants was evaluated	By identifying existing gaps, this research puts forward potential targets for future action
	<u>Chapter 2</u> : The effectiveness of alternative SARS-CoV-2 vaccination schedules (i.e., mRNA booster vaccination <i>versus</i> primary vaccination only, and heterologous viral vector–mRNA <i>versus</i> homologous mRNA–mRNA prime-	

	boost vaccination) against severe outcomes of Omicron infection was evaluated among hospitalized COVID-19 patients in Belgium	
	<u>Chapter 4</u> : The effectiveness of primary vaccination in preventing SARS-CoV-2 infections was evaluated across three sites: Aragon (Spain), Brussels and Wallonia (Belgium), and Finland	

GENERAL DISCUSSION

1. OBSERVATIONAL STUDIES TO ADVANCE EVIDENCE-INFORMED PUBLIC HEALTH POLICY

Observational studies, conducted in real-world, uncontrolled settings, span a spectrum of designs that can have an important role in informing public health decision-making. Descriptive studies, such as health surveillance statistics, can provide a snapshot of the health of the population and monitor the spread of diseases. Alternatively, correlational studies examine relationships between various factors of interest and health outcomes, offering valuable insights without establishing causality. Within the broad spectrum of evidence that can inform public health policy, evidence from observational studies leveraging RWD to infer causality is particularly valuable and challenging [1,2]. These rather complex studies enable the identification of direct drivers of health outcomes or the comparative assessment of the effectiveness of interventions, and have the potential to yield actionable results [2]. In contrast to correlational relationships between exposures and health outcomes, where a change in one factor does not necessarily lead to a change in the other, causal relationships implicate that a change in exposure directly results in a change in the linked health outcome [3]. This direct link makes causal health determinants good targets for public health policy, as interventions can be designed to effectively alter these exposures and improve health outcomes. Causal real-world evidence can thus enable policy-makers to make informed decisions and design effective, targeted interventions, ultimately improving health outcomes and/or ensuring the efficient allocation of resources.

Several elements are important for (causal) public health research to inform policy-making effectively. The research should address relevant public health issues (i.e., *relevance*), providing evidence that is available in time to inform policy decisions (i.e., *timeliness*) [4,5]. Further, research should be of good *quality*, aware of potential biases, and provide accurate and generalizable results [4]. For findings to be applicable, they should be *communicated* in a clear, concise manner, conveying the core aspects and implications, and be *accessible* to policy-makers [4,5]. *Stakeholder engagement* (including policy-makers) throughout the research process is important to enhance the chance of translating resulting evidence into policy changes [4]. These factors jointly enhance the likelihood of research evidence being integrated into public health policy-making, ultimately leading to more informed and effective health policies.

Contributions of this thesis

Chapter 1 provides critical insights for observational causal public health research by proposing best practices to limit bias and ensure that researchers can choose the most accurate and context-appropriate approach when investigating the severity of infection with SARS-CoV-2 variants. These best practices are generalizable for future research on the severity of COVID-19 and other infectious diseases in the context of emerging viral variants, thereby facilitating future research on the topic. By inferring causal links between viral variants and disease outcomes (e.g., infection-related hospitalization) in a timely manner, policy-makers can implement targeted interventions. For example, if a particular viral variant is found to directly increase hospitalization rates [6,7], policy-makers can respond by increasing hospital capacity or encouraging measures preventing infections like social distancing or the wearing of face masks [8,9]. This work contributes to guiding the future generation of *high quality and timely epidemiological evidence*, thereby enhancing the ability of public health authorities to make informed decisions, ensuring preparedness and a rapid response to future waves of variants of COVID-19 or other infectious diseases. Further, by effectively integrating RWD (i.e., existing COVID-19 registries) to assess the severity of infection with the SARS-CoV-2 Omicron variant compared to the Delta variant in a population of hospitalized COVID-19 patients under alternative approaches, *robust causal evidence* was generated.

Similarly, in Chapter 2 we illustrated the integration and analysis of vertically-partitioned RWD for observational public health research in Belgium. This work provides *relevant causal evidence for policy-makers* by evaluating how different vaccination strategies (i.e., homologous and heterologous prime-boost vaccination schemes) affect the clinical outcomes of hospitalized COVID-19 patients infected with the Omicron variant. Highlighting the most effective vaccination policies can guide public health decisions on vaccine distribution, prioritization, and communication strategies, improving future public health actions and enhancing overall vaccine effectiveness. By integrating data from different registries, detailed characteristics were available for patients in the study population. This enabled us to address for potential confounding bias and obtain more *reliable* causal inferences.

A methodological framework to address (causal) cross-border research questions by conducting observational research was additionally proposed in Chapter 3, likewise providing guidance to facilitate the execution of *timely and accurate policy-relevant (causal) research* when working with RWD. This approach addresses the most complex scenario where researchers seek to infer causality, yet also serves as a valuable guide for other observational designs that require integrating distributed RWD. Additionally, we demonstrated this approach in Chapter 3 and Chapter 4, illustrating the integration of RWD distributed across different systems and jurisdictional borders to investigate a causal cross-

border policy-relevant research question. The study *evaluated past efforts* by addressing the effectiveness of vaccination across regions or countries, and can *guide future actions*. This evidence enables policy-makers to make informed decisions about optimizing vaccination campaigns. The results were made *readily available* in the Zenodo repository and can be reproduced by deploying the analytical pipeline published in Github [10].

Challenges encountered throughout the work

It is important to note that, even when selecting the best possible approach in a specific context, where you are minimizing selection biases and carefully addressing confounding, causal estimates obtained from real-world observational studies can still be biased, e.g., resulting from residual confounding or measurement errors that aren't white noise. For example, throughout the work presented in this thesis, we were repeatedly confronted with missing information for certain identified confounders (e.g., host genetics in [Chapter 1](#) and [Chapter 2](#), individual-level socio-economic status in [Chapter 4](#)). It is important to acknowledge the limitations of both the data and adopted approach, and communicate potential residual biases, such as the unobserved confounding and selection bias originating from different selection steps to obtain the study population.

The research conducted in this thesis lacked interaction with policy frameworks and directed communication towards decision-makers. This limitation hindered the effective translation of findings into actionable insights guiding public health policy interventions. Further, the research was not consistently produced in a timely manner, often missing opportunities to inform real-time policy decisions. The absence of these elements reduced the potential impact of the generated evidence, highlighting the need for more integrated approaches involving active collaboration with policy-makers.

Way forward

Integrating heterogeneous data sources for observational studies holds a great potential to provide policy-relevant research evidence. Still, future efforts must address key methodological challenges to enhance the reliability and generalizability of findings. One critical step is the application of sensitivity analyses, evaluating how much findings are affected by residual bias (e.g., residual confounding) and the robustness of results under various assumptions or approaches [11,12], evaluating whether they hold true under alternative conditions. Transparency is essential for policy-makers in taking informed decisions, as clearly communicating the limitations of the data and the study design allows policy-makers to make more informed decision on how to include the provided evidence. Causal diagrams, such as Directed Acyclic Graphs (DAGs), can promote transparency by visually representing

researchers' assumptions about the relationships between variables and derived presumed biases [13]. Similarly, Data Quality Assessments (DQAs) can play an important role in increasing transparency by evaluating and reporting the quality (e.g., completeness, validity, uniqueness) of the data used in analyses. By systematically reporting data quality, researchers can provide a clearer picture of the strengths and limitations of their evidence, thereby improving its interpretability and trustworthiness.

To bridge the 'evidence-policy gap', capacity building is essential, for researchers to effectively communicate their findings and for policy-makers to understand and interpret the implications and limitations of scientific results [4], especially in the context of integrative causal research using complex statistical approaches. In their review of barriers and facilitators of evidence-based policy, Oliver et al. (2014) highlight the need for two-way interaction between researchers and policy-makers [14]. Preventing that researchers are isolated from the policy process is critical to identify relevant research topics and creating actionable evidence. They also emphasize the need for the integration of a broad range of evidence [14]. Triangulation of evidence (i.e., the use of diverse data sources, studies, and approaches) offers a more comprehensive view and allows for the cross-validation of findings, which strengthens the evidence and reduces the reliance on any single, potentially biased data source or estimate [1,15–17].

A cautious approach to evidence-based policy implementation is warranted when research evidence is limited, uncertain and potentially biased. This could involve pilot programs or phased rollouts [18], and ensuring that policies remain flexible and adaptable as more reliable evidence becomes available.

It is important to recognize that well-founded policy advice, may not translate into actionable policy if it conflicts with governmental priorities, economic constraints, or existing conflicts of interest. For example, research may strongly support implementing higher taxes on alcoholic beverages to reduce consumption and health-related consequences, such as liver disease and the increased risk of certain cancers associated with alcohol consumption [19–21]. However, such a policy might face resistance if it clashes with economic interests, such as revenue generated from this industry and concerns about the impact on local business and jobs, or if there might be pressure from industry lobbying groups [19]. It could also be that governments are prioritizing other political or economic goals at the time. This highlights the complex interplay between evidence-based research and the practical realities of policy making.

2. LEVERAGING VERTICALLY-PARTITIONED REAL-WORLD DATA FOR PUBLIC

HEALTH RESEARCH WITHIN JURISDICTIONAL BORDERS

In Belgium, individual-level health and health-related observational data tend to be distributed across different systems or institutions (e.g., across public health agencies like Sciensano, hospitals, or research centers) [22–24]. By integrating such scattered data, datasets containing diverse characteristics of an individual can be obtained.

Given that most public health issues are multifactorial constructs, with multiple contributing factors or component causes that elevate risks and with a variety of causal pathways, the effective integration of diverse scattered data sources is often essential to achieve a comprehensive understanding of various health determinants [3]. Public health policy-makers can acquire a more holistic view of public health issues by harmonizing, integrating, and analyzing heterogeneous observational data sources. Further, inferring reliable (causal) estimates using RWD necessitates addressing for potential biases, which if not properly addressed, can distort observed associations between an exposure and health outcome (i.e., affecting their internal validity) or limit their generalizability to the broader population (i.e., affecting their external validity) [25,26]. Detailed individual-level information is generally required to properly address biases (which are inherent to observational data). By integrating heterogeneous RWD from different fields, evidence-informed decision-making can become responsive to the complex and dynamic nature of public health challenges.

The use of vertically-partitioned RWD for (causal) public health research displays both significant opportunities and challenges.

Contributions of this thesis

In [Chapter 1](#) and [Chapter 2](#), we were able to demonstrate the successful integration and secondary use of existing RWD sources for policy-relevant causal public health research within the jurisdictional borders of Belgium.

In these studies, data on characteristics of individuals in the study population (i.e., comprising hospitalized COVID-19 patients in Belgium) was originally vertically partitioned across distributed data sources, each containing different features of the same set of individuals. However, to address the illustrative causal research questions (i.e., assessing the severity of infections with the SARS-CoV-2 Omicron variant compared to the Delta variant among hospitalized COVID-19 patients in Belgium in [Chapter 1](#), and assessing the impact of homologous versus heterologous prime-boost vaccination on clinical severity of SARS-CoV-2 Omicron infection among hospitalized COVID-19 patients in Belgium in

Chapter 2), the integration of different registries was required. In particular, it was important to enable assigning exposure and outcome, approximate confounding factors by using information available in existing registries and assess the heterogeneity of the treatment effect. Essentially, by leveraging and effectively integrating clinical, vaccination, epidemiological, and socio-economic data, we were able to investigate these policy-relevant causal research questions on the severity of infection with SARS-CoV-2 variants and vaccine effectiveness respectively.

Challenges encountered throughout the work

While we successfully integrated heterogeneous RWD sources for causal policy-relevant public health research in Belgium, challenges encountered throughout the research process in Chapter 1 and Chapter 2, and brought forward in Chapter 5, pose potential obstacles to the efficiency and timeliness of future integrative (causal) research efforts, and may limit the ability to generate results with both high internal and external validity.

(Timely) data integration

As discussed earlier, integrating data from diverse sources is crucial for gaining a comprehensive understanding of public health issues. It also plays a key role in conducting robust causal inference studies using observational data. When health and health-related data are siloed or fragmented across different systems or institutions, their utility for research is strongly limited. Integrating data in a timely manner can in turn, facilitate the generation of timely (causal) evidence to inform public health decisions.

To address the research questions outlined in Chapter 1 and 2, information was obtained by linking existing data sources within a secured pseudonymized research environment established as part of the LINK-VACC project (initial project duration: Jan 2021 – Jan 2024) [27]. This project, along with the corresponding secured research environment, was developed by Sciensano (i.e., the Belgian public health agency) [28] following the start of the COVID-19 crisis in 2020. Setting up the LINKVACC project demanded the drafting of a study protocol, submission and approval of this protocol by the medical ethics committee of the UZ Brussels university hospital – Vrije Universiteit Brussel (VUB), and authorization from the Information Security Committee (ISC) [27]. Establishing this project, going through these data collection, access, and linkage procedures, was not part of the work conducted in this thesis and precluded it in time. As researchers at Sciensano, we were granted access to conduct these studies that aligned with the objectives of the LINK-VACC project. Within the secured pseudonymized environment of the LINK-VACC project, the different data sources could be linked based on a pseudonymized person identifier, which was available in the different registries. This

availability of a unique identifier for each Belgian citizen [29] allowed for the implementation of a reliable deterministic linkage approach, substantially reducing the likelihood of linkage errors [30,31].

The research conducted in [Chapters 1 and 2](#) was expedited by having the authorization to link different pseudonymized data registries and conduct analyses within a secure research environment that was already in place in the context of the LINK-VACC project [27]. However, generally speaking, *ad hoc* procedures are required to set up data linkage projects, going through different laborious steps for every research question or topic that emerges in order to be compliant with legal and ethical requirements (i.e., from conceptualizing the project, to authorization requests, and technical and legal implementation) and involving collaboration between various stakeholders [24,31,32]. Navigating the process of obtaining ethical and legal authorization to conduct data linkages for research tends to be time-consuming and challenging [31]. Further, outlining responsibilities and data protection measures, like pseudonymization or linkage and de-identification by a Trusted Third Party (TTP) (i.e., separating identifiers from data to reduce the risk of re-identification), requires establishing agreements between all involved partners, which can further slow down the overall process of setting up a data linkage for research [31]. These aspects potentially impede the timely assessment of emerging (causal) research questions in crisis settings, where new issues can arise promptly and may require an immediate response (e.g., during emergency epidemic responses).

That data governance challenges can impede the efficient integration of distributed RWD sources for public health research was also brought forward from qualitative analyses in [Chapter 5](#), where experts' experiences and perceptions towards the integration of socio-economic and genomic data for public health research were gathered. Alternatively, competing research interests between data owners was identified in [Chapter 5](#) as a cause for reluctance toward constructive collaborations, hindering effective data integration and undermining the potential of these data to gain causal public health insights that rely on integrated datasets. Other barriers like a lack of policy support or compliance anxiety were identified. It was noted that during the COVID-19 crisis, there was greater willingness, with institutions being less guided by fear of compliance. Processes were speeded up, as data linkage projects' immediate public health benefits were clear, and there was a strong sense of urgency. However, after the COVID-19 crisis cautiousness increased, particularly as different stakeholders raised privacy and data protection concerns. Balancing the utility of data linkages for public health with ethical and legal considerations is crucial to ensure that valuable insights can be gained to improve the health of populations while safeguarding individual privacy and maintaining public trust.

