Integration of socioeconomic data in observational studies using a prototyped workflow standard to federated population health research

The BY-COVID Baseline Use Case

Author	Partner
Nina Van Goethem	Sciensano, Belgium
Marjan Meurisse	Sciensano, Belgium
Francisco Estupiñán-Romero	Institute for Health Sciences in Aragon (IACS), Spain
Enrique Bernal-Delgado	Institute for Health Sciences in Aragon (IACS), Spain
Simon Saldner	CESSDA/KNAW DANS, The Netherlands
Vasso Kalaitzi	CESSDA/KNAW DANS, The Netherlands
Lorenz Dolanski-Aghamanoukjan	Gesundheit Österreich GmbH (GÖG), Austria
Teemu Paajanen	Finnish Institute for Health and Welfare (THL), Finland

Contributors	Cees Hof, Irena Vipavc Brvar, Eduardo Antonio Bracho Montes de Oca

Keywords	Real-world data, Social Sciences and Humanities, Data Integration, Pandemic Preparedness, Vaccine Effectiveness

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Date	Version	Editor	Change
2023/10/19	0.1	Nina Van Goethem	Table of Contents, Scope
2023/10/26	0.2	Marjan Meurisse, Nina Van Goethem	Value of socioeconomic indicators for population health research, Socioeconomic indicators
2023/11/10	0.3	Nina Van Goethem, Marjan Meurisse	Socioeconomic data sources, Recommendations
2024/06/03	0.4	Nina Van Goethem, Marjan Meurisse	Integrating insights from workshop II, Adding results from the analysis
2024/06/13	0.5	Francisco Estupiñán-Romero, Lorenz Dolanski-Aghamanoukjan, Teemu Paajanen	Mapping exercise national data sources
2024/07/29	1.0	Enrique Bernal Delgado, Simon Saldner, Vasso Kalaitzi	Review

Funded by
the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Abbreviations

Abbreviation	**Description**
API	Application Programming Interface
CD	Cumulative disadvantage
CDM	Common Data Model
DCMI	Dublin Core Metadata Initiative
DDI	Data Documentation Initiative
EHR	Electronic Health Record
ESeC	European Socio-economic Classification
ESS	European Statistical System
GDPR	General Data Protection Regulation
IMD	Index of Multiple Deprivation
ISCED	International Standard Classification of Education
LAU	Local Administrative Units
NUTS	Nomenclature of Territorial Units for Statistics
RMSTD	Difference in Restricted Mean Survival Time
RWD	Real-world data
SES	Socioeconomic Status
SSH	Social Sciences and Humanities
SPE	Secure Processing Environment
VE	Vaccine effectiveness
WP	Work Package

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Table of Contents

Funded by
the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Scope

The **baseline use case**, developed in BY-COVID WP5, is **prototyping a workflow standard for population health research**. The workflow provides a structured process for **causal inference based** on **real-world observational data**[1] to respond to policy-relevant questions. Conducting causal research, which often implies the need for detailed individual-level data to mitigate confounding (1,2) across national borders, brings challenges in terms of **sensitive data access** and **interoperability** (3). This **federated analysis**[2] **workflow** is designed to leverage real-world heterogeneous data sources from different domains in multiple countries in a privacy-preserving and interoperable way. A detailed description of the proposed framework has been described by Meurisse & Estupiñán-Romero *et al.* (4).

A first policy-relevant research question on **real-world vaccine effectiveness** has been defined and is used to demonstrate the implementation of the described framework. More specifically, we aim to investigate the real-world effectiveness of SARS-CoV-2 primary vaccination as compared to partial or no vaccination in preventing SARS-CoV-2 infection in virtually all resident populations spanning different countries. For this, we have designed an observational study to **emulate the target trial**[3] to estimate the causal effect of interest. Details on the methodology can be consulted in the published study protocol (5).

Further developments in the baseline use case will address how to **integrate additional data types**. Here, we will focus on those data sources containing information on **socioeconomic status (SES)** originating from **Social Sciences**. An **initial workshop** was held in April 2023 in the Hague, on the "*Integration of socioeconomic data in observational studies on vaccine effectiveness*" placing its focus on Belgium and the Netherlands, to promote these developments and stimulate community discussion around it. Preceding the analysis execution, the workshop took an exploratory approach to the topic, concentrating primarily on data infrastructures and the discoverability of data sources within the Social Sciences and Humanities (SSH) field. The workshop report has been published (6) and forms the basis of the current work. A **second workshop** took place, online in June 2024, after the local execution of the analytical pipeline, to evaluate the data integration process and **identify barriers and facilitators related to the identification, linkage, and analysis of individual-level socioeconomic data** (7) in Europe. This evaluation is crucial for understanding the challenges faced in integrating such data and for finding solutions to overcome them. The insights from this second workshop have been used to further enrich

[1] Real-world data (RWD) in the medical and healthcare field "are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources".
[2] Federated analysis is a form of data federation in which collaborators "bring the code to the data" to analyse data locally and integrate results without sharing sensitive data (privacy by design).
[3] A hypothetical target trial is a conceptual framework used in causal research to design and emulate a randomised controlled trial using observational data.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

the current report. Our aim is to generalise these solutions in various disciplinary and geographical contexts. This broader perspective ensures that our findings can be applicable across diverse populations and settings.

Value of socioeconomic data for population health research

As highlighted during the first workshop "BY-COVID Spring 2023 Baseline Use Case workshop: Integration of socioeconomic data in observational studies on vaccine effectiveness" (6), socioeconomic indicators may be relevant at individual and area (i.e., aggregated) levels. There is a wide spectrum of socioeconomic factors and relevant data sources to be considered, depending on the research question(s) and/or the purpose of use. Despite its relevance for population health research and policy, and necessity for the definition and uptake of relevant measures, the use of social science research in public health policymaking tends to be marginal compared to biomedical research, which was also observed during the COVID-19 pandemic. The further integration of socioeconomic data in public health policymaking therefore has the potential to contribute additional research, expertise, and evidential pluralism to future public health crises (8).

Socioeconomic data at the individual level

Socioeconomic status (SES) is a major determinant of health, and associations between low SES and poor health outcomes have repeatedly been found (9–11). In addition, a relationship between SES and access to and use of health care (12), uptake of health interventions (13–15), or exposure to specific health determinants (e.g. environmental, behavioural) (16,17) has frequently been reported. As such, SES is often considered as a **potential confounder**[4] of many exposure-outcome associations, particularly in studies assessing the effect of treatments, interventions, or health determinants on health outcomes. To effectively address confounding associated with socioeconomic position, researchers should aim to control for many socioeconomic factors. Simply controlling for a single individual socioeconomic factor, like income or approximating SES by an area-based measure, is unlikely to fully eliminate all residual confounding by socioeconomic position (18). The workshop report (6) also mentions that multiple individual-level indicators are needed to discover the intersectionality of risks and cumulative disadvantages.[5]

[4] A confounder is a variable that influences both the exposure and the outcome in a study, potentially leading to a spurious association between them if not properly controlled.

[5] Intersectionality explains how different systems of oppression interact and create unique effects, recommending simultaneous analysis of inequalities. Cumulative disadvantage refers to the accumulating impact of social, behavioural, and biological processes on health over a lifetime.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Researchers may also have a particular interest in testing whether the effect of an intervention differs according to the socioeconomic status of participants (19). Indeed, interventions can have differential effectiveness within particular population subgroups (20). Research on **potential moderators[6] of intervention efficacy** is crucial for the further development of intervention programs, strategies, and priorities. For example, a study showed that neighbourhoods with varying levels of disadvantage reacted differently to the implementation and relaxation of COVID-19 mitigation policies (21). Interventions that demonstrate a net overall effect, may risk silently exacerbating health inequalities if its programme is effective for high SES populations but makes no difference to low SES populations (19). Therefore, studying SES as a moderator of intervention effectiveness would help policymakers to appropriately tailor their efforts towards reducing inequalities.