Data discovery

Diseases often have a multifactorial pathogenesis, meaning that multiple factors contribute to disease outcomes rather than a single causal factor. As such, investigating these health outcomes requires the consideration of data from various research fields, such as genetic, environmental, socio-economic, and behavioral sciences. Integrating and analyzing multidisciplinary data from different fields can be challenging as a researcher.

During the COVID-19 crisis and in the context of the LINK-VACC project, different teams within Sciensano, working on alternative COVID-19 registries, collaborated intensely to enhance the overall understanding of the data in the existing registries. This facilitated the work that was conducted in [Chapter 1](#) and [Chapter 2](#). However, without clear documentation or shared expert knowledge, researchers can face difficulties in understanding sampling and data collection procedures, the context in which this was executed, data management procedures, variable measurement and definitions, and potential biases resulting from the former aspects. This lack of transparency can make it difficult to evaluate the accuracy and completeness of the data, increasing the risk for the use of unsuitable statistical methods and misinterpretation of causal estimates. This was also highlighted in [Chapter 5](#), where the difficulty to identify suitable socio-economic variables, indicators and data sources when not being an expert in this field was pointed out.

Data availability and quality

Our research encountered notable limitations stemming from the availability of data, but also from challenges related to its quality (more specifically, missingness), which had significant implications for addressing specific biases and drawing certain conclusions. These constraints underscored the complexities of working with RWD and the importance of critically evaluating its suitability for specific analytical objectives.

For the analyses conducted in [Chapter 1](#) and [Chapter 2](#), we relied on data recorded in the Clinical Hospital Surveillance (CHS) system to obtain information on disease severity indicators of COVID-19 patients. The CHS was set up following the start of the COVID-19 epidemic in Belgium, with the purpose of collecting specific characteristics of hospitalized COVID-19 patients (e.g., risk factors such as demographic or comorbidity characteristics, disease outcomes such as the length of hospital stay or ICU admission). This system relied on individual-level information collected by hospitals at admission and discharge [33]. These data allowed us to gain information of different clinical severity outcomes of COVID-19, including a composite measure (i.e., severe COVID-19, defined as experiencing an acute respiratory distress syndrome (ARDS) event, or an ICU admission, or in-hospital mortality), a severity

measure dependent on clinical decision making (i.e., ICU admission), and a ‘hard’ objective severity measure (i.e., in-hospital mortality).

However, the CHS also faces limitations. Due to non-mandatory reporting, the CHS operates as a non-exhaustive surveillance system, with data not being reported by all Belgian hospitals [33]. Consequently, the data did not represent the full population of hospitalized COVID-19 cases. For example, due to its non-exhaustive design the lack of a recorded ICU admission in the CHS did not necessarily exclude the actual admission of a recorded COVID-19 case to the ICU. This limitation rendered the data unsuitable for making inferences for the general population, as it did not provide a complete or representative picture of all hospitalizations and associated outcomes.

The structure of the CHS data collection also introduced constraints. The data collection process relies on information documented at admission (e.g., demographic characteristics, pre-existing comorbidities, diagnostic information) and discharge (e.g., severity indicators, treatment information, discharge information) of patients, which primarily consists of administrative clinical information recorded as part of routine patient care [33]. This method of data collection has inherent limitations, as it does not include repeated measurements of the exposure, confounding variables, or the outcome over time. As a result, it is not possible to track changes in these variables over the course of a patient’s disease trajectory. This lack of longitudinal data precluded us from constructing a time-varying framework or dynamically adjust for confounding. The analysis was restricted to static snapshots of information (i.e., cross-sectional).

Beyond CHS-specific challenges, integrating multiple data sources for the analyses in [Chapter 1](#) and [Chapter 2](#) revealed further challenges. Several datasets contained missing values in critical variables, which, if unaddressed, could compromise the validity of causal inferences. To mitigate the impact of missing values in essential variables, we employed multiple imputation, i.e., a statistical method that creates multiple probable datasets, each with different estimates for the missing values, predicted based on the observed values [34,35]. These imputed datasets were then analyzed separately, followed by pooling the results to produce a more robust and accurate estimate, minimizing bias and maintaining the variability inherent to the data [35]. While this approach helps minimize bias and preserves data variability, there are also some limitations of using a multiple imputation approach to handle missingness, such as the reliance on making assumptions about the mechanisms of missingness, being prone to model specification (with improper model specification potentially leading to inaccurate prediction of missing values), and not performing well with high proportions of missing values [36].

In addition to missing values in critical variables, certain key variables, such as host genetic information, were entirely absent from the available datasets. Host genetics had been identified as a confounder for the exposure-outcome relationships examined in [Chapter 1](#) and [Chapter 2](#). The unavailability of such data likely resulted in residual confounding, where spurious associations persist, even after attempts to control for measured confounders.

To visualize hypothesized relationships between variables, identify potential biases, and guide decisions on which variables to control (i.e., those identified in a ‘minimum sufficient adjustment set’), we relied on Directed Acyclic Graphs (DAGs). However, these graphical representations are based on assumptions of researchers, which are fed by their experience, expertise, and prior evidence. This gap leaves room for unidentified biases, further complicating the validity of causal estimates. In addition, even when confounders are identified using DAGs, the reliance on existing RWD registries for information means that the measured variables might not fully encapsulate the intended constructs, leaving the potential for residual confounding to be unaddressed.

In general, RWD present recurring challenges stemming from missing or incomplete information, which can arise from various factors such as nonresponse, human error, inconsistent data entry, system limitations, or technical issues [37]. The resulting data quality concerns can weaken both the internal and external validity of causal inferences. Addressing these challenges requires not only methodological validity, but also ongoing efforts to maintain and improve data collection practices. The careful application of statistical methods, alongside critical evaluation of data limitations, is essential for deriving meaningful insights from imperfect datasets.

Potential selection bias

In [Chapter 1](#) and [Chapter 2](#), the study population was limited to hospitalized COVID-19 patients in Belgium for whom clinical data were registered in the CHS.

Eligibility for inclusion in these studies was determined based on events occurring after intervention assignment, namely based on the occurrence of COVID-19 hospitalization. This post-intervention event may itself have been influenced by the exposure/intervention (i.e., infection with a particular variant of interest in [Chapter 1](#), or completing a specific vaccination scheme in [Chapter 2](#)), creating the potential for collider bias, selecting a subgroup whose outcomes may not represent the general population [38,39]. This could introduce spurious associations when making inferences for this target population [38,39]. Hospitalized patients represent only the more severe cases of COVID-19, with this selection bias likely resulting in the underestimation of the differences between the severity of disease caused by alternative variants or vaccination schedules in the general population. For example, if one

variant (e.g., the Delta variant) caused more severe COVID-19 infections in a population compared to another variant (e.g., the Omicron variant), this difference may have been diminished within a population of hospitalized patients. This is because individuals infected with the milder variant (e.g., the Omicron variant) must have exhibited more severe disease outcomes to be hospitalized, effectively making Omicron infections appear artificially more severe. Selection into hospitalization can distort the relationship between exposure and infection severity, with the observed differences in the hospitalized study population expected to be more conservative than the effect size in the general population. However, it is important to note that the suitability of the data and adopted approach depends on the specific research question being addressed, namely whether we are investigating (a) the severity of infection once hospitalized (i.e., in a hospitalized population) or (b) the severity of infection overall (i.e., in the general population). In Chapters 1 and 2, we focus on the former (i.e., a). We do not claim to address the latter (i.e., b), nor do we assert that the findings are generalizable to that broader context.

Meanwhile, information on the variant of infection was required to identify the exposure to a certain SARS-CoV-2 variant in Chapter 1, and to investigate the Omicron sub-lineage of infection as a potential effect modifier in Chapter 2. However, by the limited coverage of samples that were characterized in the baseline surveillance (i.e., randomly selecting samples to be sequenced, covering about 5-10 % of the positive COVID-19 test samples) [40], the sample sizes obtained for analysis were reduced significantly.

The limited sample sizes, together with potential selection bias introduced from non-random selection steps and collider bias introduced by post-intervention selection, could impact both the internal and external validity of obtained causal estimates, limiting the generalizability and precision. As such, it is important to exercise caution when interpreting the resulting association estimate in causal terms for the general population. While these finding should be interpreted as indicative of associations rather than causal effects when considering the general population, emphasizing the need for further research to validate these results in these settings, they can provide valuable insights into the risk of severe clinical progression associated with a viral variant or vaccination scheme once COVID-19 patients are hospitalized.

To make inferences for the general population, we would need a representative sample of the general population with detailed information on the severity of COVID-19 disease available to study these research questions. However, considering the limitation of data availability, it could still be possible to derive insights by the use of statistical techniques, e.g., Inverse Probability Weighting (IPW) [41]. IPW can adjust for selection bias by weighting individuals based on the inverse of the probability of being

selected in the study population (here the inverse probability of being hospitalized) [41,42]. Hospitalized patients could according to this strategy be reweighted to present the general population, including the not included non-hospitalized COVID-19 cases. This approach includes (1) modeling the probability of hospitalization based on covariates associated with selection (such as age, comorbidities, vaccination status, etc.), (2) assigning a weight to each hospitalized patient (equal to the inverse of their probability of hospitalization), and (3) using the weights in the analysis to estimate the causal effect of interest in the general population [41]. If executed appropriately, IPW allows the generation of a pseudo-population, that could be used to mimic the general population, reducing selection bias induced by restricting the study population to hospitalized patients [41]. This could enable the estimation of (causal) relationships that closely approximate those that would be observed in target population if a full set of samples were available for analysis [41]. However, it is important to note that this approach requires an accurate estimation of selection probabilities (here the probability of being hospitalized). Further, IPW techniques demand that the factors influencing the selection are available for both participants and nonparticipants, to enable the estimation of the probability of being selected and the calculation of weights [41]. Likewise, IPW does not account for unknown or unmeasured factors that have an effect on the selection into the study population [41].

Confounding adjustment

Obtaining valid causal estimates from observational data requires adjusting for confounding factors to approach comparability between exposure groups. To adjust for confounding in the studies described in [Chapter 1](#) and [Chapter 2](#), we employed regression standardization, a technique that involves constructing a model to estimate the risk of the outcome under different exposure conditions [35,36]. This approach allowed us to account for identified and measured confounding factors, by predicting what the outcome would have been had all individuals in the study population been exposed to a level or category of the exposure [43]. While regression standardization is a valuable method for causal inference, it is not without limitations. A major concern is the potential miss-specification of the outcome model, which would result in biased estimates [43,44].

In addition, during periods of rapidly changing contextual factors (such as shifting testing strategies, social behaviors, and immunization degrees during the COVID-19 crisis), accurately adjusting for biases and inferring valid causal estimates becomes increasingly challenging [7].

Target trial emulation

In [Chapter 2](#), the study was designed to emulate a target trial, aiming to imitate the design and analytical rigor of a RCT using observational data [45]. However, limitations presented by the available

data raise questions about whether a target trial can be effectively emulated and whether meaningful causal estimated can be reliably obtained.

Emulating target trials using real-world observational data presents significant challenges [46]. These data are not collected with the explicit intent of answering specific research questions. Consequently, key information generally required for target trial emulation, such as precise timings of exposures and outcomes, detailed confounder measurements, or repeated and consistent follow-up times, is often missing or incomplete [46]. Moreover, observational datasets frequently only capture a subset of the target population, and are typically not designed to ensure a representative sample. These challenges highlight the need for careful design, robust methodological approaches to address potential biases, and transparent reporting when emulating target trials using RWD.

Way forward

While the integration of RWD within jurisdictional borders has enabled significant advancements in generating insights, challenges (such as those related to data availability and quality, data governance, lack of documentation, and confounder adjustment) can in the future be addressed to improve the robustness and timeliness of (causal) observational evidence that can guide efficient public health policy.

Surveillance strategies

By conducting thoughtful and structured surveillance, the utility of RWD in observational research can significantly be enhanced.

One example of how effective surveillance can have an impact on the utility of data for observational research, is genomic surveillance. The European Centre for Disease Prevention and Control (ECDC) has published a document providing guidance on implementing genomic surveillance for EU/EEA Member States [47]. Two complementary sampling strategies for genomic surveillance of SARS-CoV-2 were recommended by ECDC: (1) representative selection of RT-PCR positive cases from established population-based surveillance systems and (2) targeted selection of RT-PCR positive cases in specific settings or populations (e.g., from individuals with a travel history or a severe disease outcome). Different options exist for specimen selection as part of representative sampling of SARS-CoV-2 cases for genomic monitoring, each with their own advantages and drawbacks: systematic/consecutive sampling, alternative systematic/consecutive sampling, simple random sampling, and alternative simple random sampling [47]. For either option, selection should be consistent over time (i.e., a high enough sequencing coverage), across geographic regions (i.e., coverage across all regions or other administrative areas), and across all levels of the healthcare system [47]. In the case of representative

sampling for monitoring (monitoring prevalence and spread of known variants), ECDC recommends sequencing adequate numbers of samples to detect a difference in the relative proportion of a certain variant from one unit of time to the next (i.e., 2.5-5%). It is essential to establish clear and consistent criteria for selecting samples for sequencing, in this way aiming to avoid ad hoc or convenience-based selection. Further, it is important to ensure that sequencing processes are not disproportionately influenced by technical and logistical factors, such as unequal access to sequencing resources or sample quality issues [48]. Standardized sample collection and testing protocols and comparable access to sequencing resources across regions and populations can be of help. Lastly, it is essential to monitor the representativeness of the sampling approach regularly [49].

Balancing the utility of data for re-use and cost of implementation is an important consideration in developing sustainable surveillance systems. The utility of data depends on its accuracy, completeness, timeliness, and representativeness, all of which demand ongoing efforts in data collection and curation [50]. At the same time, overly expensive or complex systems may become unsustainable or inaccessible. Prioritizing scalable solutions, and employing cost-effective technologies can help maximize data re-use while minimizing expenses [51].

Regardless, it is essential to thoroughly document surveillance strategies and consider them during analysis to minimize bias, and when bias is present, appropriately address it in the analysis and interpretation of the data [47].

Missing data

Preventing and effectively addressing missing data in RWD requires a broad approach. Initially, proactive measures should be taken to reduce the intrusion of missing data through improved and standardized data collection procedure. Standard Operating Procedures (SOPs) for data collection can play a role in this, providing detailed guidelines that standardize how data should be gathered, recorded, and managed across systems [52]. By adhering to these pre-defined procedures, data collectors can reduce variability and avoid errors that can arise from inconsistent data collection methods, thereby contributing to improved data quality and reduced missingness [53,54]. Additionally, investing in capacity building for individuals collecting data (e.g., healthcare professionals) is important, equipping them with the skills and knowledge to adhere to these procedures and to recognize potential sources of missing data early on [55,56]. Along with that, communicating the benefits of high-quality RWD for public health and clinical practice can enhance the motivation of collectors to adhere to best practices [57]. By fostering a culture of attention to detail in data collection, RWD's overall quality and completeness can significantly improve.

Another risk in data collection is the oversight of key variables for research. Minimal datasets can play an important role in reducing this risk by establishing minimal data collection standards, e.g., specifying variables that must be consistently collected across institutions or studies, helping to address the common issue of missing vital information [58–60]. Establishing a core set of standardized variables in minimal datasets, ensures that the most critical information is captured uniformly, reducing the risk of incomplete or inconsistent data, protecting against the absence of vital information, and ensuring that researchers have access to the foundational data necessary to draw accurate conclusions [58,59,61]. Further, the concept of minimal datasets aligns with data minimization principles, ensuring that only data required for the intended purposes is collected. However, establishing these minimal datasets across diverse fields, purposes, and research questions can be challenging. One difficulty lies in identifying and prioritizing variables that are universally relevant yet flexible enough to address the unique requirements of specific research domains. For example, the Minimal Dataset for Cancer developed under the 1+Million Genomes Initiative focusses on capturing standardized genomic and clinical data for cancer research [62], but adapting this dataset to align with the needs of different cancer (sub)types, disease stages, or treatment modalities can be complex [62]. Developing minimal datasets requires a thorough, inclusive, and iterative approach, that actively involves multidisciplinary stakeholders [62]. These stakeholders can bring diverse perspectives that are essential for ensuring that minimal datasets are both practical and aligned with the needs of end users. It is important to acknowledge that a single universal minimal dataset cannot address the wide array of research objectives across all domains. Instead, tailored minimal datasets that reflect the specific needs and contexts of individual fields are necessary. Furthermore, data requirements are not static. They evolve as scientific knowledge, technologies, and priorities change over time. These considerations highlight the importance of continuing collaboration and building frameworks that support the dynamic and context-specific nature of research.

Once missing values in measured variables occur, it is equally important to address them using appropriate statistical methods, such as multiple imputation, to limit the impact on the results and maintain the integrity of the research [36]. In recent years, tools using artificial intelligence have emerged, presenting promising approaches to data imputation [63]. Machine Learning (ML) algorithms such as K-nearest neighbors (KNN) and decision tree (DT) can be employed to predict missing values based on patterns within the dataset [63–65], often outperforming traditional imputation techniques in capturing complex relationships [66]. By leveraging these capabilities, AI-based methods have the potential to not only improve the accuracy of imputations, but also enhance the reliability of downstream analyses.

Governance

Proper governance structures are essential for effectively leveraging RWD in a timely and efficient manner.

First, developing and implementing standardized and verified procedures for initiating and conducting data linkages is essential to ensure transparency and consistency throughout the research process. This involves creating a clear framework that outlines the necessary steps for setting up data linkages and ensuring that all involved parties adhere to the uniform guidelines. These guidelines should cover the whole process, involving study conceptualization, ethical considerations, data security measures, and quality checks. Standardized and verified procedures can streamline and speed up the process for researchers, alongside preventing errors and discrepancies that may emerge from irregular practices for handling data. Uniform procedures can additionally stimulate trust and strong collaboration between stakeholders, such as data owners, researchers, and policy-makers. Further, setting a maximum response time for institutions to respond to applications is recommended, alongside the requirement of sufficient workforce to process such applications in a timely manner. Next to *ad hoc* (or project-based) data linkages, requiring each project to go through laborious and time-consuming procedures and potentially taking months or even years to establish, routine (or systematic) data linkages offer a proactive alternative [31]. Routinely linking data could support researchers to generate timely evidence, by setting up data linkage systems that are readily available for research projects. In several high-income countries, advanced data linkage platforms have been set up, offering the ability to integrate diverse datasets in near real-time. For example, the Secure Anonymized Information Linkage (SAIL) Databank, containing de-identified linked data covering the Wales population, provides an exemplary model of secure data linkage [67–69]. Secure remote access to the data by approved users for research can be provided through the SAIL Gateway [70]. During the COVID-19 crisis, two population-level cohorts were created in Wales to conduct near real-time analytics, using linked data from the SAIL Databank [71]. Similarly, Findata, Finland’s data permit authority for social and health data, offers authorized researchers secure access to linked healthcare and genomic data in the population-wide FinRegistry [72,73], enabling detailed studies on variant transmissibility, vaccine effectiveness, and health disparities. The application process to request a data permit involving the FinRegistry tends to be shorter and less expensive, than the process of requesting a regular data permit [73]. This is because separate requests with the data controllers can be avoided [73]. The establishment of these advanced data linkage platforms highlight the significant strides that have already been made in harnessing RWD for public health research. Initiatives like SAIL and Findata demonstrate how streamlined data governance, secure access mechanisms, and innovative

infrastructure can accelerate research, while still complying with data privacy and ethical principles. Population-based cohorts enabled by such platforms have already played a pivotal role in addressing public health challenges in the past [71,74,75], and can provide important opportunities for future research [24]. Moving forward, these platforms can serve as a model for other regions aiming to leverage RWD for public health purposes.

Political willingness was identified as being a critical enabler for successful data integration for public health research. Without strong political support, efforts to integrate data that are distributed across institutions can be obstructed by legal, financial and logistical barriers. Political engagement is important for providing the necessary resources and introducing legislative changes that may be required to overcome privacy and data protection regulations or bureaucratic obstacles that restrict data linkage [76,77]. Further, political leadership can foster collaboration between stakeholders, like healthcare providers, government agencies, and research institutions. Political support is essentially required to establish an environment where data integration is not only feasible but also prioritized to facilitate essential public health research [76].

In Belgium, establishment of the federal Health Data Agency (HDA) in 2023 reflects a growing political commitment towards harnessing the potential of RWD for observational public health research [78,79]. The HDA was created to address longstanding challenges in identifying, accessing, and combining diverse health data sources. One of its primary objectives is to streamline the often complex and time-consuming processes to access health data [78]. At the same time, the HDA ensures adherence to privacy and data protection regulations, upholding the ethical standards necessary for public trust [45]. This new initiative, though still in its early stages, demonstrates the government's recognition of the critical role that high-quality, well-linked health data play in tackling complex public health challenges and represents a crucial milestone in Belgium's journey towards establishing a more data-informed healthcare system.

Data discovery and data provenance

By evaluating the accuracy of RWD, researchers can more effectively acknowledge and address biases and limitations, ultimately leading to more reliable (causal) estimates. However, detailed documentation to assess the accuracy of data is often lacking. High-quality, standardized metadata and data provenance information have an important role in addressing the lack of documentation often encountered when leveraging existing real-world data for secondary use in research. Metadata is information accompanying datasets, describing the data (e.g., their structure, content) and providing details on their collection [81,82]. Data provenance represents a type of metadata, tracking a data

object's (physical and digital) history, documenting its journey from collection to its state used for analysis [83,84]. This includes documenting data manipulations or data processing steps that the data have undergone, revealing how changes in one part of the process can impact others, enhancing reproducibility, and allowing the monitoring of data quality [85]. Broad adoption of a standardized provenance framework, like the Common Provenance Model [85], could significantly enhance researchers understanding of RWD and its limitations. In combination, metadata and data provenance information guarantee transparency and reproducibility, enabling researchers to trace back and verify the integrity of the data. As a consequence, it can help mitigate biases, enhancing the reliability of research findings and supporting more robust causal inference.

Confounder adjustment

Accurate adjustment for measured confounders is essential to obtain valid (causal) estimates. However, this process can be challenging due to potentially complex relationships among explanatory variables. Traditional methods like regression standardization (i.e., constructing an outcome model) or propensity score matching (i.e., constructing a propensity score model) are commonly used [43]. More sophisticated methods, however, can advance the execution of confounder adjustment, better addressing these complexities. Double-robust methods, combining both outcome and propensity score modelling, offer protection against model misspecification (requiring only one of the models to be to be 'valid' to obtain an unbiased effect estimate under the strong assumption of there being no unmeasured confounders) [86,87]. Machine Learning (ML) approaches can further enhance causal estimation by flexibly modeling complex relationships in the data [88]. Among these, Targeted Maximum Likelihood Estimation (TMLE), an advanced technique first described by van der Laan and Rubin (2006) [89], combines ML and traditional statistical models to provide robust estimations. TMLE offers properties such as double-robustness, minimized assumptions on the distribution of the data and an optimized estimation process, in this way minimizing bias from measured confounders and achieving estimates closer to the target value compared to classic parametric models [90]. It is also important to acknowledge the potential existence of unmeasured confounding. E-values provide a measure quantifying the degree to which obtained causal estimates may still be influenced by unmeasured or unconsidered confounding in observational studies [91].

3. LEVERAGING HORIZONTALLY- AND VERTICALLY-PARTITIONED REAL-WORLD

DATA FOR PUBLIC HEALTH RESEARCH ACROSS JURISDICTIONAL BORDERS

To conduct policy-relevant, cross-jurisdictional, observational public health research, existing RWD are typically stored in decentralized systems. For cross-border research, in addition to data being scattered across diverse sources and parties, they are also distributed across these jurisdictional borders (e.g., across several regions or countries). Leveraging these data for observational public health research reveals unique challenges, particularly due to the limitations imposed by the inability to centralize data and heterogeneity across jurisdictions.

Contributions of this thesis

Federated analysis, in which algorithms or workflows are distributed and applied to local datasets without sharing data, offers a promising solution to address cross-border research questions, with vertically-partitioned data distributed across different jurisdictions [92,93]. The federated approach ensures that (sensitive) personal data remain secure under supervision of the data controller, ensuring data privacy and regulatory compliance by design, while maximizing the utility of cross-national data for valuable observational public health research [92,93]. However, comprehensive guidance on the full methodological process for approaching causal inference in a federated context was still lacking, including specifying data requirements and ensuring interoperability in federated research.

To fill this gap, a step-by-step methodological framework to approach cross-national causal inference using RWD was described in [Chapter 3](#), starting 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. This approach allows to address public health research questions in observational studies in a privacy-preserving way, while tackling interoperability challenges. It provides guidelines to address complex causal research questions, while supporting other observational study designs. By clearly outlining this framework, we aimed to equip researchers to respond more rapidly to emerging public health questions.

As described in [Chapter 3](#) and [Chapter 4](#), we demonstrated the proposed methodological framework by effectively integrating RWD in three nations to investigate a policy-relevant research question, namely the real-world effectiveness of primary vaccination campaigns in preventing SARS-CoV-2 infections. The study emulated conditions of a randomized controlled trial in a real-world setting, enhancing the reliability of its findings by minimizing avoidable biases. Throughout this process, we

also demonstrated the FAIRification of research objects generated throughout the process (i.e., making objects like data or workflows Findable, Accessible, Interoperable and Reusable), which is essential to improve efficiency, reproducibility, and transparency of scientific research, making research outcomes more reliable to guide public health policy [94,95]. The FAIR principles aim to avoid the duplication of efforts and promote collaboration and knowledge exchange [96]. To achieve findability, robust human- and machine-readable metadata are essential [96]. Many research objects lack comprehensive metadata, making it challenging for researchers to locate and understand them. Here, we illustrated the use of RO-Crate's developments [97], by packaging both the reproducible workflow (available in the Github repository) and local results (available in Zenodo) together with their data into a standardized format [10,98], alongside referencing the workflow in WorkflowHub [99,100]. By publishing these objects in open-source repositories like Github and Zenodo, we additionally made them accessible for other researchers. We addressed different aspects of interoperability (i.e., legal, organizational, semantic, technical) [101]. Legal compliance was obtained by using a federated, or 'data visiting', approach (i.e., keeping data under the control of the data holders) and following data minimization principles (i.e., specifying data requirements to answer a research question and just mobilizing minimal required data). Organizational interoperability was pursued by extensively collaborating with different partners, bridging gaps between institutions and researchers from diverse fields throughout the process and clearly specifying responsibilities. By building a CDM, we addressed semantic interoperability aspects, providing standardized semantic and syntactic specifications for the required data, and crosswalks or translations to different encoding systems. We also addressed technical interoperability challenges by providing a reproducible workflow and deployment environment. Reusability was promoted by providing proper documentation and descriptions [102,103], by adopting the standardized RO-Crate metadata format [97] and providing data provenance metadata using the common provenance model [85]. Publication of the workflow in Github [10] enables other researchers to rapidly adopt and adapt the analytical pipeline scripts.

By piloting the described framework, in addition to challenges that arise within jurisdictional borders (e.g., local governance issues, missing data), we identified key limitations that should be addressed to allow adoption of the framework for timely cross-border policy-relevant research. These insights were complemented by results from qualitative analyses in [Chapter 5](#), which also revealed common challenges and requirements to reach successful integration of socio-economic and pathogen genomic data with other health-related data to address public health research questions in Europe.

Challenges encountered throughout the work

Labor-intensive manual processes

Firstly, the proposed and demonstrated methodological approach still involved extensive manual processing and interactions. Building a Common Data Model (CDM), which involves specifying data requirements, was executed by manually generating an Excel file that captures required variables along with their labeling, description, format. It potentially requires the specification of crosswalks between classification systems. Generating an accurate and thorough CDM description is time-consuming, and despite the provision of a customizable template, is lacking a standardized approach. Additionally, synthetic data were generated manually based on the CDM specification, avoiding data-driven generation methods (i.e., based on existing real-world data) to avert concerns of re-identification. This manual generation process, which involves capturing the structure, syntactic, and semantic requirements of the CDM and ideally reflecting real population-level distributions, is also demanding and requires technical skills to program the data generation.

Further, data access applications within the sites, the integration and mapping of local data to the CDM specifications, and extensive collaboration, requiring repeated back-and-forth between the organizing partner (i.e., Coordination Node) and nodes participating and having access to the required individual-level data (i.e., Participant Nodes), result in an overall process that spans multiple months to finalize, often making it unsuitable for generating timely evidence. For example, in the demonstrated workflow, the study protocol was published in January 2023. IACS led the way in locally deploying the reproducible analytical pipeline, delivering preliminary results within three months after publication of the study protocol, by April 2023 [104]. THL accessed the required individual-level data in October 2023. After addressing issues through an iterative back-and-forth debugging process, the analytical pipeline was successfully executed by February 2024, 13 months after the protocol's release. Similarly, a dedicated researcher at Sciensano obtained access to the required data for the specified analyses in December 2023 and exported initial results by February 2024, also taking 13 months to reach this stage.

Complying with data requirements

The collaboration among different partners across Europe was facilitated by the European BY-COVID project (project duration: October 2021 – September 2024) [105], involving different research institutes having access to individual-level health and health-related data. Initially, six partners were identified: the Belgian Institute of Public Health (Sciensano), Instituto Aragonés de Ciencias de la Salud (IACS), the Finnish Institute for Health and Welfare (THL), the Dutch National Institute for Public Health and the Environment (RIVM), the University of Tartu (Estonia), and the Austrian National Public Health Institute (GÖG) [106]. The causal model, CDM, and study protocol were drafted, engaging these different partners. However, eventually only Sciensano, THL, and IACS were able to acquire access to

the required individual-level data of the respective sites Brussels and Wallonia (Belgium), Finland, and Aragon (Spain) for this specific research in question, and were therefore considered as ‘Participant Nodes’ in this study [106].