Socioeconomic data at the area level

Given that individual-level SES data are sensitive, these data are often not available in many population-based datasets. In addition, participants may hesitate to respond or tend to exaggerate their self-reported SES in studies requesting these data (22). Therefore, area-level measures are often used to **approximate individual-level SES** when individual-level data are unavailable. Area-level SES are obtained by linking an individual to a geographically defined area and measuring, for example, the median household income or percent of residents with at least a certain education level. These measures of area-level SES are often publicly available from nationally representative federal resources (22). The justification behind this reasoning is that generally, high (low) area-level SES tends to be associated with higher (lower) individual-level SES. However, misclassification of individual-level SES using area-level measures can easily occur. Also, the interpretation of trends across area-level SES groups may be vulnerable to the ecological fallacy[7] (23,24). For instance, a low level of agreement has been observed between individual and area-based income measures (25). This suggests that these metrics may capture distinct populations, each with substantive differences in their demographic profiles. In general, efforts to approximate individual-level SES may be more successful using smaller geographical areas (26).

Area-level measures of SES hold significance in their regard, as they can reflect other health-related characteristics, such as conditions in the **social and physical environment** (22). For example, several studies suggested that individual-level SES variables are stronger predictors of several outcomes than are area-level SES variables but that area-level variables retain important predictive power for vascular disease mortality even after controlling for individual-level SES (27). The persistence of predictive power in

[6] A moderator is a variable that influences the strength or direction of the relationship between an independent variable and a dependent variable in a study.
[7] Ecological fallacy is a type of cross-level bias that results from the assumption that inferences can be made about individual patterns based on patterns observed in groups.

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

area-level SES variables, even after adjustment for individual-level SES variables, implies two possible scenarios. First, they may capture residual confounding at the individual level not fully addressed by individual-level SES. Alternatively, these **ecological variables** might possess independent predictive power by representing community-wide factors influencing disease outcomes, such as access to medical care and the stress from widespread poverty. Indeed, deprivation of an area can also have an independent effect on an individual's health beyond that of the individual socio-economic position. Hence, area-level measures are independently meaningful for health outcomes, and their combined use with individual-level measures allows for a comprehensive contextualisation of SES.

Area-level socioeconomic data can be categorised into various administrative levels, providing structured frameworks for analysis. The **Nomenclature of Territorial Units for Statistics (NUTS)** divides the economic EU territory into regions: NUTS 1 regions encompass major socio-economic entities like states or provinces, NUTS 2 regions divide these further for policy application, and NUTS 3 regions offer detailed data at the county or district level. **Local Administrative Units (LAU)** refine this further, delineating municipalities (LAU 1) and smaller units (LAU 2) within NUTS 3 regions. **Statistical sectors** are specifically created to collect and analyse data on population, housing, economic activities, and other social indicators, aligning with census data collection needs. Their boundaries are designed to be homogeneous with respect to key characteristics, making them ideal for detailed statistical and demographic analyses. In contrast, **postal codes** are defined for mail delivery efficiency, leading to irregular and non-standard boundaries. Mapping postal codes to NUTS levels is possible but complex, highlighting the advantages of using purpose-built statistical units.

Socioeconomic indicators

SES can be operationalised in a multitude of ways, with different indicators measuring different social or economic aspects and capturing an alternative dimension of SES. In public health research studies, SES is often approached by using information on income, education, and/or occupation (28,29), each one measuring different components and aspects of the social environment (30). Different aspects can be more or less relevant depending on the study context (e.g. the research question, study population, geographical area, health outcome of interest, data availability). Therefore, researchers should carefully consider the advantages and disadvantages of the different indicators when assessing their useability. Elements to take into consideration are the **validity** of the indicator (i.e. does the indicator reflect the underlying socioeconomic mechanisms), the **level of measurement** of the indicator (e.g. individual-level, household-level, regional-level, national-level), **selection and combination of indicators** (e.g. using an individual indicator or index combining multiple indicators, weighting of different indicators to construct an index), and the **interpretation and potential of translation to policy** of the indicator. In this section of

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

the report (i.e. 'Socioeconomic indicators') we describe individual- and area-level socio-economic indicators, while in the next section of the report (i.e. 'Socioeconomic data sources') we describe data sources which could potentially be leveraged to retrieve these indicators.

Individual-level socioeconomic indicators

Individual-level socioeconomic indicators provide measures of **an individual's social and economic status** relative to other individuals. Factors such as the occupational status and type of occupation (31–33), personal income (21), years or highest level of education (34,35), and housing tenure (36–38) can serve as indicators of an individual's SES. Table S1 presents a non-exhaustive list of individual-level indicators of SES and possible classification systems. Alternative indicators might have different relations with a health outcome of interest, and as such the choice of a fitting indicator is crucial and dependent on the study context.

Indicators at the **household level**, capturing the social and economic well-being of the household as a whole, can in some instances more adequately reflect the resources of individual household members. For example, for households consisting of a married couple and only one partner receiving an income, the household income will provide a more accurate reflection of the individuals' SES.

Individuals may experience several types of social and economic disadvantages, which can have a **cumulative effect**. The cumulative disadvantage (CD) is a measure of accumulated social, economic, and person-related stressors due to unequal access to resources and opportunities, which increases a person's biological risk for disease (39). As such, **combined indices** of SES can also be used, incorporating different social and economic aspects of an individual into one scale. For example, Kuppuswamy's SES scale combines information on the education level, occupation and monthly income of an individuals' family to provide a measure of SES in India (40).

Area-level socioeconomic indicators

Area-level socioeconomic indicators provide measures of an **average social and economic level of the population in an area**, allowing for the classification of individuals based on the area in which they live. Similar to individual-level indicators, these indicators can be approached by focussing on an individual social or economic factor (e.g. income, education, occupation) or by constructing **compound indicators** combining different factors. Table S2 represents a non-exhaustive list of area-level indicators of SES.

Townsend introduced the concept of '**deprivation**' of a geographical area in the UK in 1987 as a lack of reasonable material (e.g. ability to afford appropriate housing, and essential goods and services) or social (e.g. ability to take part in social activities, social rights, integration) resources of individuals living in this area (41). By measuring the socioeconomic deprivation of an area, contextual factors of the area in which an individual lives can be taken into account and deprived areas can be identified. Townsend proposed a compound index of deprivation constructed based on four elements, namely, unemployment, car and home ownership, and household overcrowding (42,43). However, a deprivation index can be created using different compositions and different standardisation, transformation and weighting approaches. A variety of alternative measures of deprivation have been proposed, such as the Carstairs and Jarman (underprivileged area) index of deprivation (44,45) or the more recent Index of Multiple Deprivation (IMD) (46), which have been frequently used in health research. The IMD can be calculated differently across countries.

Socioeconomic data sources

Individual-level linkages (sensitive)

Individual-level indicators can be obtained from different data sources, such as data from surveys or administrative data. **Administrative data**[8] can cover large populations, tend to be updated on a regular basis, can be obtained at a relatively low cost and effort, and might be more reliable and generalizable than self-reported survey data (47,48). However, administrative data might not be collected in a standardised and comparable way and/or might not contain all information required to adequately build the indicators of interest. **Surveys** on the other hand, can collect specific information required for the construction of socioeconomic indicators (which might not be available in administrative records), with the additional possibility of collecting information in a standardised and comparable way (i.e. using international standard classification systems). However, surveys may come at a high cost, introduce accuracy uncertainties, and yield smaller sample sizes subjected to non-response bias. The use of either administrative or survey data can among other reasons be determined by the data availability, the cost and quality of data, and the research question.