To meet the data requirements to address the specified research question on vaccine effectiveness in the Belgian population, multiple data sources could be linked by an authorized researcher within a secured processing environment established under the LINK-VACC project (see above for more information) [27]. Following the initial crisis mode of the pandemic, reasonable cautiousness increased, ensuring that strict reliance to the authorized use is maintained. To verify that research questions align with the post-authorization surveillance of vaccines objectives of the LINK-VACC project and to ensure continuous involvement of data owners, a Governance Board of data owners was established [107]. However, involving diverse data owners was at times challenging, particularly resulting in a lack of timely authorization to process vaccination data from the Flanders region. Consequently, the Belgian population cohort was limited to a recorded population of the regions Brussels and Wallonia.

Perfect adherence to the CDM was not a prerequisite to participate in the study. The Participant Nodes eventually involved in the study each did not fully comply with the specified data requirements. For example, information on the covariates *foreign* and *pregnancy* was lacking for the population cohort of Brussels and Wallonia (Belgium). However, such non-compliance with required variables restricts the ability to adequately adjust for confounding, potentially impairing causal interpretation.

While the ability to achieve the linkage of diverse data sources for the study on vaccine effectiveness in Aragon (Spain), Brussels and Wallonia (Belgium), and Finland was a significant accomplishment, complying with detailed data requirements for causal inference remains a significant challenge, as illustrated in this study. Limitations related to aspects like data availability, accessibility, and governance can hinder participation or full adherence to predefined data models and study protocols. Despite efforts to harmonize data through the use of the CDM, disparities in availability and completeness of datasets underscore the difficulties in achieving uniformity. These challenges reflect broader issues of heterogeneity in data collection practices, legal frameworks governing data access, and the technical infrastructure of Participant Nodes.

Centralized trust

The demonstrated framework relied on a central Coordination Node for providing required research objects (such as a research protocol, CDM, and reproducible analytical pipeline) and offering guidance, coordinating the process and communication with the ‘Participant Nodes’, and ensuring

harmonization across the federation. This approach aligns with a '*centralized trust*' architecture in federated computing [92]. Within this structure, the different nodes within the federation (referred to here as '*Participant Nodes*') are considered data controllers, independently managing and processing the personal data available to it. Each node ensures compliance with legal and regulatory requirements for its data-processing activities, operating independently of other nodes [92]. The central '*Coordination Node*' serves as the authority facilitating orchestration, which can simplify coordination, enforce standards, and enhance operational efficiency.

This approach was well-suited to the context of the work described in [Chapter 3 and 4](#), as it addressed a complex research question, aimed to estimate a causal effect, and involved only three Participant nodes. However, the centralized trust model has notable limitations. As the federation scales up (i.e., more Participant Nodes join the study), the central Coordination Node may encounter performance bottlenecks [92]. Moreover, the reliance on a single central authority requires all participants to place their trust in it. In contexts where trust is limited, this dependency may preclude some organizations from participating in the federation [92].

Technical barriers

Technical elements can also present significant barriers, especially when dealing with large-scale datasets (or '*big data*'). In our study, where the goal was to obtain population-wide estimates at each participating site, we were confronted with large population cohorts, resulting in the need to manage and analyze vast volumes of data, with potentially complex data structures. While effective statistical methods might be available to address or minimize biases, applying these methods to large datasets can be particularly difficult and computationally demanding. For example, when attempting to handle missing values through imputation in our demonstrated workflow, we encountered challenges with runtime efficiency. Applying a multiple imputation approach across such large datasets would have dramatically inflated the computational load and processing time, which made us rely on the suboptimal approach of single imputation [108]. Similarly, ML techniques designed for confounder adjustment, such as the earlier mentioned TMLE, offer sophisticated solutions, but can become computationally intensive when working with large data volumes. To overcome some of the efficiency hurdles, we adopted the data management tool '*DuckDB*', to allow for faster querying, and reading and writing of large datasets when working with workflow with different modules, making it easier to handle computational demands [109]. However, it remains important to explore alternative solutions to improve computational efficiency further when using complex statistical approaches.

Additionally, complex analytical processes designed to account for potential biases often involve multiple human decision points. This raises a critical question: can such nuanced analytical procedures be fully automated? Achieving full automation becomes especially challenging when considering the situation where we want to distribute reproducible analysis workflows in a federated infrastructure, where data remains decentralized. Automation needs robust algorithms capable of making informed decisions at each analysis step.

Alternatively, ensuring compatibility across different platforms, software versions, and computational environments is another important barrier towards deploying reproducible workflows. Our study adopted containerization strategies (i.e., Docker, Singularity) to generate the same conditions under which the analysis is performed in different sites. However, ensuring proper management of dependencies, such as software packages or libraries, can be cumbersome, especially when updates or deprecations occur. In addition, advanced programming skills can be required to create modular, well-documented workflows that are not only reproducible but also scalable, and to package a workflow together with its dependencies in a container. Further, these solutions are not always well-known, trusted, or supported by local IT departments or infrastructure maintainers, which can impede their effective adoption and integration into existing systems. For example, within the secured processing environment of the LINK-VACC project, where data from the Belgian population cohort was linked and analyzed, container deployment was not supported.

Heterogeneity in data availability, data quality, and legislations across sites

As touched upon above, heterogeneity across sites presents a significant barrier towards conducting federated analyses. Differences between alternative systems, institutes, or sites can introduce bias, complicating efforts to obtain valid evidence from observational public health studies. Even when following the step-by-step methodological approach (i.e., working with fit-for-purpose bespoke CDMs that uniformize variable definitions, semantics and the structure of data, and implementing a reproducible workflow and deployment environment), differences between sites may still exist through elements like inconsistent data collection and integration methods, data storage, population characteristics, and health care practices. Further, variability in data quality and missingness between sites, can limit the accuracy and comparability of local results. For example, when applying the methodological framework as outlined in [Chapter 3](#) and [Chapter 4](#), information on *pregnancy* status was unavailable for Brussels and Wallonia (Belgium) and Finland, whereas it was available for Aragon (Spain). These unavailabilities can introduce bias and limit comparability across these sites. When different centers in a federated analysis measure different confounders, the challenge lies in ensuring that the analysis accounts for these discrepancies, while maintaining consistency and accuracy across

the study sites. When these discrepancies persist, researchers should report and interpret these limitations transparently and consider their potential implications for the results (e.g., potential for residual confounding when key variables are missing), which can entail limiting causal interpretation of the results. It is important to state how these limitations might influence the accuracy, generalizability, and comparability of the findings.

When pooling or comparing estimates from different sites, it is important to carefully consider the specifics of the local analyses. Non-collapsibility refers to the property of certain measures (e.g., odd ratios) whereby the measure changes when adjusting for covariates through regression, even if those covariates are not true confounders [110,111]. This phenomenon can affect the comparability of results across sites, especially if different sites account for different sets of confounders. In scenarios where non-collapsibility could introduce inconsistencies or bias, the use of collapsible measures (e.g., risk or rate differences) may be preferred [111,112]. These measures are not influenced by the inclusion or exclusion of non-confounding covariates, making them more robust for aggregating and interpreting results across sites [112]. In the workflow demonstrated in [Chapter 3](#) and [Chapter 4](#), we computed the collapsible Restricted Mean Survival Time Difference (RMSTD) as an estimation of the average treatment effect [113,114].

Additionally, heterogeneity in the local interpretation of European data privacy regulations, national legislations, and data access application processes, can hinder the seamless integration and analysis of data from multiple regions. While the GDPR provides an overarching legal framework on privacy and data protection across European countries (EU/EEA), its interpretation and implementation can vary between member states [115]. Each country has the flexibility to apply the regulation differently, reflecting their national legislations and local policies, leading to inconsistencies in data governance practices across countries [115]. This variation creates obstacles for cross-border research, highlighting the growing need for harmonization of GDPR implementation across EU/EEA member states to ensure smoother collaboration, and a more consistent approach across the European region. The European Health Data Space (EHDS), proposed by the European Commission in 2022 and adopted by the European Parliament in April 2024, is a promising initiative, with the potential to address these challenges by aiming to provide a standardized framework for health data sharing and access, and aspiring to harmonize data protection practices and processes for data linkage across member states [116,117].

These barriers must be carefully addressed to ensure that cross-jurisdictional research is not only methodologically valid, but also complying with varying regulations, allowing us to obtain reliable (causal) estimates, generalizable across all sites involved.

Statistical approaches

In the demonstrated analytical pipeline, we adopted a day-by-day matching approach (i.e., a longitudinal matching approach, sequentially enrolling individuals [118]). This means that on each day of the enrolment period, the exposed individuals (i.e., those who have completed a primary vaccination schedule on that specific day) were 1:1 matched to unexposed individuals (i.e., those who have not (yet) completed a primary vaccination schedule) based on a set of confounding factors. As such, not all individuals in the study population are enrolled at one time point. By conducting daily matching based on these confounding factors, we ensured that exposed and unexposed individuals were comparable based on the considered characteristics [119]. Daily matching also balanced time-dependent phenomena between exposed and unexposed individuals, like the COVID-19 epidemiological waves. This approach was inspired by the methodology adopted in the study of Dagan et al. (2021) [120], and was also adopted elsewhere to study SARS-CoV-2 vaccine effectiveness [121–123]. A disadvantage of matching includes the loss of information when unmatched individuals are discarded, potentially reducing statistical power [124]. The use of a rigorous matching algorithm in this study, involving exact matching whenever possible (i.e., whenever exact matches could be found) or based on a distance-based algorithm (i.e., greedy nearest neighbor propensity-score matching, with propensity scores estimated using a generalized linear model), enhanced matching efficiency, minimizing the loss of information. When matching is not perfect (i.e., not exact) matching estimators can exhibit bias in finite samples [124,125]. However, in this study, large matched samples were successfully obtained, and covariates achieved a good balance following the matching process. This indicated that the matching approach effectively limited confounding in considered factors, enhancing the reliability of the resulting estimates despite inherent limitations of imperfect matching. Alternatively, G-computation is a more flexible approach that can model complex relationships, include treatment as a time-dependent covariate, and account for time-varying confounding without excluding data [43], but it often requires advanced statistical expertise and strong assumptions about the underlying data-generating process [125]. Ultimately, our choice for daily matching reflects a balance between methodological considerations, simplicity of implementation and interpretation, and the expertise available within our research team.

The adopted iterative daily matching process generated day-specific matched cohorts. During local execution of the analysis, data from these matched cohorts, covering all days of the enrollment period, were combined into one consolidated dataset. This ‘pooled’ dataset was subsequently used to calculate survival probabilities at each time point using the Kaplan-Meier (KM) estimator, resulting in a single KM curve. It is important to note that these time points represent the time elapsed since

enrollment rather than calendar time. The derived KM curve corresponds to the locally matched study population, reflecting individuals pooled across time through daily matching over the enrollment period. Its generalizability to broader populations depends on the representativeness of the matched population. A comparison of baseline characteristics between the matched population and the population eligible for matching could provide valuable insights into the extent of representativeness and potential limitations in generalizing the results more broadly.

Not or partially vaccinated individuals matched to a primary vaccinated individual can change intervention group when they complete a primary vaccination schedule themselves on a future day. To account for individuals being included multiple times, robust variance estimators can be used (i.e., adjust for clustering in the matched population, accounting for the fact that some individuals contribute multiple times to the analysis) [126,127]. Here, this aspect was not accounted for, potentially resulting in an underestimation of the obtained standard errors. This can potentially lead to narrower confidence intervals and as such overly optimistic conclusions about the precision of the estimate [127].

FAIRification

The FAIRification of research objects generated throughout the process was demonstrated, using key repositories or tools like GitHub, Zenodo, RO-Crate and WorkflowHub [97,99]. Continuous development of tools aimed at improving FAIRness of research projects has significantly enhanced the FAIRification process. However, an additional step of checking the achieved level FAIRness after project completion was not performed here.

Way forward

Federated approaches not only help to overcome the barriers associated with centralizing data, but also ensure that robust (causal) insights can still be derived from fragmented yet valuable RWD across borders. Ultimately, strategies to refine and upgrade the proposed methodological framework can further enhance the potential of RWD in public health research, elevating its capacity to inform evidence-based policy in a timely manner and address complex health issues on a large, cross-border scale.

Automation

To limit extensive manual processing and interactions, further automation is required. For example, following the work described in [Chapter 3 and 4](#), a Common Data Model Builder was developed by González-Galindo et al. (2023) [128], comprising a Python library that enables the user-driven,

standardized generation of data models, and associated reproducible folder structure, data validation rules, and metadata. The library comes with a user-interface, lowering the barrier for use, and enabling a broader range of researchers to engage with the tool and methodology, without needing advanced programming expertise. Further validation and broad adoption of these type of tools can facilitate future work.

Machine learning techniques hold potential for automating decision-making in the context of federated (causal) analysis, where it is curial to make informed decisions throughout the analysis while maintaining a reproducible pipeline. These techniques can help automate key processes, such as selecting appropriate models, or handling missing data, thus reducing reliance on manual interventions and human decision points while improving performance and reliability of results [125,129,130]. For example, Targeted Maximum Likelihood Estimation (TMLE) [125] allows researchers to validate and refine theoretical models in an automated way [131].

Alternatively, tools for data-driven synthetic data generation, designed to preserve the statistical distributions and relationships present in the original RWD, can enable the automated generation of realistic datasets for analysis and modeling [132,133]. Generative AI can play an important role in advancing this process. AI solutions, such as Generative Adversarial Networks (GANs), are effective in capturing complex patterns and relationships within RWD, enabling the generation of highly realistic synthetic datasets [132]. However, it is important to carefully consider potential privacy risks associated with data-driven synthetic data generation. While synthetic data can reduce the likelihood of directly re-identifying individuals, there remains a possibility for information leakage, especially if the synthetic data closely mimics sensitive data patterns or rare combinations present in the original dataset [132,133]. When adopting data-driven data generation, advanced privacy preserving techniques like differential privacy can be incorporated [132,133]. However, it always sensible to carefully balance the realism of synthetic data with the privacy concerns [132].

Technical solutions

Several solutions can be employed to overcome technical barriers that occur when analyzing RWD distributed across jurisdictional borders to obtain population-wide (causal) estimates in public health research, particularly when working with population-level datasets and when the adoption of complex statistical methods is required.

One key approach is the use of parallel computing, which allows for the efficient processing of large datasets and reducing computational load and runtime by distributing tasks across multiple processors or nodes [134]. For example, the iterative multiple imputation approach could benefit from parallel

processing by allowing the simultaneous processing of different subsets of the data on alternative processors [135]. By leveraging parallel computing techniques, researchers can apply complex analyses more efficiently, reducing processing limitations, minimizing the computational burden, and enabling faster results. Alternatively, developing and adopting statistical techniques that have increased performance and efficiency can play an important role in overcoming technical barriers when processing large datasets. For example, XGBoost presents a method for the imputation of missing values delivering high computational efficiency even when confronted with large volumes of data [136]. In addition, collaboration with and support by local IT departments or infrastructure maintainers throughout the research process is critical. These partnerships can facilitate the rapid adoption of novel technological advancements and efficiently address system-specific challenges, improving computational capacity and workflow efficiency.