Although individual measures of socioeconomic position are included in some health data sets, an integrated analysis often requires the **linkage** of individual-level health data to other data types, such as lifestyle, environmental or social data, based on a **common identifier**. One of the greatest strengths of data linkage is that it can create comprehensive datasets for a fraction of the cost of collecting primary data. One technical challenge is the

[8] Administrative data are data which are routinely collected by organisations or governmental agencies for administrative purposes.

availability of accurate identifiers that can be used to link the same person across multiple data sources. Further, in order to protect patient and public **privacy**, it is crucial to implement strict data security measures, including data de-identification, access control to the database stored within a Secure Processing Environment (SPE), and careful considerations on how to safely export non-sensitive results.

Area-level linkages (open)

As for individual-level indicators, area-level indicators can be constructed using **survey or administrative data**, with modern indices more frequently adopting an approach based on routinely collected administrative data, which can be more frequently updated. During the first workshop (6), it was raised that for area level indicators, challenges arise in deciding how to aggregate data and choose variables (e.g. aggregating individual level data vs. using publicly available aggregated data).

Discovering data sources with publicly available socioeconomic data can be demanding. There is a diversity of national statistical agencies providing publically available aggregated data across Europe, however navigating websites from these agencies and extracting data can be a laborious process, due to for instance the lack or robust search functionalities, language barriers or formats of the data (e.g. difficult to manipulate). Different data catalogues exist in Europe, allowing the discovery of data sources related to socioeconomic indicators in Europe. For example, the European Health Information Portal (49) or COVID-19 Data Portal (50) can serve as a resource for the discovery of national socioeconomic data sources in the context of health-related research and policy. Nevertheless, data from national agencies can adopt varying data collection methods, use alternative definitions, and standards, hindering accurate cross-country comparisons.

Eurostat, the statistical office of the European Union, provides a wealth of data on various aspects of European society, economy, and environment. The data covers a wide range of topics such as population, employment, education, health, and income. It is made available for reuse under the terms of the European Union's open data policy, which promotes transparency and accessibility of information. Eurostat organises data according to the Nomenclature of Territorial Units for Statistics (NUTS), which divides countries into regions for statistical purposes ranging from broad regional groupings (e.g., NUTS 1) to smaller administrative units (e.g., NUTS 2 and NUTS 3). Eurostat aims to provide comparable data across member states, by striving for harmonisation and standardisation of statistical methodologies. The comparability of Eurostat data between member states depends on several factors, such as data collection methods, definitions, classifications, and quality assurance procedures. Through partnerships like the European Statistical System (ESS), they target the implementation of common guidelines and frameworks.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Application to the Baseline Use Case

Socioeconomic data requirements to estimate vaccine effectiveness

The BY-COVID WP5 Baseline Use Case aimed to assess the **real-world effectiveness of SARS-CoV-2 primary vaccination** compared to partial or no vaccination in preventing SARS-CoV-2 infection in populations spanning different countries by applying a **federated causal research methodology** (4,5). Non-randomised studies assessing the effectiveness of COVID-19 vaccines must account for various factors that may generate spurious estimates due to bias (51). The data requirements for the proposed research question were captured in the **Common Data Model (CDM)**[9] (52). More specifically, all nodes that have been specified in the prespecified **causal model** (see Figure 1), e.g., variables measuring the exposure, outcome and the minimal sufficient adjustment set[10], as well as variables required to achieve secondary objectives of the study or to perform supplementary or exploratory analyses, were listed within the model description of the CDM. The analysis subsequently implemented the daily matching of exposed (i.e. vaccinated) to unexposed (i.e. un- or partially vaccinated) individuals on variables corresponding to nodes in the minimal sufficient adjustment set following the causal model, thereby attempting to close non-causal backdoor paths[11] and limit bias (4).

SES has been identified as an important predictor for COVID-19 vaccine uptake (15,53,54). In addition, people who are socially deprived have a disproportionately greater risk of contracting SARS-CoV-2 infection (55–58). Given that both the exposure (vaccination) and outcome (SARS-CoV-2 infection) may be associated with SES, the potential for **confounding** may exist when estimating **vaccine effectiveness** (59).

By daily matching on the **residence area** as a proxy, the proposed analysis assumed to account for differences in SES. However, controlling for SES is only reasonable if the residence area variable can capture a sufficiently small area (e.g. cities or even neighbourhoods). For example, the NUTS 3 level may be too large to meaningfully control for SES as these areas potentially include large variations in SES. Indeed, matching on the

[9] A common data model is a standardised framework capturing data requirements, thereby detailing syntactic and semantic considerations to achieve interoperability and enable sound comparability between different systems or domains within the federation.
[10] In causal research, a "minimal sufficient adjustment set" is the smallest set of variables that need to be controlled to accurately estimate the causal effect of an exposure on an outcome by blocking all confounding paths.
[11] Non-causal backdoor paths do not represent actual causal relationships but rather arise due to correlations or shared causes between variables that create a spurious association between the exposure and the outcome. Identifying and blocking these non-causal backdoor paths is crucial in causal inference to accurately estimate the true causal effect.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

statistical region of residence only ensures that eligible controls are similar to enrolled cases at the residence area level, while differences may remain between the groups at smaller area levels or at the individual level (59). This means that even after matching on the residence area, **confounding by individual SES** may remain. Therefore, individual-level SES was added as a node in the causal model and identified as a potential confounder. Subsequently, the CDM has included **SES at individual-level** as a variable (*socecon_lvl_cd*) in the data requirements.

The integrated dataset with information on SES individually-linked to the other data (vaccination, infection, comorbidities, etc.), should be prepared by each '*Participant Node*' (i.e. institution being able to access the required individual-level data) in the federated structure and will stay within their SPE given its sensitive nature.

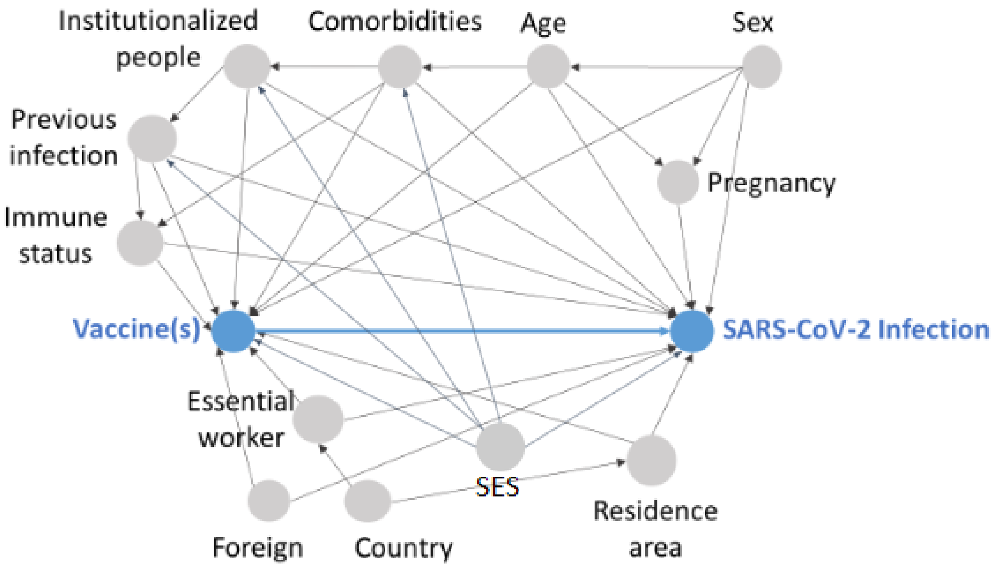

Figure 1. Causal model, established using a Directed Acyclic Graph (DAG), responding to the research question on real-world vaccine effectiveness. SES: socioeconomic status. Source: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4869107.