Ensuring sufficient resources, particularly in terms of computational power and memory, is another key factor. Research involving large datasets requires robust infrastructure, capable of handling the data volume and processing demands [137]. Investments in scalable, secure, and user-friendly infrastructures are essential. This includes building or adopting platforms that enable standardized data collection, secure storage, streamlined access, and efficient analysis across distributed sites. Such an infrastructure would allow researchers to leverage data from multiple jurisdictions while maintaining consistency, ensuring privacy, and fostering collaboration.

Ultimately, these combined efforts to overcome technical challenges can enhance the capacity to perform rigorous population-wide (causal) observational studies in public health research.

Standardization

Standardization efforts are essential for the seamless integration of data across systems and for obtaining truly comparable estimates. Standardization is required at multiple stages of the data lifecycle, from data collection to data storage, access, and secondary use.

Given the diversity in *data collection* methods across regions or countries, it is important to establish a common data standard framework [138,139]. Standardization of variables ensures that data from different sources can be meaningfully compared and integrated. Harmonizing data formats, coding systems (e.g., International Classification of Diseases or ICD codes [140–142], SNOMED classification [143,144]) and units of measurement reduces discrepancies and improves consistency of collected data. The role of minimal datasets to guide standardized data collection has formally been described above. Complementary, increased adoption of metadata standard like DCAT-AP [145] or Schema.org, and using approaches like RO-Crate [97], describing data and other research objects in a standardized

way, can increase the ability to locate and interpret them efficiently. Further, standardized *data storage* facilitates reuse for public health research. Pre-defined general CDMs, like the OMOP CMD [146], can provide a guide to map existing data in storage to a common standard [147,148]. However, mapping data entails a heavy investment from data holders and may risk oversimplification [147,148].

Similarly, developing and adopting internationally recognized data governance protocols (e.g., guiding data access procedures) would facilitate more efficient and timely data access and reuse, while ensuring compliance with regulatory frameworks related to privacy and data protection. Continued efforts to achieve a uniform implementation of the GDPR across Europe would contribute to more timely responses by allowing to build clear, standardized processes, ensuring that compliance is built into everyday operations, and reducing compliance uncertainty [149]. The EHDS envisions these aspects, that aims to create a unified framework and streamlining procedures for data sharing and access, and ensuring that data security principles are uniformly applied across countries [117,150].

By broadly implementing these standards, researchers can enhance the quality, reliability, and interoperability of distributed datasets, enabling the generation of robust, generalizable observational evidence to inform public health policy on a broad, cross-border scale. A top-down approach is recommended to implement change and advance large-scale harmonization, e.g., looking at institutions like the European Commission. Legal enforcement and accreditation of efforts can facilitate broad adoption of standardization efforts.

Decentralized trust

A centralized trust approach was adopted in the work described in [Chapter 3 and 4](#). However, as highlighted earlier, besides having considerable advantages, this architecture has important limitations, such as potential performance bottlenecks and the requirement for trust in a centralized authority, which may raise concerns about transparency and fairness. An alternative to the centralized trust model is a ‘decentralized trust’ architecture, characterized by the absence of a central authority [92]. In such an architecture, trust and agreement among federation members are established using methods like peer-to-peer majority voting (i.e., where members collectively agree on protocols or analytical approaches) and zero-knowledge proofs (i.e., which allows members to prove compliance or correctness without revealing sensitive data) [92]. These decentralized, peer-to-peer federations ensure fair and transparent involvement, with equal rights among federation members [151]. A notable example of a decentralized trust architecture is Swarm learning, which uses blockchain technology and distributed machine learning to enable federated nodes to collaboratively build models, while maintaining data privacy and autonomy [92,151]. In this system, trust is distributed

across the nodes within the federation, with consensus mechanisms ensuring that all participants adhere to agreed protocols without relying on a central coordinator [92,151].

While decentralized trust architectures offer several advantages, including enhances fairness, transparency, and resilience to centralized failures, implementing decentralized trust architectures for federated causal inference requires overcoming several challenges. For instance, the consensus mechanisms and cryptographic tools must be computationally efficient to support population-wide data processing and complex causal analyses [92]. Additionally, the absence of a central authority may complicate the resolution of disputes or inconsistencies in the data or analysis process [92].

When considering centralized versus decentralized trust architectures, it is essential to weigh the advantages and disadvantages of each approach, with the optimal approach depending on the specific requirements, scale, and goals of the research project.

Evaluating FAIRness

As mentioned previously, significant efforts were made to make sure that research outputs obtained in [Chapter 3 and 4](#) adhere to the FAIR principles. However, a unified approach to implementing the FAIR data principles was not employed. Tools like F-UJI [152] or FAIR-Checker [153] could have an added value by automating the evaluating of FAIRification efforts and identifying areas for future improvement to comply with the FAIR principles. Incorporating these evaluation tools into future research workflows could represent an important step toward more robust and comprehensive FAIRification. This step would contribute to establish standardized FAIRification practices, aligning with the broader goals of data-driven research, ensuring that research outputs can support reproducibility, collaboration, and efficiency. Ultimately, adopting such tools would further increase the impact of our research.

4. GENERAL INSIGHTS

Centralization versus decentralization

The debate between centralization and decentralization in policy, governance practices, infrastructure, and resources reflects competing priorities of efficiency, privacy, and control. Both processes have advantages and drawbacks, and different aspects or levels of centralization/decentralization can be considered. When evaluating data storage and analysis capacity, decentralization offers advantages such as the reduced data security risks in the event of a data breach [154,155]. However, decentralization also poses challenges, including increased costs, as each node

must maintain infrastructure capable of locally storing and analyzing data [154,155]. Additionally, decentralization tends to require intense collaboration and communication among diverse partners, which can slow down response times and impede rapid action. In contrast, centralization can streamline operations by standardizing processes, simplifying information access, and consolidating resources, which can lead to greater efficiency and cost-effectiveness in achieving public health goals [154,155]. These benefits must be carefully weighed against the potential vulnerabilities associated with centralizing sensitive data.

A hybrid model, achieving data centralization at a national or regional level, in combination with federated analysis capabilities, could align with Europe's legal, ethical, and research priorities. This approach would balance the efficiency of centralized systems with the privacy and resilience of decentralized ones, fostering robust and collaborative research across Europe. Regardless of the approach, harmonization across Europe is essential. Clear guidelines for data collection, data sharing, consent, and governance can reduce friction and enhance collaboration.

Preparedness

The successful integration of heterogeneous data sources requires a robust system capable of addressing key challenges related to infrastructure, data governance, quality, and interoperability. Building and maintaining such systems should be prioritized during non-crisis periods to ensure preparedness, enabling a swift and effective response when emergencies occur. System preparedness entails several elements, such as a sustainable infrastructure, workforce capacity, and capacity building, and to facilitate all this, sustainable funding is needed.

A broad range of initiatives has already been undertaken to enhance pandemic preparedness, both at the national and European level. These efforts acknowledge existing challenges, and active measures are being taken to address them. For example, the HDA that was established in Belgium in 2023, aims to streamline process to access health data in Belgium (see above) [78]. At the European level, the EHDS regulation supports standardization initiatives, fosters cross-border collaboration, and regulatory-compliant data access for research [116,117]. In addition to these initiatives, significant enhancements in research infrastructures are playing a crucial role in strengthening preparedness. For example, a centralized European discovery portal can serve as a discovery hub, where researchers can efficiently locate datasets and contact points. The establishment of such a portal has also been envisioned under the EHDS [156–158]. Specifically, Article 57 outlines the requirement to compile metadata records into interoperable national dataset catalogues to enhance discoverability [156–158] and calls for the European Commission to establish an EU Datasets Catalogue, connecting these

national datasets catalogues, fostering seamless data discovery and access across member states. The European Health Information Portal (HIP) has previously been established under the Information for Action (InfAct) Joint Action and the Population Health Information Research Infrastructure (PHIRI) project [159], and serves as a practical implementation step towards realizing the EHDS goals, providing an infrastructure for dataset discovery across Europe [159]. It can be seen as a precursor to the EHDS vision to create a cohesive infrastructure for dataset discovery and access across Europe [159]. Another noteworthy example includes the Genomic Data Infrastructure (GDI) project, which focusses on establishing a sustainable infrastructure to support the sharing and accessibility of genomic data, along with related phenotypic and clinical information, across Europe [160].

Significant lessons have been learned from the COVID-19 pandemic, which has accelerated the establishment of systems required for rapid and effective responses to large-scale public health emergencies [161–164]. Improvements in digital infrastructure, real-time surveillance, and data-sharing have been particularly notable [165–168]. Nevertheless, sustainable commitment is required to address gaps and ensure long-term preparedness.

This thesis contributes to preparedness by providing insights that can enhance the reuse of RWD for policy-relevant public health research. By establishing and demonstrating a methodological framework to develop federated cross-border (causal) observational studies in a privacy-preserving and interoperable way, by formulating guidelines to limit bias, and by generating FAIR research objects, the outcomes of this thesis can help researchers tackle emerging research question. Further, by identifying existing gaps, this research puts forward potential targets for future action.

Utility of the generated insights for other domains

Outcomes from this thesis stem from infectious diseases research, but can also provide valuable insights for other public health topics such as chronic disease or health system research, as well as for other domains like the social sciences or environmental field [169]. These research fields often face similar challenges related to governance, interoperability, standardization, collaboration, and data privacy, making these findings broadly applicable.

For example, in the social sciences field, studies assessing the effects of social policies on vulnerable populations (e.g., the impact of housing policies on homelessness [170–172]) often require integrating sensitive personal data from diverse sources like administrative records and surveys [172]. Similarly, in the environmental field, research on climate change or disaster preparedness often involves integrating sensitive individual-level data such as energy usage or health outcomes with large environmental datasets like satellite images, sensor data, and health data [169,173,174].

By identifying and addressing challenges in managing and analyzing (sensitive) personal data, this work contributes to establishing insights that are relevant across domains, advancing data-driven research in diverse fields, ultimately fostering policy-relevant insights while upholding ethical and legal standards.

RECOMMENDATIONS

RECOMMENDATIONS TO ACHIEVE THE SUCCESSFUL AND EFFICIENT INTEGRATION OF REAL-WORLD DATA FOR POLICY-RELEVANT PUBLIC HEALTH RESEARCH WITHIN AND ACROSS JURISDICTIONAL BORDERS IN EUROPE

- **Standardization is key. Standards should be developed and widely adopted at the European level, with national regulations aligned to reflect European recommendations.** Standardization is required at different levels, from data collection, to storage and access procedures, analysis, infrastructures and documentation.
- **Robust multidisciplinary collaborative networks should be established to conduct observational public health research, which involves integrating data from diverse fields, and requires complex statistical procedures, federated networks, and robust digital infrastructures.** It is essential to build bridges, bringing together professionals like clinicians, researchers from different fields, IT specialists, and policy-makers. Collaboration with policy-makers throughout the research project is essential to ensure that the resulting evidence is both relevant and actionable.
- **Robust and scalable systems should be put in place during non-crisis periods to ensure preparedness and enable a rapid and coordinated response in the event of a large-scale public health crisis.** This proactive approach allows for the rapid mobilization of researchers when urgent action is needed. System preparedness entails establishing robust digital and physical infrastructures, standardized protocols and working procedures, maintaining workforce capacity, and providing support systems.
- **With the adoption of the European Health Data Space (EHDS), the European Commission must take a central role in advancing the secondary use of health and health-related data for public health research.** The European Commission should drive the development and enforcement of standards and guidelines, promote collaborative initiatives through European projects, and establish robust systems. These coordinated efforts are crucial to enable (causal) cross-border, policy-relevant public health research by integrating existing real-world data, while ensuring compliance with privacy and data protection regulations.

REFERENCES

1. Brownson RC, Chiqui JF, Stamatakis KA. Understanding Evidence-Based Public Health Policy. *Am J Public Health*. 2009;99(9):1576–1583.
2. Victora CG, Habicht J-P, Bryce J. Evidence-Based Public Health: Moving Beyond Randomized Trials. *Am J Public Health*. 2004;94(3):400–405.
3. Lucas RM, McMichael AJ. Association or causation: evaluating links between “environment and disease”. *Bull World Health Organ*. 2005;83(10):792–795.
4. World Health Organization. Rapid response: Knowledge translation mechanisms to translate evidence into public health policy in emergencies [Internet]. Copenhagen; 2021. Available: <https://iris.who.int/bitstream/handle/10665/342575/WHO-EURO-2021-2719-42477-58997-eng.pdf>
5. Thomson H. Improving Utility of Evidence Synthesis for Healthy Public Policy: the Three Rs (Relevance, Rigor, and Readability [and Resources]). *Am J Public Health*. 2013;103(8):e17–23.
6. Bast E, Tang F, Dahn J, Palacio A. Increased risk of hospitalisation and death with the delta variant in the USA. *Lancet Infect Dis*. 2021;21(12):1629–1630.
7. Bager P, Wohlfahrt J, Rasmussen M, Albertsen M, Krause TG. Hospitalisation associated with SARS-CoV-2 delta variant in Denmark. *The Lancet Infectious Diseases*. 2021;21(10):1351.
8. Peters JA, Farhadloo M. The Effects of Nonpharmaceutical Interventions on COVID-19 Cases, Hospitalizations, and Mortality: A Systematic Literature Review and Meta-analysis. *AJPM Focus*. 2023;2(4):100125.
9. Taccone FS, Goethem NV, Pauw RD, Wittebole X, Blot K, Oyen HV, et al. The role of organizational characteristics on the outcome of COVID-19 patients admitted to the ICU in Belgium. *The Lancet Regional Health – Europe*. 2021;2. doi:10.1016/j.lanepe.2020.100019
10. Meurisse M, Estupiñan-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E. BY-COVID_WP5_T5.2_baseline-use-case [Internet]. GitHub; Available: https://by-covid.github.io/BY-COVID_WP5_T5.2_baseline-use-case/
11. Frank KA, Lin Q, Xu R, Maroulis S, Mueller A. Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science. *Social Science Research*. 2023;110:102815.
12. Brown JP, Hunnicutt JN, Ali MS, Bhaskaran K, Cole A, Langan SM, et al. Quantifying possible bias in clinical and epidemiological studies with quantitative bias analysis: common approaches and limitations. *BMJ*. 2024;385:e076365.
13. Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. *JAMA*. 2022;327(11):1083–1084.
14. Oliver K, Lorenc T, Innvær S. New directions in evidence-based policy research: a critical analysis of the literature. *Health Research Policy and Systems*. 2014;12:34.
15. Moon MD. Triangulation: A Method to Increase Validity, Reliability, and Legitimation in Clinical Research. *Journal of Emergency Nursing*. 2019;45(1):103–105.