In addition to relationships at the individual level, there is an impact of political and socio-economic factors on SARS-CoV-2 vaccination trends that are observed in different countries at a population or group level (60). Countries having similar vaccination coverage have shown divergent epidemiological outcomes which may be explained by multiple factors, including differences in the socioeconomic status of the countries being compared (61). Therefore, **area-level SES** may help to explain the observed differences in vaccine effectiveness between countries. The socioeconomic level has been defined in the CDM at area-level as a *recommended* variable (*socecon_lvl_area_nm*). Heterogeneity of vaccine effectiveness (VE) between the different areas and countries is expected. Indeed, VE may vary according to the SES of the recipient. Many studies have found a positive association of VE and one's SES (62–64). Therefore, area-level SES can be used during a comparative analysis in order to attempt **explaining differences in the estimated VE between countries/areas**.

Individual-level socioeconomic indicators in the context of vaccine effectiveness studies

A frequently identified individual-level determinant of vaccine uptake is related to access or ability to understand vaccination information, which is varying by education level (65). However, whilst **education** may be associated with vaccine uptake globally (53,66–68), its role appears to vary by country-context (65). Other frequently cited socio-economic factors associated with vaccine hesitancy are **employment status** and **family economic status** (69,70). Individuals with lower income or unstable employment may face barriers such as lack of access to healthcare services, transportation issues, or time constraints, hindering vaccine uptake. Further, **cultural beliefs** and **trust** in healthcare providers and institutions can influence vaccine acceptance.

Availability and integration of individual-level socio-economic data sources and indicators in the Participant Nodes

Tables S3-6 provide a mapping exercise of individual-level socioeconomic data availability among BY-COVID Baseline Use Case partners: the Belgian Institute for Health (Sciensano), the Finnish Institute for Health and Welfare (THL), Instituto Aragonés de Ciencias de la Salud (IACS), and the Austrian National Public Health Institute (GÖG).

Belgium

The **Belgian LINK-VACC project**[12] links selected variables from existing national registries for COVID-19 vaccine surveillance to ensure the monitoring of COVID-19 vaccines in the phase following their marketing authorization (post-authorization surveillance). This includes the measurement of uptake and coverage of the vaccination, the estimation of vaccine effectiveness, and continuous monitoring of the vaccine's safety. For these purposes, existing pseudonymised data on COVID-19 laboratory test results, hospitalised COVID-19 patients, COVID-19 vaccinations, underlying health problems, socio-demographic and -economic factors, and healthcare worker status are individually linked. The socioeconomic information (civil status, employment status, income decile, ...) originates from **Statistics Belgium (Statbel)**. In general, researchers can request[13] microdata from Statbel if the data is necessary for their statistical or scientific research. In

[12] LINK-VACC project: www.sciensano.be/en/projects/linking-registers-covid-19-vaccine-surveillance
[13] Statbel data request procedures:
https://statbel.fgov.be/en/about-statbel/what-we-do/microdata-research

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

the context of the LINK-VACC project, individual-level data was available[14] for linkage in a SPE on the net taxable income on household level (categorised in deciles), education (using the International Standard Classification of Education - ISCED), employment status, household type (derived from the civil status), and migration background. Finally, three **income** classes, being Low income (decile 1-4), Middle income (decile 5-7), and High income (decile 8-10), were defined and used during the analysis. As such, the LINK-VACC database enables to meet the requirements of the CDM specified for the Baseline Use Case, including the specification of individual-level SES.

Finland

In **Finland**, socioeconomic information originates from **Statistics Finland**. This national institution is responsible for producing and disseminating a broad range of statistical data to support policy decisions and research. It offers comprehensive data on various aspects of Finnish society, including, education, economy and social conditions. Aggregated data (e.g. for the entire country, by municipality, by postal code area) are openly available and can be used freely, as long as the source is cited. Under the Statistics Act, Statistics Finland is authorised to share sensitive data with other statistical authorities, including the Finnish Institute for Health and Welfare (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 licence is required[15]. For their participation in the Baseline Use Case, individual-level data was available for linkage in a SPE on occupation using the European Socio-economic Classification (ESeC). The **occupations** are classified into three groups during the analysis: Higher occupations (ESeC class 1-3), Intermediate occupations (ESeC class 4-6), and Routine and manual occupations (ESeC class 7-9).

Spain

The **INE (National Statistical Institute) in Spain** provides a wide range of statistical data covering various aspects of the country's economy, society, and environment. 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. The INE evaluates each request for access and if approved, only necessary information for the research project will be provided, with access requiring adherence to usage conditions and a commitment to preserving Statistical Secrecy. Currently, individual-level socioeconomic data is not directly linked to patient data

[14] More information on the data landscape for infectious diseases surveillance in Belgium and use of socio-economic indicators is available on https://zenodo.org/records/7988733

[15] Application process Statistics Finland: https://guides.stat.fi/remote-access-to-research-data/application-process-and-agreement-practices

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

in Spain, limiting its availability for research purposes. To link socioeconomic data with patient data, coordination between the Health System and the National or Regional Statistics Institute is required. This involves submitting data requests to both entities and processing requests to the National or Regional Statistics Institute to create a common pseudonymization. In the absence of individual-level socioeconomic data, a proxy variable based on drug copayment levels, which vary depending on income and labour status, is used. However, these categories are considered too broad for meaningful analysis in the Baseline Use Case, as they only differentiate income levels in three broad categories (i.e. less than 18k€, between 18k€ and 100k€ and over 100k€). While individual-level socioeconomic data is not accessible in healthcare, patients in the National Health System are assigned to a reference primary care setting, representing the smallest healthcare administrative area. Socioeconomic information from smaller census areas can be aggregated to characterise the population assigned to each primary care setting.

Austria

In **Austria**, obtaining individual-level socioeconomic 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 to generate 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-practised and trusted enough to be efficient and exploit the full potential of this design. Consequently, these issues limit the ability to conduct detailed studies that require linking socioeconomic information with other types of individual-level data.

Potential residual confounding

A **sensitivity analysis** was conducted for a *Participant Node* having access to individual-level SES, namely Sciensano, to estimate the impact of (not) accounting (as a matching factor) for individual-level SES. In this sensitivity analysis, the unadjusted (acquired by excluding individual-level SES as matching factor) and adjusted (acquired by including individual-level SES as matching factor) vaccine effectiveness (VE) estimates in the Belgian population cohort (including residents of Brussels and Wallonia) were compared. VE was approached by calculating the Difference in Restricted Mean Survival Time (RMSTD) between the contrasted exposure groups. The unadjusted (RMSTD = 57.734 [57.378; 58.090]) and adjusted VE estimate (RMSTD = 59.626 [59.260; 59.991]) revealed a limited percentage change (3.2%) (71). This analysis does not give a strong indication for the presence of residual confounding by individual-level SES after already matching for

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

residence area in the population cohort of Brussels and Wallonia (Belgium). Extrapolation of these results to other settings has to be done cautiously.

Recommendations for future preparedness

How to make socioeconomic data relevant for policy-making?

A disadvantaged socioeconomic position has been established as a potential determinant of infectious diseases (30). Although the complex and interrelated influence of socioeconomic factors on disease transmission, incidence and its outcomes is often unknown and subject to investigation, people with a disadvantaged SES should in general be considered as **high-risk populations** at the time of any infectious disease outbreak (72). However, in order to make socioeconomic data integration relevant for policy-making, it is essential to ensure that the data is actionable, accessible, and aligned with the needs of policymakers.