16. Rutherford GW, McFarland W, Spindler H, White K, Patel SV, Aberle-Grasse J, et al. Public health triangulation: approach and application to synthesizing data to understand national and local HIV epidemics. *BMC Public Health*. 2010;10:447.
17. Warren JR, White BM. A translational science approach to community-based participatory research using methodological triangulation. *J Public Health (Berl)*. 2022;30:447–458.
18. He AJ. Scaling-up through piloting: dual-track provider payment reforms in China’s health system. *Health Policy and Planning*. 2023;38(2):218–227.
19. Elder RW, Lawrence B, Ferguson A, Naimi TS, Brewer RD, Chattopadhyay SK, et al. The Effectiveness of Tax Policy Interventions for Reducing Excessive Alcohol Consumption and Related Harms. *Am J Prev Med*. 2010;38:217–29.
20. Daley JL, Stahre MA, Chaloupka FJ, Naimi TS. The Impact of a 25 Cent-Per-Drink Alcohol Tax Increase: Who Pays the Tab? *Am J Prev Med*. 2012;42:382–9.
21. Kilian C, Rovira P, Neufeld M, Ferreira-Borges C, Rumgay H, Soerjomataram I, et al. Modelling the impact of increased alcohol taxation on alcohol-attributable cancers in the WHO European Region. *Lancet Reg Health - Eur*. 2021;11:100225.
22. Van Goethem N. Towards genomic-informed pathogen surveillance and control in Belgium : integration of genomic information for the surveillance and control of infectious diseases within the national public health institute in Belgium [Internet]. UCL - Université Catholique de Louvain. 2022. Available: <https://dial.uclouvain.be/pr/boreal/object/boreal:258870>
23. Maes I, Kok E, Dewulf G. Recommendations on a Real-World Data Strategy for Belgium: Multi-stakeholder initiative on reuse of routine care health data. 2022 Jun.
24. Schutte N, Saelaert M, Bogaert P, De Ridder K, Van Oyen H, Van der Heyden J, et al. Opportunities for a population-based cohort in Belgium. *Archives of Public Health*. 2022;80:188.
25. Listl S, Jürges H, Watt RG. Causal inference from observational data. *Community Dentistry and Oral Epidemiology*. 2016;44(5):409–415.
26. Degtiar I, Rose S. A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application*. 2023;10:501–524.
27. Linking of registers for COVID-19 vaccine surveillance. In: European Health Information Portal [Internet]. [cited 18 Sep 2024]. Available: <https://www.healthinformationportal.eu/health-information-sources/linking-registers-covid-19-vaccine-surveillance>
28. Sciensano. Linking of registers for COVID-19 vaccine surveillance. In: sciensano.be [Internet]. [cited 25 Nov 2021]. Available: <https://www.sciensano.be/en/projects/linking-registers-covid-19-vaccine-surveillance>
29. Fairchild A, de Vuyst B. The Evolution of the e-ID card in Belgium: Data Privacy and Multi-Application Usage. Valencia, Spain; 2012.
30. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. An Overview of Record Linkage Methods. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. Agency for Healthcare Research and Quality (US); 2014. Available: <https://www.ncbi.nlm.nih.gov/books/NBK253312/>

31. Berete F, Demarest S, Charafeddine R, De Ridder K, Van Oyen H, Van Hoof W, et al. Linking health survey data with health insurance data: methodology, challenges, opportunities and recommendations for public health research. An experience from the HISlink project in Belgium. *Archives of Public Health*. 2023;81:198.
32. March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, et al. Good Practice Data Linkage (GPD): A Translation of the German Version. *Int J Environ Res Public Health*. 2020;17(21):7852.
33. Van Goethem N, Vilain A, Wyndham-Thomas C, Deblonde J, Bossuyt N, Lernout T, et al. Rapid establishment of a national surveillance of COVID-19 hospitalizations in Belgium. *Archives of Public Health*. 2020;78:121.
34. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. *Biometrical Journal*. 2021;63:915–947.
35. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7:147–177.
36. de Goeij MCM, van Diepen M, Jager KJ, Tripepi G, Zoccali C, Dekker FW. Multiple imputation: dealing with missing data. *Nephrology Dialysis Transplantation*. 2013;28(10):2415–2420.
37. Agiwal V, Chaudhuri S. Methods and Implications of Addressing Missing Data in Health-care Research. *Current Medical Issues*. 2024;22(1):60-62.
38. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol*. 2018;47:226–35.
39. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annu Rev Sociol*. 2014;40:31–53.
40. Cuypers L, Dellicour S, Hong SL, Potter BI, Verhasselt B, Vereecke N, et al. Two Years of Genomic Surveillance in Belgium during the SARS-CoV-2 Pandemic to Attain Country-Wide Coverage and Monitor the Introduction and Spread of Emerging Variants. *Viruses*. 2022;14(10):2301.
41. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. *Acta Obstet Gynecol Scand*. 2018;97:407–16.
42. Carry PM, Vanderlinden LA, Dong F, Buckner T, Litkowski E, Vigers T, et al. Inverse Probability Weighting is an Effective Method to Address Selection Bias During the Analysis of High Dimensional Data. *Genet Epidemiol*. 2021;45:593–603.
43. Hernán MA, Robins JM. *Causal Inference: What If*. 1st edition. Boca Raton: Chapman & Hall/CRC; 2020.
44. Lee S, Lee W. Application of Standardization for Causal Inference in Observational Studies: A Step-by-step Tutorial for Analysis Using R Software. *J Prev Med Public Health*. 2022;55(2):116–124.
45. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*. 2016;183(8):758–764.
46. Fu EL. Target Trial Emulation to Improve Causal Inference from Observational Data: What, Why, and How? *J Am Soc Nephrol*. 2023;34:1305.

47. European Centre for Disease Prevention and Control. Guidance for representative and targeted genomic SARS-CoV-2 monitoring – 3 May 2021. In: ECDC: Stockholm [Internet]. 2021 [cited 17 Dec 2024]. Available: <https://www.ecdc.europa.eu/en/publications-data/guidance-representative-and-targeted-genomic-sars-cov-2-monitoring>
48. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. Geneva: World Health Organization; 2022 pp. 1–21. Available: <https://www.who.int/publications/i/item/9789240046979>
49. Twohig KA, Harman K, Zaidi A, Aliabadi S, Nash SG, Sinnathamby M, et al. Representativeness of whole-genome sequencing approaches in England: the importance for understanding inequalities associated with SARS-CoV-2 infection. *Epidemiol Infect.* 2023;151:e169.
50. Ibrahim NK. Epidemiologic surveillance for controlling Covid-19 pandemic: types, challenges and implications. *J Infect Public Health.* 2020;13:1630–8.
51. de Vries L, Koopmans M, Morton A, van Baal P. The economics of improving global infectious disease surveillance. *BMJ Glob Health.* 2021;6:e006597.
52. Eskandarzadeh S, Fahimnia B, Hoberg K. Adherence to standard operating procedures for improving data quality: An empirical analysis in the postal service industry. *Transportation Research Part E: Logistics and Transportation Review.* 2023;176:103178.
53. Francis RCE, Schmidbauer W, Spies CD, Sörensen M, Bubser F, Kerner T. Standard operating procedures as a tool to improve medical documentation in preclinical emergency medicine. *Emergency Medicine Journal.* 2010;27(5):350–354.
54. Rwanda Ministry of Health. Standard Operating Procedures for Management of Routine Health Information at Referral/Provincial and District Hospitals (Public and Privates) [Internet]. 2019. Available: https://www.moh.gov.rw/fileadmin/user_upload/Moh/Publications/Health_Data/Standard_Operating_Procedures_for_Management_of_Routine_Health_Information_Hospital_2019.pdf
55. Chen H, Hailey D, Wang N, Yu P. A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health.* 2014;11(5):5170–5207.
56. McDonald PL, Foley TJ, Verheij R, Braithwaite J, Rubin J, Harwood K, et al. Data to knowledge to improvement: creating the learning health system. *BMJ.* 2024;384: e076175.
57. Haldane V, Jung A-S, Foo CD, Bonk M, Jamieson M, Wu S, et al. Strengthening the basics: public health responses to prevent the next pandemic. *BMJ.* 2021;375:e067510.
58. Bernardi FA, Mello de Oliveira B, Bettiol Yamada D, Artifon M, Schmidt AM, Machado Scheibe V, et al. The Minimum Data Set for Rare Diseases: Systematic Review. *J Med Internet Res.* 2023;25: e44641.
59. Rojek AM, Moran J, Horby PW. Core Minimal Datasets to Advance Clinical Research for Priority Epidemic Diseases. *Clinical Infectious Diseases.* 2020;70(4):696–697.
60. Qazizadeh Z, Shahbaznejad L, Navaeifar MR, Rezai MS. Developing a National Minimum Data Set for Kawasaki Disease Registry in Iran. *Front Pediatr.* 2022;10.

61. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. *J Am Med Inform Assoc.* 2015;22(1):76–85.
62. Riba M, Sala C, Culhane AC, Flobak Å, Patocs A, Boye K, et al. The 1+Million Genomes Minimal Dataset for Cancer. *Nat Genet.* 2024;56:733–6.
63. Rahman MM, Davis DN. Machine Learning-Based Missing Value Imputation Method for Clinical Datasets. In: Yang G-C, Ao S, Gelman L, editors. *IAENG Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2012.* Dordrecht: Springer Netherlands; 2013. p. 245–57.
64. Rizvi STH, Latif MY, Amin MS, Telmoudi AJ, Shah NA. Analysis of Machine Learning Based Imputation of Missing Data. *Cybern Syst.* :1–15.
65. Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. *Data Technol Appl.* 2021;55:558–85.
66. Alabadla M, Sidi F, Ishak I, Ibrahim H, Affendey LS, Che Ani Z, et al. Systematic Review of Using Machine Learning in Imputing Missing Values. *IEEE Access.* 2022;10:44483–502.
67. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res.* 2009;9: 157. doi:10.1186/1472-6963-9-157
68. Jones K, Ford D, Thompson S, Lyons R. A Profile of the SAIL Databank on the UK Secure Research Platform. *Int J Popul Data Sci.* 4: 1134. doi:10.23889/ijpds.v4i2.1134
69. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak.* 2009;9: 3. doi:10.1186/1472-6947-9-3
70. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. *Journal of Biomedical Informatics.* 2014;50: 196–204. doi:10.1016/j.jbi.2014.01.003
71. Lyons J, Akbari A, Torabi F, Davies GI, North L, Griffiths R, et al. Understanding and responding to COVID-19 in Wales: protocol for a privacy-protecting data platform for enhanced epidemiology and evaluation of interventions. *BMJ Open.* 2020;10: e043010. doi:10.1136/bmjopen-2020-043010
72. About Findata. In: Findata [Internet]. [cited 2 Jan 2025]. Available: <https://findata.fi/en/about-findata/>
73. Viippola E, Kuitunen S, Rodosthenous RS, Vabalas A, Hartonen T, Vartiainen P, et al. Data Resource Profile: Nationwide registry data for high-throughput epidemiology and machine learning (FinRegistry). *International Journal of Epidemiology.* 2023;52: e195–e200. doi:10.1093/ije/dyad091
74. Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR, et al. Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol.* 2013;42:968–77.
75. Goldberg M, Carton M, Descatha A, Leclerc A, Roquelaure Y, Santin G, et al. CONSTANCES: a general prospective population-based cohort for occupational and environmental epidemiology: cohort profile. *Occup Environ Med.* 2017;74:66–71.

76. Panteli D, Polin K, Webb E, Allin S, Barnes A, Degelsegger-Márquez A, et al. Health and Care Data: Approaches to data linkage for evidence-informed policy. *Health Syst Transit*. 2023;25(2):1–248.
77. Erismann S, Pesantes MA, Beran D, Leuenberger A, Farnham A, Berger Gonzalez de White M, et al. How to bring research evidence into policy? Synthesizing strategies of five research projects in low- and middle-income countries. *Health Research Policy and Systems*. 2021;19:29.
78. Health Data Agency. Belgian Health Data Agency: For Better Healthcare, Research, & Policy Making [Internet]. [cited 3 Oct 2024]. Available: <https://www.hda.belgium.be/en>
79. Belgische Kamer van Volksvertegenwoordigers. Legislative proposal on the establishment and organisation of the Health data agency. 2023. Available: <https://www.dekamer.be/FLWB/PDF/55/3065/55K3065008.pdf>
80. Saelaert M, Mathieu L, Van Hoof W, Devleeschauwer B. Expanding citizen engagement in the secondary use of health data: an opportunity for national health data access bodies to realise the intentions of the European Health Data Space. *Archives of Public Health*. 2023;81(1):168.
81. Matters MD, Lekachvili A, Savel T, Zheng Z-J. Developing Metadata to Organize Public Health Datasets. *AMIA Annu Symp Proc*. 2005;2005:1047.
82. Badawy R, Hameed F, Bataille L, Little MA, Claes K, Saria S, et al. Metadata Concepts for Advancing the Use of Digital Health Technologies in Clinical Research. *Digit Biomark*. 2019;3(3):116–132.
83. Mayernik MS, DiLauro T, Duerr R, Metsger E, Thessen AE, Choudhury GS. Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation. *Data Science Journal*. 2013;12:158-171.
84. Viglas SD. Data Provenance and Trust. *Data Science Journal*. 2013;12: GRDI58-64.
85. Wittner R, Soiland-Reyes S, Leo S, Meurisse M, Hermjakob H. BY-COVID D4.3 Provenance model for infectious diseases [Internet]. Zenodo; 2024 Apr. Available: <https://zenodo.org/doi/10.5281/zenodo.10927252>
86. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. *Am J Epidemiol*. 2011;173(7):761–767.
87. Emsley R, Lunt M, Pickles A, Dunn G. Implementing Double-robust Estimators of Causal Effects. *The Stata Journal*. 2008;8(3):334–353.
88. Sanchez P, Voisey JP, Xia T, Watson HI, O’Neil AQ, Tsafaris SA. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*. 2022;9(8):220638.
89. Laan MJ van der, Rubin D. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*. 2006;2(1).
90. Rytgaard HCW, van der Laan MJ. Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Anal*. 2024;30(1):4–33.
91. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*. 2017;167(4):268–274.