Individual-level SES are often not collected in healthcare systems as they are not considered as data of clinical interest by most clinicians and as such are absent from medical records. Given that high-quality data on socioeconomic factors are needed to identify groups who are most likely to have poor outcomes, which has implications in the development of public health measures, one could advocate that socioeconomic data should be **routinely recorded** in, for example, medical records (30). However, healthcare providers may lack standardised protocols, resources and training for collecting socioeconomic information. Moreover, **linking** socioeconomic data from external sources allows for the inclusion of a broader range of variables and indicators than may be feasible to collect within medical records alone. On the other hand, several challenges have been identified regarding the linkage of sensitive social science data for health research by the participants of the second public workshop “BY-COVID Spring 2024 Baseline Use Case workshop: integration of individual-level socioeconomic data for infectious diseases research and prevention in Europe”[16]. Reusing existing statistics implies methodological dependencies that may limit the scope and depth of meaningful research, as the original data collection methods and purposes might not align with current research needs. Additionally, the timeliness of socioeconomic data often does not match the timeliness of other data sources, leading to discrepancies that can affect the accuracy and relevance of analyses. Surveys and other traditional data sources may not be updated regularly enough to provide current information, reducing their utility for policy research. Sampling methods used in these sources might also lack relevance or reduce representativeness, limiting their

[16] https://by-covid.org/news-events/spring-2024-baseline-usecase-workshop/

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

applicability to health services and policy research. To address these challenges, there is a need for the use of continuously updated administrative data sources and clear documentation on linking procedures.

Approximation of individual SES by linking people's addresses or postcodes to area-based SES through geolocalisation are not an accurate reflection of individual circumstances and are best used in parallel within individual-level variables to reflect geographical or aggregate-level exposures (30). This view was also shared by the participants during the "BY-COVID Spring 2024 Baseline Use Case workshop", stressing that area-level data only provide a rough estimate of SES and often omit crucial individual-level information. Furthermore, the level of granularity achieved through aggregated data may not be meaningful for the specific research interests, limiting its utility in capturing nuanced socioeconomic differences. Indeed, the selection of the **geographical unit of analysis** is critical and should reflect meaningful units for policy decisions. However, as raised during the workshop, a significant issue arises when healthcare areas do not align with the administrative areas for which socioeconomic data is available. This misalignment can create challenges in data integration and analysis, as health data collected based on healthcare regions may not correspond directly to the administrative regions used for policy and socioeconomic planning. Consequently, the populations captured by both areas might differ, affecting the measures describing such populations. Additionally, for estimating systematic variation between countries, constructing homogeneous units of analysis (i.e., comparable distribution of the population across areas) supports more robust cross-country comparisons (73).

How to make socioeconomic data integration generalisable for future preparedness?

Data governance presents a multitude of challenges that hinder efficient and effective integration of sensitive individual-level socioeconomic data. Even with a federated approach, which offers a potential solution by distributing data management responsibilities, the issue related to data linkage persists within individual nodes. 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 anonymisation or pseudonymisation policies add further layers of complexity. As such, the participants of the "BY-COVID Spring 2024 Baseline use case workshop: integration of individual-level socioeconomic data for infectious diseases research and prevention in Europe"[17] advocated that researchers might benefit from more support structures. Despite mandates for data provision, data holders

[17] https://by-covid.org/news-events/spring-2024-baseline-usecase-workshop/

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

are often ill-prepared for the demands of researchers and may prioritise other tasks, particularly after the urgency of the pandemic wanes. 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. Furthermore, data governance does not always prioritise research use cases, highlighting the need for greater alignment between governance practices and research needs. The workshop participants proposed standardisation 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, while they could also benefit from collaborations across domains at national, european and international levels. Data sharing may be encouraged by initiating national projects where data holders are involved in the process from the start. However, as touched upon in the BY-COVID Fest,[18] consisting of a series of training and knowledge exchange sessions on data sharing and reuse under the General Data Protection Regulation (GDPR) (74), data sharing faces significant challenges due to differences in the interpretation and implementation of the **GDPR** across institutions and countries, leading to inconsistent data sharing practices and compliance uncertainties. This variability complicates the establishment of standardised protocols for data exchange, creating legal and administrative barriers. Additionally, concerns over data privacy, potential misuse, and the risk of re-identification contribute to a general reluctance among stakeholders to share data. Researchers and institutions may fear legal repercussions or damage to their reputation, further hindering collaborative efforts. As a result, valuable data sources often remain siloed, impeding the advancement of research and the potential for comprehensive, multi-faceted insights that could emerge from broader data integration and sharing. The event provided insight from experts across domains and offered training on practices and tools that can be used in order to overcome those challenges and limitations.

Establishing standardised data formats facilitates that socioeconomic data from diverse sources can be integrated for analysis. This includes adopting common data models and metadata standards to enable interoperability among federated entities. **Common data models** provide a structured framework for organising and representing socioeconomic variables consistently across different sources, ensuring that data attributes, relationships, and semantics are uniformly defined and understood. Meanwhile, **metadata standards** define descriptive details regarding the organisation, content, and background of socioeconomic data, encompassing definitions, formats, and conventions. However, challenges such as lack of documentation, semantic standards across data sources (especially cross-border), and the absence of socioeconomic metadata standards in public administration (e.g., statistical offices) persist. Additionally, the level of detail of the data descriptions is often insufficient, making it difficult for researchers, who may not be experts

[18] https://by-covid.org/events/by-covid-fest/

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

on all data sources, to fully comprehend the data. To address these challenges, initiatives like the DANS Data Station Life Sciences[19] aim to establish domain-specific standards by identifying existing standards and cross-walks, and creating a cross-disciplinary knowledge base. Adoption of universal metadata standards, such as Data Documentation Initiative (DDI) or Dublin Core Metadata Initiative (DCMI), fosters improved data discovery, comprehension, and usability across federated systems. An example of interoperability, harmonisation and standardisation is the integration of socioeconomic data sources in the COVID-19 Data Portal[20] in the context of the BY-COVID project.

Data discovery tools enabling efficient data discovery at its source[21], such as the Beacon API (75) used in human biomedical research, are notably absent in the social sciences field. Implementing a similar tool for social science datasets could revolutionise how researchers discover and access socioeconomic data. Such a tool would enable researchers to query extensive databases for relevant socioeconomic information, identify contact points or access locations, and understand the conditions for data usage. This streamlined approach would significantly reduce the time and effort required to locate and verify the relevance of datasets, allowing researchers to evaluate the suitability of data for their studies before embarking on the often lengthy and complex data access application process.

For researchers outside the SSH field, it can be challenging to decide on which socioeconomic indicators and data sources to use. Therefore, 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 integration of data. Further, such an approach supports data minimisation principles, limiting data collection to what is necessary, thereby reducing the potential for misuse of sensitive information. This could be accomplished by defining a set of ‘target’ variables required to answer research questions of interest (76) (i.e. agreeing on a minimal dataset) and that are universally recognised and accepted within the research community. For this, researchers and domain experts need to collaborate to identify SES indicators that capture key dimensions of socioeconomic status and that have been widely used and validated in previous studies, demonstrating their relevance, reliability, and validity in measuring socioeconomic 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. Also, **composite SES indicators** could be created by aggregating multiple individual indicators to simplify comparative analysis across entities, while still providing a comprehensive and holistic measure of socioeconomic status. This approach allows for

[19] https://lifesciences.datastations.nl/
[20] https://www.covid19dataportal.org/search/social-sciences
[21] More information on enabling data discovery at source using beacon-like mechanisms can be found in the BY-COVID WP2 Deliverable 2.3

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

capturing multidimensional aspects of SES, such as income, education, occupation, and wealth, in a single composite index. It was stressed by the workshop's participants that 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.