92. Casaletto J, Bernier A, McDougall R, Cline MS. Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer. *Annu Rev Genomics Hum Genet.* 2023;24:347–368.
93. González-García J, Estupiñán-Romero F, Tellería-Orriols C, González-Galindo J, Palmieri L, Faragalli A, et al. Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct. *Archives of Public Health.* 2021;79:221.
94. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, Jauer M-L, Deserno T, Chronaki C, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. *Methods of Information in Medicine.* 2020;59(S01):e21–32.
95. Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, et al. A Generic Workflow for the Data FAIRification Process. *Data Intelligence.* 2020;2(1-2):56–65.
96. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
97. University of Technology Sydney, The University of Manchester UK and RO-Crate contributors. Research Object Crate (RO-Crate). In: Research Object Crate (RO-Crate) [Internet]. [cited 3 May 2023]. Available: <https://www.researchobject.org/ro-crate/>
98. Meurisse M, Estupiñán-Romero F, Van Goethem N, Perola M, Paajanen T, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Local outputs and comparative analysis [Internet]. 2024 [cited 11 Jun 2024]. Available: doi:10.5281/zenodo.11209362
99. Silva RF da, Pottier L, Coleman T, Deelman E, Casanova H. WorkflowHub: Community Framework for Enabling Scientific Workflow Research and Development. 2020 *IEEE/ACM Workflows in Support of Large-Scale Science (WORKS).* 2020. pp. 49–56.
100. Bernal-Delgado E, Van Goethem N, Estupiñán-Romero F, Meurisse M, González-Galindo J, Martínez-Lizaga N, et al. BY-COVID WP5 T5.2 Baseline Use Case [Internet]. WorkflowHub. 2024. Available: <https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.502.3>
101. Croatian Institute of Public Health (CIPH), Instituto Aragonés de Ciencias de la Salud (IACS). LOST* and found: Report on interoperability landscape in Europe. pp. 1–55. Available: <https://www.inf-act.eu/sites/inf-act.eu/files/2020-11/D10.1.pdf>
102. Meurisse M, Van Goethem N, Estupiñán-Romero F, González-Galindo J, Royo-Sierra S, Martínez-Lizaga N, et al. BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol [Internet]. 2023 [cited 31 Jan 2023]. Available: doi:10.5281/zenodo.7560731
103. Meurisse M, Estupiñán-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E. BY-COVID_WP5_T5.2_baseline-use-case/vaccine_effectiveness_analytical_pipeline/documentation at main · by-covid/BY-COVID_WP5_T5.2_baseline-use-case [Internet]. [cited 17 Jun 2024]. Available: https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case/tree/main/vaccine_effectiveness_analytical_pipeline/documentation
104. Estupiñán-Romero F, González-Galindo J, Meurisse M, Martínez-Lizaga N, Royo-Sierra S, Van Goethem N, et al. BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment preliminary outputs from Aragon, Spain. Zenodo; 2023 May. doi:10.5281/zenodo.7930373

105. BY-COVID. [cited 30 Mar 2023]. Available: <https://by-covid.org/>
106. Van Goethem N, Bernal-Delgado E, Estupiñán-Romero F, Meurisse M. BY-COVID D5.1 Enriched report viral variants and health outcomes. Zenodo; 2024 Sep. Available: <https://zenodo.org/records/13870075>
107. Martínez-Lizaga N, Meurisse M, Estupiñán-Romero F, Goethem NV, Bernal-Delgado E. BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Data Management Plan [Internet]. 2023 [cited 2 May 2023]. Available: doi:10.5281/zenodo.7625783
108. Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*. 2015;314(18):1966–1967.
109. Raasveldt M, Mühleisen H. DuckDB: an Embeddable Analytical Database. *Proceedings of the 2019 International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1981–1984.
110. Schuster NA, Twisk JWR, ter Riet G, Heymans MW, Rijnhart JJM. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC Medical Research Methodology*. 2021;21:136.
111. Whitcomb BW, Naimi AI. Defining, Quantifying, and Interpreting “Noncollapsibility” in Epidemiologic Studies of Measures of “Effect.” *Am J Epidemiol*. 2020;190:697–700.
112. Didelez V, Stensrud MJ. On the logic of collapsibility for causal effect measures. *Biometrical Journal*. 2022;64:235–242.
113. Ni A, Lin Z, Lu B. Stratified Restricted Mean Survival Time Model for Marginal Causal Effect in Observational Survival Data. *Annals of epidemiology*. 2021;64:149-154.
114. Lin Z, Ni A, Lu B. Matched design for marginal causal effect on restricted mean survival time in observational studies. *J Causal Inference*. 2023;11.
115. Doetsch JN, Dias V, Indredavik MS, Reittu J, Devold RK, Teixeira R, et al. Record linkage of population-based cohort data from minors with national register data: a scoping review and comparative legal analysis of four European countries. *Open Res Eur*. 2021;1:58.
116. European Commission. Commission welcomes European Parliament’s adoption of the European Health Data Space and regulation on substances of human origin [Internet]. [cited 15 Aug 2024]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_24_2250
117. Horgan D, Hajduch M, Vrana M, Soderberg J, Hughes N, Omar MI, et al. European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare. *Healthcare*. 2022;10(9):1629.
118. Thomas LE, Yang S, Wojdyla D, Schaubel DE. Matching with time-dependent treatments: A review and look forward. *Stat Med*. 2020;39: 2350–2370.
119. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. 2010;25: 1–21.
120. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *New England Journal of Medicine*. 2021;384: 1412–1423.
121. Dagan N, Barda N, Biron-Shental T, Makov-Assif M, Key C, Kohane IS, et al. Effectiveness of the BNT162b2 mRNA COVID-19 vaccine in pregnancy. *Nat Med*. 2021;27: 1693–1695.

122. Fabiani M, Mateo-Urdiales A, Sacco C, Rota MC, Petrone D, Bressi M, et al. Relative effectiveness of a 2nd booster dose of COVID-19 mRNA vaccine up to four months post administration in individuals aged 80 years or more in Italy: A retrospective matched cohort study. *Vaccine*. 2023;41: 76–84.
123. Methi F, Gran JM, Valberg M, Kinge JM, Telle K, Magnusson K. Third dose mRNA vaccination against SARS-CoV-2 reduces medical complaints seen in primary care: a matched cohort study. *BMC Medicine*. 2023;21: 157.
124. Greifer N, Stuart EA. Matching Methods for Confounder Adjustment: An Addition to the Epidemiologist's Toolbox. *Epidemiol Rev*. 2021;43: 118–129.
125. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Stat Med*. 2018;37(16):2530–2546.
126. Austin PC, Cafri G. Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Stat Med*. 2020;39: 1623–1640.
127. Barlow WE. Robust Variance Estimation for the Case-Cohort Design. *Biometrics*. 1994;50: 1064–1072.
128. González-Galindo J, Estupiñan-Romero F, Royo-Sierra S, Bernal-Delgado E. *cienciadedatosysalud/cdmb: Common Data Model Builder v1.2.1* [Internet]. Zenodo; 2023. Available: doi:10.5281/zenodo.10168928
129. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell*. 2020;2:369–375.
130. Wang H, Tang J, Wu M, Wang X, Zhang T. Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*. 2022;22:13.
131. Petersen AH, Ekstrøm CT, Spirtes P, Osler M. Constructing Causal Life-Course Models: Comparative Study of Data-Driven and Theory-Driven Approaches. *Am J Epidemiol*. 2023;192: 1917–1927.
132. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: State of the art in health care domain. *Computer Science Review*. 2023;48: 100546.
133. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit Med*. 2023;6: 1–8.
134. Dobre C, Xhafa F. Parallel Programming Paradigms and Frameworks in Big Data Era. *Int J Parallel Prog*. 2014;42: 710–738.
135. Buuren S van. *Flexible Imputation of Missing Data, Second Edition*. 2nd ed. New York: Chapman and Hall/CRC; 2018.
136. Deng Y, Lumley T. Multiple Imputation Through XGBoost. *Journal of Computational and Graphical Statistics*. 2024;33(2):352–363.
137. Anagnostopoulos I, Zeadally S, Exposito E. Handling big data: research challenges and future directions. *J Supercomput*. 2016;72:1494–1516.

138. Unim B, Mattei E, Carle F, Tolonen H, Bernal-Delgado E, Achterberg P, et al. Health data collection methods and procedures across EU member states: findings from the InfAct Joint Action on health information. *Arch Public Health*. 2022;80:17.
139. Feldman K, Johnson RA, Chawla NV. The State of Data in Healthcare: Path Towards Standardization. *J Healthc Inform Res*. 2018;2:248–271.
140. Slee VN. The International Classification of Diseases: Ninth Revision (ICD-9). *Ann Intern Med*. 1978;88(3):424–426.
141. Giles TD, Sander GE. The New International Classification of Diseases (ICD-10): The Hypertension Community Needs a Greater Input. *J Clin Hypertens (Greenwich)*. 2012;14(1):1–2.
142. Almeida MSC, Sousa Filho LF de, Rabello PM, Santiago BM. International Classification of Diseases – 11th revision: from design to implementation. *Rev Saúde Pública*. 2020;54:104.
143. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak*. 2008;8(Suppl_1):S2.
144. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*. 2021;28(9):2017–2026.
145. Barthelemy F, Cochez M, Dimitriadis I, Karim N, Loutas N, Magnisalis I, et al. Towards a standard-based open data ecosystem: analysis of DCAT-AP use at national and European level. *Electronic Government, an International Journal*. 2022;18:137–180.
146. Observational Health Data Sciences and Informatics (OHDSI). Standardized Data: The OMOP Common Data Model [Internet]. 2024 [cited 23 Sep 2024]. Available: <https://www.ohdsi.org/data-standardization/>
147. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med Res Methodol*. 2021;21:238.
148. González-García J, González-Galindo J, Estupiñán-Romero F, Thißen M, Lyons RA, Telleria-Orriols C, et al. PHIRI: lessons for an extensive reuse of sensitive data in federated health research. *European Journal of Public Health*. 2024;34(Supplement_1):i43–49.
149. Hölbl M, Kežmah B, Kompara M. Data Protection Heterogeneity in the European Union. *Applied Sciences*. 2021;11(22):10912.
150. European Commission. European Health Data Space [Internet]. 24 Apr 2024 [cited 15 Aug 2024]. Available: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
151. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature*. 2021;594:265–70.
152. Devaraju A, Huber R. F-UJI – An Automated FAIR Data Assessment Tool. 2024.
153. Gaignard A, Rosnet T, De Lamotte F, Lefort V, Devignes M-D. FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. *J Biomed Semant*. 2023;14:7.

154. Abughazalah M, Alsaggaf W, Saifuddin S, Sarhan S. Centralized vs. Decentralized Cloud Computing in Healthcare. *Appl Sci*. 2024;14:7765.
155. Islam S, Apu KU. DECENTRALIZED VS. CENTRALIZED DATABASE SOLUTIONS IN BLOCKCHAIN: ADVANTAGES, CHALLENGES, AND USE CASES. *Glob Mainstream J Innov Eng Emerg Technol*. 2024;3:58–68.
156. Kessissoglou IA, Cosgrove SM, Abboud LA, Bogaert P, Peolsson M, Calleja N. Are EU member states ready for the European Health Data Space? Lessons learnt on the secondary use of health data from the TEHDAS Joint Action. *European Journal of Public Health*. 2024;34: 1102–1108.
157. Bernal-Delgado E, Cascini F, Dinis S, Estupiñan-Romero F, Fonseca Y, González-García J, et al. Report on architecture and infrastructure options to support EHDS services for secondary use of data. TEHDAS; 2023 Mar pp. 1–56. Available: <https://tehdas.eu/app/uploads/2023/03/tehdas-report-on-architecture-and-infrastructure-options-to-support-ehds-services.pdf>
158. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. Strasbourg: European Commission; 2022 May pp. 1–121. Report No.: COM(2022) 197 final. Available: https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en
159. Tolonen H, Saso M, Unim B, Palmieri L, Schutte N, Peyroteo M, et al. European Health Information Portal: a one-stop shop for health information. *Eur J Public Health*. 2024;34 Supplement_1:i29–34.
160. Pascucci E, Causio FA, Pesole G, Chiara M, Morelli M, Tonon G, et al. Progressing towards personalised medicine: the Genomic Data Infrastructure (GDI) project. *Eur J Public Health*. 2024;34 Supplement_3:ckae144.1956.
161. European Centre for Disease Prevention and Control. ECDC presents lessons learned from the COVID-19 pandemic. 2023 [cited 8 Jan 2025]. Available: <https://www.ecdc.europa.eu/en/news-events/ecdc-presents-lessons-learned-covid-19-pandemic>.
162. Narayanasamy S, Curtis LH, Hernandez AF, Woods CW, Moody MA, Sulkowski M, et al. Lessons From COVID-19 for Pandemic Preparedness: Proceedings From a Multistakeholder Think Tank. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2023;77:1635–43.
163. Assefa Y, Gilks CF, Reid S, van de Pas R, Gete DG, Van Damme W. Analysis of the COVID-19 pandemic: lessons towards a more effective response to public health emergencies. *Glob Health*. 2022;18:10.
164. Fauci AS, Folkers GK. Pandemic Preparedness and Response: Lessons From COVID-19. *J Infect Dis*. 2023;228:422–5.
165. Getachew E, Adebeta T, Muzazu SGY, Charlie L, Said B, Tesfahunei HA, et al. Digital health in the era of COVID-19: Reshaping the next generation of healthcare. *Front Public Health*. 2023;11:942703.
166. Think Tank | European Parliament. The rise of digital health technologies during the pandemic. 2021 [Cited 8 Jan 2025]. Available: [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2021\)690548](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)690548).
167. Lau DT, Sosa P, Dasgupta N, He H. Impact of the COVID-19 Pandemic on Public Health Surveillance and Survey Data Collections in the United States. *Am J Public Health*. 2021;111:2118–21.

168. Clark EC, Neumann S, Hopkins S, Kostopoulos A, Hagerman L, Dobbins M. Changes to Public Health Surveillance Methods Due to the COVID-19 Pandemic: Scoping Review. *JMIR Public Health Surveill.* 2024;10:e49185.
169. Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. *J Biol Res-Thessalon.* 2015;22:9.
170. Culhane D. The Potential of Linked Administrative Data for Advancing Homelessness Research and Policy. HPRI. https://hpri.usc.edu/homeless_research/potential-linked-administrative-data-advancing-homelessness-research-policy/. Accessed 8 Jan 2025.
171. Fatai LO, Salau OF, Gaisie LYK, Muchira K. Leveraging data analytics to optimize government interventions for homelessness, substance abuse, and mental health: A case study in evidence-based policy design. *World J Adv Res Rev.* 2023;20:2025–47.
172. Kube AR, Das S, Fowler PJ. Community- and data-driven homelessness prevention and service delivery: optimizing for equity. *J Am Med Inform Assoc.* 2023;30:1032–41.
173. Gao B, Yang J, Chen Z, Sugihara G, Li M, Stein A, et al. Causal inference from cross-sectional earth system data with geographical convergent cross mapping. *Nat Commun.* 2023;14:5875.
174. Colf LA, McAleavy T. Health consequences of disasters: Advancing disaster data science. *PNAS Nexus.* 2024;3:pgae211.

Summary

Real-world data, routinely collected in vast amounts, are revolutionizing public health research and advancing evidence-informed public health policy. In a complex landscape where data are distributed across isolated data sources and jurisdictional borders, researchers face growing expectations to deliver timely, accurate, and actionable evidence on public health concerns. To successfully conduct observational public health research, they need to find their way through legislations, take into consideration ethical and legal concerns related to privacy and data security, tackle interoperability challenges, and effectively address biases.