Conclusions

In population health research, socioeconomic status (SES) of individuals and their surroundings is often considered a **potential confounding factor** in many exposure-outcome relationships under investigation. The Baseline Use Case addresses a policy-relevant question about **vaccine effectiveness** by developing a workflow for **federated analysis** using **heterogeneous real-world observational data sources**. One major challenge involves mobilising individual-level data across federated entities. The required health-related data specified in the **Common Data Model (CDM)** typically includes Electronic Health Records (EHR), laboratory tests, administrative data, healthcare insurance claims, pharmacy records, and specific registries (e.g., vaccination registry). Integrating additional types of data from other domains, such as **Social Sciences and Humanities (SSH)**, could further **enrich** the proposed analyses.

To ensure the generalizability of **socioeconomic data integration** for future preparedness, addressing **data governance** challenges like data linkage issues, diverse ownership, and siloed systems is essential. There is an increased need for support structures to assist researchers in navigating administrative and legal complexities. **Standardising** data formats through common data models and metadata standards would enhance interoperability among federated entities and promote the adoption of a common set of socioeconomic indicators at the European level. Further, implementing **federated querying at the variable level** (e.g., Beacon-like systems) would enhance the reuse of individual-level SES data. **Collaborative** projects like BY-COVID, involving interdisciplinary teams and diverse perspectives, can provide valuable insights to improve data integration and analysis capabilities for future preparedness.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

References

1. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758–64.
2. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. Annu Rev Public Health. 2013;34:61–75.
3. González-García J, Estupiñán-Romero F, Tellería-Orriols C, González-Galindo J, Palmieri L, Fagaralli 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 Dec 9;79(1):221.
4. 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 Oct 23;23(1):248. DOI: https://doi.org/10.1186/s12874-023-02068-3
5. 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 Jan 23 [cited 2023 Feb 9]; Available at DOI: https://zenodo.org/doi/10.5281/zenodo.7551181
6. Kalaitzi, V., Van Goethem, N., Saldner, S. BY-COVID Spring 23 Use Cases Workshop: Integration of socioeconomic data in observational studies on vaccine effectiveness. [cited 2023 Oct 20]; Available at DOI: https://doi.org/10.5281/zenodo.8234104
7. Meurisse M, Saldner S, Van Goethem N, Kalaitzi V, Estupiñán-Romero F, Bernal-Delgado E. BY-COVID Spring 24 Baseline Use Case Workshop. 2024 Jun 19 [cited 2024 Jul 16]; Available from: https://zenodo.org/records/12168495
8. 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 Oct 22;11(4):99.
9. Steptoe A, Zaninotto P. Lower socioeconomic status and the acceleration of aging: An outcome-wide analysis. Proceedings of the National Academy of Sciences. 2020 Jun 30;117(26):14911–7. DOI: https://doi.org/10.1073/pnas.1915741117
10. Chetty R, Stepner M, Abraham S, Lin S, Scuderi B, Turner N, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA. 2016 Apr 26;315(16):1750–66. DOI: https://doi.org/10.1001%2Fjama.2016.4226
11. Lewer D, Jayatunga W, Aldridge RW, Edge C, Marmot M, Story A, et al. Premature mortality attributable to socioeconomic inequality in England between 2003 and 2018: an observational study. Lancet Public Health. 2020 Jan;5(1):e33–41.
12. Alamneh TS, Teshale AB, Yeshaw Y, Alem AZ, Ayalew HG, Liyew AM, et al. Socioeconomic inequality in barriers for accessing health care among married reproductive aged women in sub-Saharan African countries: a decomposition analysis. BMC Women's Health. 2022 Apr 25;22(1):130.
13. Pouliasi II, Hadjikou A, Kouvari K, Heraclides A. Socioeconomic Inequalities in COVID-19 Vaccine Hesitancy and Uptake in Greece and Cyprus during the Pandemic. Vaccines. 2023 Aug;11(8):1301. DOI: https://doi.org/10.3390%2Fvaccines11081301
14. Bozhar H, McKee M, Spadea T, Veerus P, Heinävaara S, Anttila A, et al. Socio-economic inequality of utilization of cancer testing in Europe: A cross-sectional study. Preventive Medicine Reports. 2022 Apr 1;26:101733.

Funded by
the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

15. Cavillot L, Van Loenhout J, Catteau L, Van den Borre L, De Pauw R, Blot K, et al. COVID-19 vaccination uptake in Belgium: socioeconomic and sociodemographic disparities. European Journal of Public Health. 2022 Oct 1;32(Supplement_3):ckac129.046.
16. Hajat A, Hsia C, O'Neill MS. Socioeconomic Disparities and Air Pollution Exposure: A Global Review. Curr Environ Health Rep. 2015 Dec;2(4):440–50.
17. Fairburn J, Schüle SA, Dreger S, Karla Hilz L, Bolte G. Social Inequalities in Exposure to Ambient Air Pollution: A Systematic Review in the WHO European Region. Int J Environ Res Public Health. 2019 Sep;16(17):3127.
18. Blakely T, Hunt D, Woodward A. Confounding by socioeconomic position remains after adjusting for neighbourhood deprivation: an example using smoking and mortality. Journal of Epidemiology & Community Health. 2004 Dec 1;58(12):1030–1.
19. Western MJ, Armstrong MEG, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act. 2021 Nov 9;18:148.
20. Beauchamp A, Backholer K, Magliano D, Peeters A. The effect of obesity prevention interventions according to socioeconomic position: a systematic review. Obes Rev. 2014 Jul;15(7):541–54.
21. Chang HY, Tang W, Hatef E, Kitchen C, Weiner JP, Kharrazi H. Differential impact of mitigation policies and socioeconomic status on COVID-19 prevalence and social distancing in the United States. BMC Public Health. 2021 Jun 14;21(1):1140.
22. Moss JL, Johnson NJ, Yu M, Altekruse SF, Cronin KA. Comparisons of individual- and area-level socioeconomic status as proxies for individual-level measures: evidence from the Mortality Disparities in American Communities study. Population Health Metrics. 2021 Jan 7;19(1):1.
23. Meurisse M, Lajot A, Devleesschauwer B, Van Cauteren D, Van Oyen H, Van den Borre L, et al. The association between area deprivation and COVID-19 incidence: a municipality-level spatio-temporal study in Belgium, 2020–2021. Archives of Public Health. 2022 Apr 2;80(1):109.
24. MacRae K. Socioeconomic deprivation and health and the ecological fallacy. BMJ. 1994 Dec 3;309(6967):1478–9.
25. Buajitti E, Chiodo S, Rosella LC. Agreement between area- and individual-level income measures in a population-based cohort: Implications for population health research. SSM - Population Health. 2020 Apr 1;10:100553.
26. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. Am J Epidemiol. 2002 Sep 1;156(5):471–82.
27. Steenland K, Henley J, Calle E, Thun M. Individual- and Area-Level Socioeconomic Status Variables as Predictors of Mortality in a Cohort of 179,383 Persons. American Journal of Epidemiology. 2004 Jun 1;159(11):1047–56.
28. Galobardes B, Shaw M, Lawlor DA, Lynch JW. Indicators of socioeconomic position (part 1). J Epidemiol Community Health. 2006 Jan;60(1):7–12.
29. Darin-Mattsson A, Fors S, Kåreholt I. Different indicators of socioeconomic status and their relative importance as determinants of health in old age. International Journal

for Equity in Health. 2017 Sep 26;16(1):173.
30. Khalatbari-Soltani S, Cumming RC, Delpierre C, Kelly-Irving M. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. J Epidemiol Community Health. 2020 Aug 1;74(8):620–3.
31. Fujishiro K, Xu J, Gong F. What does "occupation" represent as an indicator of socioeconomic status?: Exploring occupational prestige and health. Social Science & Medicine. 2010 Dec 1;71(12):2100–7.
32. CeLSIUS data dictionary [Internet]. [cited 2023 Oct 31]. Available from: https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone////varDetail.php?tabid=ME91&varid=ECONPO9
33. Gregorio DI, Walsh SJ, Paturzo D. The effects of occupation-based social position on mortality in a large American cohort. Am J Public Health. 1997 Sep;87(9):1472–5.
34. Mirowsky J, Ross CE. Education, Personal Control, Lifestyle and Health: A Human Capital Hypothesis. Res Aging. 1998 Jul 1;20(4):415–49.
35. Laaksonen M, Rahkonen O, Karvonen S, Lahelma E. Socioeconomic status and smoking: Analysing inequalities with multiple indicators. European Journal of Public Health. 2005 Jun 1;15(3):262–9.
36. Dalstra JAA, Kunst AE, Mackenbach JP. A comparative appraisal of the relationship of education, income and housing tenure with less than good health among the elderly in Europe. Social Science & Medicine. 2006 Apr 1;62(8):2046–60.
37. Pollack CE, Knesebeck O von dem, Siegrist J. Housing and health in Germany. Journal of Epidemiology & Community Health. 2004 Mar 1;58(3):216–22.
38. Macintyre S, Ellaway A, Der G, Ford G, Hunt K. Do housing tenure and car access predict health because they are simply markers of income or self esteem? A Scottish study. Journal of Epidemiology & Community Health. 1998 Oct 1;52(10):657–64.
39. Latham-Mintus K, Weathers TD, Bigatti SM, Irby-Shasanmi A, Herbert BS, Tanaka H, et al. Racial Differences in Cumulative Disadvantage Among Women and Its Relation to Health: Development and Preliminary Validation of the Cumulative Stress Inventory of Women's Experiences. Health Equity. 2022 Jun 15;6(1):427–34.
40. Wani RT. Socioeconomic status scales-modified Kuppuswamy and Udai Pareekh's scale updated for 2019. J Family Med Prim Care. 2019 Jun;8(6):1846–9.
41. Jarman B, Townsend P, Carstairs V. Deprivation indices. BMJ. 1991 Aug 31;303(6801):523.
42. Adams J, Ryan V, White M. How accurate are Townsend Deprivation Scores as predictors of self-reported health? A comparison with individual level data. Journal of Public Health. 2005 Mar 1;27(1):101–6.
43. Yousaf S, Bonsall A. UK Townsend Deprivation Scores from 2011 census data [Internet]. UK Data Service; 2017 Jul p. 0–36. Available from: http://statistics.digitalresources.jisc.ac.uk.s3.amazonaws.com/dkan/files/Townsend_Deprivation_Scores/UK%20Townsend%20Deprivation%20Scores%20from%202011%20census%20data.pdf
44. Carstairs V, Morris R. Deprivation: explaining differences in mortality between Scotland and England and Wales. BMJ. 1989 Oct 7;299(6704):886–9.
45. Talbot RJ. Underprivileged Areas And Health Care Planning: Implications Of Use Of Jarman Indicators Of Urban Deprivation. BMJ: British Medical Journal. 1991;302(6773):383–6.
46. Jordan H, Roderick P, Martin D. The Index of Multiple Deprivation 2000 and

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

accessibility effects on health. Journal of Epidemiology & Community Health. 2004 Mar 1;58(3):250–7.

47. Virnig BA, McBean M. Administrative Data for Public Health Surveillance and Planning. Annual Review of Public Health. 2001;22(1):213–30.

48. Burgun A, Bernal-Delgado E, Kuchinke W, Staa T van, Cunningham J, Lettieri E, et al. Health Data for Public Health: Towards New Ways of Combining Data Sources to Support Research Efforts in Europe. Yearb Med Inform. 2017 Aug;26(1):235–40.

49. Homepage | European Health Information Portal [Internet]. [cited 2024 Jun 14]. Available from: https://www.healthinformationportal.eu/

50. COVID-19 Data Portal. COVID-19 Data Portal - accelerating scientific research through data [Internet]. [cited 2021 Nov 16]. Available from: https://www.covid19dataportal.org/

51. Ioannidis JPA. Factors influencing estimated effectiveness of COVID-19 vaccines in non-randomised studies. BMJ Evidence-Based Medicine. 2022 Dec 1;27(6):324–9.

52. 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 Jan 26 [cited 2023 Feb 9]; Available from: https://zenodo.org/record/7572373

53. Lillebråten A, Todd M, Dimka J, Bakkeli NZ, Mamelund SE. Socioeconomic status and disparities in COVID-19 vaccine uptake in Eastern Oslo, Norway. Public Health Pract (Oxf). 2023 May 28;5:100391.

54. Saban M, Myers V, Ben-Shetrit S, Wilf-Miron R. Socioeconomic gradient in COVID-19 vaccination: evidence from Israel. International Journal for Equity in Health. 2021 Nov 8;20(1):242.

55. Vandentorren S, Smaïli S, Chatignoux E, Maurel M, Alleaume C, Neufcourt L, et al. The effect of social deprivation on the dynamic of SARS-CoV-2 infection in France: a population-based analysis. The Lancet Public Health. 2022 Mar 1;7(3):e240–9.

56. Godefroy R, Lewis J. What explains the socioeconomic status-health gradient? Evidence from workplace COVID-19 infections. SSM - Population Health. 2022 Jun 1;18:101124.

57. Patel JA, Nielsen FBH, Badiani AA, Assi S, Unadkat VA, Patel B, et al. Poverty, inequality and COVID-19: the forgotten vulnerable. Public Health. 2020 Jun;183:110–1.

58. Geranios K, Kagabo R, Kim J. Impact of COVID-19 and Socioeconomic Status on Delayed Care and Unemployment. Health Equity. 2022 Feb 2;6(1):91–7.

59. Link-Gelles R, Westreich D, Aiello AE, Shang N, Weber DJ, Holtzman C, et al. Bias with respect to socioeconomic status: A closer look at zip code matching in a pneumococcal vaccine effectiveness study. SSM - Population Health. 2016 Dec 1;2:587–94.

60. Chauhan R, Varma G, Yafi E, Zuhairi MF. The impact of geo-political socio-economic factors on vaccine dissemination trends: a case-study on COVID-19 vaccination strategies. BMC Public Health. 2023 Nov 2;23(1):2142.

61. Alhinai ZA, Elsidig N. Countries with similar COVID-19 vaccination rates yet divergent outcomes: are all vaccines created equal? International Journal of Infectious Diseases. 2021 Sep 1;110:258–60.

62. Gyaase S, Asante KP, Adeniji E, Boahen O, Cairns M, Owusu-Agyei S. Potential effect modification of RTS,S/AS01 malaria vaccine efficacy by household socio-economic status. BMC Public Health. 2021 Jan 28;21(1):240.