To facilitate their work, this thesis aims to develop, describe, demonstrate, and evaluate methodological approaches to integrate heterogeneous real-world data (RWD) sources at the Belgian and cross-jurisdictional border level, enabling researchers to respond to emerging (causal) public health research questions in a more timely and accurate manner, thereby pursuing to provide insights that can enhance elements of pandemic preparedness and response strategies for potential future public health crises. The effective integration of distributed heterogeneous data sources for policy-relevant observational public health research in Belgium ([Chapter 1](#), [Chapter 2](#)) and Europe ([Chapter 4](#)) is demonstrated. It generates valuable real-world evidence, but also highlights potential biases and uncovers challenges for future research. Further, we propose best practices to identify and mitigate identified potential biases ([Chapter 1](#)), and propose a methodological framework to set up cross-border (causal) observational studies ([Chapter 4](#)).

In [Chapter 1](#), we describe potential limitations and biases that researchers are confronted with and evaluate different methodological approaches to study the severity of infection with SARS-CoV-2 variants, based on findings from a review. The impact of different methodological choices is illustrated by using real-world data of Belgian hospitalized COVID-19 patients. We observe different ways of defining COVID-19 disease severity (e.g., admission to the hospital or intensive care unit *versus* the occurrence of severe complications or death) and exposure to a variant (e.g., linkage of the sequencing or genotyping result with the patient data through a unique identifier *versus* categorization of patients based on time periods). Different potential selection biases (e.g., overcontrol bias, endogenous selection bias, sample truncation bias) and factors fluctuating over time (e.g., medical expertise and therapeutic strategies, vaccination coverage and natural

immunity, pressure on the healthcare system, affected population groups) according to the successive waves of COVID-19, dominated by different variants, are identified. Using data of Belgian hospitalized COVID-19 patients, we can document (i) the robustness of the analyses when using different variant exposure ascertainment methods, (ii) indications of the presence of selection bias and (iii) how important confounding variables are fluctuating over time. Best practices are formulated to identify and mitigate potential biases arising from the study design, data analysis methods, underlying surveillance strategies and/or the data infrastructure.

In [Chapter 2](#), another policy-relevant research question was addressed, specifically, the effectiveness of alternative SARS-CoV-2 vaccination schedules (i.e., mRNA booster vaccination *versus* primary vaccination only, and heterologous viral vector–mRNA *versus* homologous mRNA–mRNA prime-boost vaccination) against severe outcomes of Omicron infection was evaluated among hospitalized COVID-19 patients in Belgium, using observational data from national COVID-19 registries. The estimated risk for severe COVID-19 and in-hospital mortality was significantly lower with an mRNA booster vaccination as compared to only a primary vaccination schedule, while no significance difference was observed in the estimated risk for severe COVID-19, ICU admission and in-hospital mortality with a heterologous compared to a homologous prime-boost vaccination schedule. These results support evidence that mRNA booster vaccination reduced the risk of severe COVID-19 disease during the Omicron-predominant period. This analysis additionally reveals challenges and limitations, highlighting areas for future improvement.

Going beyond the jurisdictional borders of Belgium, a methodological framework is proposed in [Chapter 3](#), offering researchers a systematic approach to develop federated cross-border (causal) observational studies using distributed personal data in a privacy-preserving and interoperable way. 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 initial conceptual and instrumental phase of the framework is demonstrated and an analytical pipeline implementing federated causal inference is 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.

In [Chapter 4](#), the implementation phase of the methodological framework is subsequentially demonstrated to address a policy-relevant research question by leveraging RWD from three

countries. Namely, we performed a reproducible cross-border comparison of the real-world effectiveness of primary vaccination in preventing SARS-CoV-2 infections across three sites: Aragon (Spain), Brussels and Wallonia (Belgium), and Finland. This federated population-based observational study shows the effectiveness of the SARS-CoV-2 primary vaccination campaign in prolonging the mean time to infection in the Aragon (Spain) and Brussels & Wallonia (Belgium) population cohorts. Only a minor difference over this time frame is observed in Finland's population cohort.

Given the difficulties in integrating data from other fields in the work conducted in [Chapter 4](#), such as pathogen genomic and socio-economic data, a qualitative assessment of experts' experiences and perceptions towards integrating these data sources for public health research is conducted in [Chapter 5](#). Insights were gathered through conducting an online workshop and in-depth interviews, with qualitative analyses revealing common challenges, concerns, and potential ways forward among participants, such as elements related to data governance, data quality, representativeness, and standardization. This convergence of insights highlights the potential for developing unified strategies that address the complexities inherent in combining heterogeneous datasets for public health research, laying the groundwork for future implementation.

Throughout this work, we identified challenges and requirements for successfully integrating heterogeneous RWD sources for policy-relevant observational public health research. We proposed approaches and formulated recommendations to mitigate potential biases and optimize the integration of RWD for observational public health research. These methodologies and insights can thus enhance various observational research designs, making them applicable across a broad spectrum of public health inquiries. Finally, we outline a way to enhance further the quality and efficiency of (causal) observational public health research, ensuring more robust and actionable insights for public health policy.

Résumé

Les données du monde réel, collectées en grande quantité de manière routinière, révolutionnent la recherche en santé publique et font progresser les politiques de santé publique fondées sur des preuves. Dans un paysage complexe où les données sont réparties entre des sources isolées et des frontières juridictionnelles, les chercheurs font face à des attentes croissantes pour fournir des preuves opportunes, précises et exploitables sur les problèmes de santé publique. Pour mener efficacement des recherches observationnelles en santé publique, ils doivent naviguer à travers les législations, prendre en compte les préoccupations éthiques et juridiques liées à la confidentialité et à la sécurité des données, relever les défis d'interopérabilité, et traiter efficacement les biais.

Pour faciliter leur travail, cette thèse vise à développer, décrire, démontrer et évaluer des approches méthodologiques permettant d'intégrer des sources de données hétérogènes du monde réel (RWD) à l'échelle belge et transfrontalière, permettant aux chercheurs de répondre à des questions de recherche en santé publique émergentes (causales) de manière plus rapide et précise. Ce faisant, elle poursuit l'objectif de fournir des informations susceptibles d'améliorer les éléments de préparation et de réponse aux pandémies dans le cadre de crises de santé publique futures. L'intégration effective de sources de données hétérogènes pour des recherches observationnelles en santé publique pertinentes pour les politiques en Belgique ([Chapitre 1](#), [Chapitre 2](#)) et en Europe ([Chapitre 4](#)) est démontrée. Elle génère des preuves précieuses, tout en mettant en évidence des biais potentiels et en identifiant des défis pour les recherches futures. Par ailleurs, nous proposons des pratiques exemplaires pour identifier et atténuer les biais potentiels identifiés ([Chapitre 1](#)) et un cadre méthodologique pour la mise en place d'études observationnelles transfrontalières (causales) ([Chapitre 4](#)).

Dans le [Chapitre 1](#), nous décrivons les limites potentielles et les biais auxquels les chercheurs sont confrontés et évaluons différentes approches méthodologiques pour étudier la gravité des infections aux variants du SARS-CoV-2, sur la base des résultats d'une revue. L'impact de différents choix méthodologiques est illustré en utilisant des données réelles de patients belges hospitalisés pour COVID-19. Nous identifions des façons variées de définir la gravité de la maladie (par exemple, admission à l'hôpital ou en unité de soins intensifs *versus* survenue de complications

graves ou décès) et l'exposition à un variant (par exemple, association du résultat de séquençage avec les données des patients via un identifiant unique versus catégorisation des patients selon les périodes). Différents biais de sélection potentiels (par exemple, 'overcontrol bias', biais de sélection endogène, biais de troncature de l'échantillon) et des facteurs fluctuants dans le temps (par exemple, expertise médicale, stratégies thérapeutiques, couverture vaccinale, immunité naturelle, pression sur le système de santé) au fil des vagues successives de COVID-19 dominées par différents variants sont identifiés. Ces analyses permettent de documenter (i) la robustesse des analyses en fonction des méthodes utilisées pour déterminer l'exposition au variant, (ii) les indications de la présence de biais de sélection et (iii) les fluctuations des variables de confusion importantes dans le temps. Des pratiques exemplaires sont formulées pour identifier et atténuer les biais potentiels découlant du design des études, des méthodes d'analyse des données, des stratégies de surveillance sous-jacentes et/ou de l'infrastructure des données.

Dans le [Chapitre 2](#), une autre question de recherche pertinente pour les politiques a été abordée: l'efficacité des schémas alternatifs de vaccination contre le SARS-CoV-2 (c'est-à-dire, vaccination de rappel avec un vaccin à ARNm *versus* schéma de vaccination primaire uniquement, et schéma hétérologue vecteur viral-ARNm *versus* schéma homologue ARNm-ARNm) contre les formes graves de l'infection par le variant Omicron a été évaluée parmi les patients hospitalisés pour COVID-19, à l'aide de données observationnelles issues des registres nationaux belges. Les résultats montrent que le risque estimé de COVID-19 sévère et de mortalité hospitalière est significativement réduit avec une vaccination de rappel par ARNm, par rapport à un schéma de vaccination primaire uniquement, tandis qu'aucune différence significative n'a été observée entre un schéma hétérologue et un schéma homologue. Ces résultats soutiennent l'idée que la vaccination de rappel par ARNm réduit le risque de forme grave de COVID-19 pendant la période de prédominance du variant Omicron. L'analyse met également en évidence des défis et des limites, suggérant des améliorations futures.

Au-delà des frontières belges, un cadre méthodologique est proposé dans le [Chapitre 3](#), offrant aux chercheurs une approche systématique pour développer des études observationnelles transfrontalières fédérées (causales) utilisant des données personnelles distribuées, de manière respectueuse de la confidentialité et interopérable. Ce cadre fournit des orientations pas à pas, allant de la définition d'une question de recherche à l'établissement d'un modèle causal, en passant par l'identification des besoins en données dans un modèle de données commun, la génération de données synthétiques, et le développement d'un pipeline analytique interopérable

et reproductible pour un déploiement distribué. Une phase conceptuelle et instrumentale initiale est démontrée, et un pipeline analytique mettant en œuvre des inférences causales fédérées est prototypé à l'aide de logiciels open source, en préparation pour évaluer l'efficacité réelle de la vaccination primaire contre le SARS-CoV-2 dans plusieurs pays.

Dans le [Chapitre 4](#), la phase de mise en œuvre du cadre méthodologique est ensuite démontrée pour répondre à une question de recherche pertinente pour les politiques, en exploitant des RWD provenant de trois pays. Une étude observationnelle transfrontalière fédérée reproductible a été menée pour comparer l'efficacité réelle de la vaccination primaire contre les infections à SARS-CoV-2 dans trois sites: Aragon (Espagne), Bruxelles et Wallonie (Belgique), et la Finlande. Cette étude montre que la vaccination primaire a prolongé le délai moyen avant infection en Espagne et en Belgique, avec une différence mineure observée en Finlande.

Étant donné les difficultés rencontrées dans l'intégration d'autres types de données au [Chapitre 4](#) (par exemple, données génomiques ou socio-économiques), une évaluation qualitative des expériences et perceptions des experts sur l'intégration de ces sources de données a été menée dans le [Chapitre 5](#). À travers un atelier en ligne et des entretiens approfondis, des analyses qualitatives ont révélé des défis communs, des préoccupations et des solutions potentielles, notamment en matière de gouvernance des données, de qualité, de représentativité et de standardisation. Ces résultats soulignent le potentiel de stratégies unifiées pour surmonter la complexité de l'intégration de données hétérogènes en recherche en santé publique.

Tout au long de ce travail, nous avons identifié des défis et des exigences pour intégrer avec succès des RWD hétérogènes dans des recherches observationnelles pertinentes pour les politiques de santé publique. Nous avons proposé des approches et formulé des recommandations pour atténuer les biais potentiels et optimiser l'intégration des RWD. Ces méthodologies et enseignements peuvent ainsi améliorer divers designs de recherche, les rendant applicables à un large éventail de problématiques de santé publique. Enfin, nous proposons des moyens d'améliorer davantage la qualité et l'efficacité des recherches observationnelles (causales) en santé publique, garantissant des informations plus robustes et exploitables pour les politiques de santé publique.

Acknowledgements

I would like to express my heartfelt gratitude to all those who have supported me throughout my PhD journey.

First, I am thankful to my promotor, Annie Robert, for her invaluable support, guidance, and encouragement throughout this PhD journey. Thank you Annie, for the advice and insights that have helped me to shape my research, and the opportunities to present my work. I am grateful for the warm welcome I received at UCLouvain.

I would like to address a special thanks to my co-promotor Nina Van Goethem. Nina, I am very thankful for all your help along the way. You were there every step of the way, giving guidance, providing valuable and thoughtful feedback, and being available to discuss ideas or helping to find solutions to encountered problems. I am very grateful for the opportunity to have worked under your guidance on the BY-COVID project. And importantly, I would like to thank you for being a great companion throughout this journey, available for a chat – about work or anything else.

I would also like to extend my appreciation to the members of my steering committee, Prof. Niko Speybroeck, Prof. Miikka Vikkula, Dr. Jérôme Ambroise, Prof. Herman Van Oyen and Dr. Emilie Cauët, for their constructive feedback and expert advice, which significantly enriched this work. Your time and dedication have helped shape this research. Thank you also to the external jury members, Prof. Stijn Vansteelandt and Prof. Kristel Van Steen, for evaluating my work. I truly appreciate the time and consideration you dedicated to this process, and the invaluable feedback during the final stages of this PhD.

Thank you to the colleagues at IACS, particularly Enrique Bernal Delgado, Francisco Estupiñán-Romero, Javier González-Galindo, and Santiago Royo-Sierra, from whose expertise, ideas, and discussions I have learned a lot, and who have made my research stay in Zaragoza an unforgettable experience. The friendship and collaboration shared with you all have been a cherished part of this process.

I would also like to give my thanks to colleagues at Sciensano, with whom I was lucky to work on manuscripts as co-authors. I am grateful to have learned from your expertise and to have been part of a collaborative environment.

On a more personal note, I wish to thank my friends and family for their support. Bedankt moeke, voor je onvoorwaardelijke steun. Bedankt vake, ik weet dat je fier zou zijn. Oma en opa, bedankt voor de aanmoediging en de gezellige koffiemomentjes. Hanne, Celine, Carole, bedankt voor de goede vriendschap, en de vele fijne momenten de voorbije jaren. Anna, Amber, Laure, bedankt voor de vele relaxerende dagjes in de wellness, die zorgden voor de ultieme ontspanning. Charlotte, bedankt voor alle fijne momenten in de Rondestraat, de cocktailavondjes, en de onvoorwaardelijke vriendschap. Helene, bedankt voor de fijne lunchmomentjes en je fantastische relativeringsvermogen. Bram, bedankt om mijn leven beter te maken. Bedankt voor alle mooie reizen en om deel te maken van een fantastische ervaring in Zaragoza. Maar ook bedankt om mijn steunpilaar te zijn wanneer het wat moeilijker ging, voor het geduld, de aanmoediging, de vele theetjes voor het slapen. You are my rock.

Lastly, to all everyone who I have not mentioned by name but who have contributed to my journey in one way or another, I sincerely thank you. Your kindness, company, wisdom, and support have played an essential role in helping me reach this goal.

Thank you all.