63. Lopman BA, Pitzer VE, Sarkar R, Gladstone B, Patel M, Glasser J, et al. Understanding Reduced Rotavirus Vaccine Efficacy in Low Socio-Economic Settings. PLOS ONE. 2012 Aug 6;7(8):e41720.
64. Kelly C, Arnold R, Galloway Y, O'Hallahan J. A Prospective Study of the Effectiveness of the New Zealand Meningococcal B Vaccine. American Journal of Epidemiology. 2007 Oct 1;166(7):817–23.
65. Sacre A, Bambra C, Wildman J, Thomson K, Bennett N, Sowden S, et al. OP126 Are there socioeconomic inequalities in vaccine uptake? An umbrella review. J Epidemiol Community Health. 2023 Aug 1;77(Suppl 1):A123–A123.
66. Bulusu A, Segarra C, Khayat L. Analysis of COVID-19 vaccine uptake among people with underlying chronic conditions in 2022: A cross-sectional study. SSM Popul Health. 2023 May 2;22:101422.
67. Robertson E, Reeve KS, Niedzwiedz CL, Moore J, Blake M, Green M, et al. Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study. Brain, Behavior, and Immunity. 2021 May 1;94:41–50.
68. Vardavas C, Nikitara K, Aslanoglou K, Lagou I, Marou V, Phalkey R, et al. Social determinants of health and vaccine uptake during the first wave of the COVID-19 pandemic: A systematic review. Preventive Medicine Reports. 2023 Oct 1;35:102319.
69. Marzo RR, Sami W, Alam MdZ, Acharya S, Jermsittiparsert K, Songwathana K, et al. Hesitancy in COVID-19 vaccine uptake and its associated factors among the general adult population: a cross-sectional study in six Southeast Asian countries. Tropical Medicine and Health. 2022 Jan 5;50(1):4.
70. Hartonen T, Jermy B, Sõnajalg H, Vartiainen P, Krebs K, Vabalas A, et al. Nationwide health, socio-economic and genetic predictors of COVID-19 vaccination status in Finland. Nat Hum Behav. 2023 Jul;7(7):1069–83.
71. Meurisse, M., Estupiñán-Romero, F., Perola, M., Paajanen, T., Gonzalez-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 [Internet]. SSRN Preprint; 2024. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4869107
72. O'Sullivan TL, Phillips KP. From SARS to pandemic influenza: the framing of high-risk populations. Nat Hazards (Dordr). 2019;98(1):103–17.
73. Thygesen LC, Baixauli-Pérez C, Librero-López J, Martínez-Lizaga N, Ridao-López M, Bernal-Delgado E, et al. Comparing variation across European countries: building geographical areas to provide sounder estimates. European Journal of Public Health. 2015 Feb 1;25(suppl_1):8–14.
74. Saldner S, Bezjak S, Kalaitzi V, Kondyli D, Schulz C, Vipavc Brvar I, et al. Data sharing and reuse under GDPR - BY-COVID Fest workshop report [Internet]. Zenodo; 2024 May [cited 2024 Jun 18]. Report No.: 10.5281/zenodo.11220596. Available from: https://zenodo.org/records/11220597
75. The Global Alliance for, Genomics and Health. Beacon [Internet]. [cited 2024 Jun 18]. Available from: https://www.ga4gh.org/product/beacon-api/
76. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013 Nov 21;10(1):12.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Supplementary material

Table S1. Individual-level indicators of socioeconomic status.

Theme	**Approach**	**Encoding(s)**	**Description**
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, ...) *versus* economically inactive (seeking employment, retired, student, long-term sick, ...)
	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
	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, ...
	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.
		Statistical classification of economic activities in	Classification of economic activities, derived from ISIC (more detailed than ISIC at lower levels, but the same highest level classification), European level.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

		the European Community (NACE)	
		National versions of NACE	Classification of economic activities, national level.
		International Standard Classification of Occupations (ISCO)	Classification of occupations
		Standard Occupational Classification (SOC)	Classification of occupations
		European Socioeconomic Classification (ESeC)	Classification of occupations
Income or wealth	Personal income	-	Indicator of material resources
	Household income	-	Indicator of material resources
	Equivalised income	-	Indicator of material resources adjusts household income for family size and its associated costs of living
	Car ownership	-	Indicator of material resources can be used when a direct measure of income is not available
Education	Years of education	-	Continuous variable

	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), ...
Housing	Housing tenure	Tenure of Household - Standard Classification (TENHOUSE)	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, ...
	Habitat	Social Standing of the Habitat (SSH)	Indicator based the building where the person lives, its surrounding area and the neighbourhood E.g. high, medium-high, medium, medium-low, low

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Table S2. 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 yo, 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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Table S3. Individual-level socioeconomic 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 socioeconomic status?	Yes
Individual-level socioeconomic 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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	- 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 standardised 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 socioeconomic data can involve lengthy and

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	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 socioeconomic data integration	
Linkage capabilities with relevant population health data	Based on the national registry number (NISS)
Pseudonymisation procedures	Only pseudonymised 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 recognised (Royal Decree of 13 June 2014) as a trusted third party and can guarantee the linking of both data and their pseudonymisation. Statbel keeps the key of the pseudonymisation 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 pseudonymisation, and a trusted research environment for data sharing and analysis.
Individual-level socioeconomic data use	
Examples of national studies/projects integrating individual-level socioeconomic 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)

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

Socioeconomic 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 operationalised ● Net taxable income on individual and household levels expressed as deciles

Table S4. Individual-level socioeconomic 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 socioeconomic status?	No. No individual-level socioeconomic 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 socioeconomic data availability	
Data source(s)	Administrative data sources (existing data from public or private institutions) ● Census ● Municipal register ● Mortality data with cause of death

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	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 socioeconomic 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 recognised 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 standardised 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 socioeconomic data with patient data using a common unique identifier, individual-level socioeconomic information is not currently linked to patient data; thus, it is not easily available for reuse in research. In the current framework, linking socioeconomic 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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	reference primary care setting as the smallest healthcare administrative area. Socioeconomic information from the census area level (smaller than the municipality level) can be summarised to characterise the reference population assigned at each primary care setting.
Individual-level socioeconomic data integration	
Linkage capabilities with relevant population health data	Based on the National Identity Document (DNI) and the health insurance card.
Pseudonymisation procedures	Only pseudonymised 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 recognised as a trusted third party and can guarantee the linking of both data and their pseudonymisation. The INE keeps the key of the pseudonymisation so that possibly additional data can be added at a later date.
Data integration barriers/facilitators	Need of a INE (*or Regional Statistics Institute*) as trusted third party for data pseudonymisation; thus requiring coordination between the Health System and INE to share data and produce a common standard pseudonym.
Individual-level socioeconomic data use	
Examples of national studies/projects integrating individual-level socioeconomic data to answer policy-relevant research questions in the field of infectious diseases	Available examples are based on ad-hoc survey consultations on socioeconomic 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. Gaceta sanitaria, 36(1), 70–73.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	https://doi.org/10.1016/j.gaceta.2021.10.003 Or based on socioeconomic 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). International Journal of Environmental Research and Public Health. 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. BMC Public Health 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. Health & place, 70, 102580. https://doi.org/10.1016/j.healthplace.2021.102580
Socioeconomic 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 operationalised ● 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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Table S5. Individual-level socioeconomic 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 socioeconomic status?	Yes
Individual-level socioeconomic 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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	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 licence 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 socioeconomic data integration	
Linkage capabilities with relevant population health data	Personal identity codes and created IDs
Pseudonymisation procedures	Pseudonymised 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 pseudonymised, and the datasets are analysed using IDs instead of personal identification codes.
Data integration barriers/facilitators	Need for a common identifier for data linkage, a trusted party for data pseudonymisation, and a trusted research environment for data sharing and analysis.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Individual-level socioeconomic data use	
Examples of national studies/projects integrating individual-level socioeconomic 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 socioeconomic factors influenced the spread and incidence of the virus, utilising register-based data. ● **National Vaccination Program** Research on the National Vaccination Program uses individual-level socioeconomic data to understand how different socioeconomic groups respond to vaccinations and how vaccination coverage varies across populations.
Socioeconomic 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 operationalised ● 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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Table S6. Individual-level socioeconomic 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 socioeconomic status?	No. No individual-level socioeconomic data is available in healthcare. Only aggregate variables on taxable income (per district) and level of education (per municipality) are available.
Individual-level socioeconomic data availability	
Data source(s)	Administrative data sources: ● National Labor Market Service (unemployment register) ● Coordinated Labor Statistics and Register: combining census, workplace 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)

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

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 organisations.
Individual-level socioeconomic 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.
Pseudonymisation procedures	Pseudonymised 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 socioeconomic data use	
Examples of national studies/projects integrating individual-level socioeconomic 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 socioeconomic disparities in COVID-19 vaccination uptake mediate subsequent disparities in COVID-19 mortality? A test of fundamental cause

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

	theory with nation-wide individual-level register data.
Socioeconomic 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)

Funded by
the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.