BMJ Open

Social inequalities and long-term health impact of COVID-19 in Belgium: protocol of the HELICON population data linkage

Robby De Pauw,[1,2] Laura Van den Borre ,[1,3] Youri Baeyens,[4] Lisa Cavillot,[1,5] Sylvie Gadeyne,[3] Jinane Ghattas,[5] Delphine De Smedt,[6] David Jaminé,[7] Yasmine Khan,[3,6] Patrick Lusyne,[4] Niko Speybroeck,[5] Judith Racape ,[8,9] Andrea Rea,[9] Dieter Van Cauteren,[1] Sophie Vandepitte,[6] Katrien Vanthomme,[3,6] Brecht Devleesschauwer[1,10]

To cite: De Pauw R, Van den Borre L, Baeyens Y, *et al*. Social inequalities and long-term health impact of COVID-19 in Belgium: protocol of the HELICON population data linkage. *BMJ Open* 2023;**13**:e069355. doi:10.1136/bmjopen-2022-069355

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-069355).

RDP and LVdB contributed equally.

Received 19 October 2022
Accepted 23 March 2023

For numbered affiliations see end of article.

Correspondence to
Dr Laura Van den Borre; laura.van.den.borre@vub.be

ABSTRACT

Introduction Data linkage systems have proven to be a powerful tool in support of combating and managing the COVID-19 pandemic. However, the interoperability and the reuse of different data sources may pose a number of technical, administrative and data security challenges.

Methods and analysis This protocol aims to provide a case study for linking highly sensitive individual-level information. We describe the data linkages between health surveillance records and administrative data sources necessary to investigate social health inequalities and the long-term health impact of COVID-19 in Belgium. Data at the national institute for public health, Statistics Belgium and InterMutualistic Agency are used to develop a representative case-cohort study of 1.2 million randomly selected Belgians and 4.5 million Belgians with a confirmed COVID-19 diagnosis (PCR or antigen test), of which 108 211 are COVID-19 hospitalised patients (PCR or antigen test). Yearly updates are scheduled over a period of 4 years. The data set covers inpandemic and postpandemic health information between July 2020 and January 2026, as well as sociodemographic characteristics, socioeconomic indicators, healthcare use and related costs. Two main research questions will be addressed. First, can we identify socioeconomic and sociodemographic risk factors in COVID-19 testing, infection, hospitalisations and mortality? Second, what is the medium-term and long-term health impact of COVID-19 infections and hospitalisations? More specific objectives are (2a) To compare healthcare expenditure during and after a COVID-19 infection or hospitalisation; (2b) To investigate long-term health complications or premature mortality after a COVID-19 infection or hospitalisation; and (2c) To validate the administrative COVID-19 reimbursement nomenclature. The analysis plan includes the calculation of absolute and relative risks using survival analysis methods.

Ethics and dissemination This study involves human participants and was approved by Ghent University hospital ethics committee: reference B.U.N. 1432020000371 and the Belgian Information Security Committee: reference Beraadslaging nr. 22/014 van 11 January 2022, available via https://www.ehealth.fgov.be/ehealthplatform/file/view/AX54CWc4Fbc33iE1rY5a?filename=22-014-n034-HELICON-project.pdf. Dissemination activities include peer-reviewed publications, a webinar series and a project website.

The pseudonymised data are derived from administrative and health sources. Acquiring informed consent would require extra information on the subjects. The research team is prohibited from gaining additional knowledge on the study subjects by the Belgian Information Security Committee's interpretation of the Belgian privacy framework.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Data are drawn from data collections including individual-level data on testing and hospitalisation from the COVID-19 health surveillance (Sciensano), on the socioeconomic and sociodemographic context (Statistics Belgium), and on healthcare use and reimbursement (InterMutualistic Agency).

⇒ The case-cohort study design allows evaluating multiple outcomes, that is, the overall COVID-19 testing strategy (positive and negative tests), COVID-19 infections and COVID-19 hospitalisations for relatively small population groups.

⇒ Limitations of the study design include a potential selection bias resulting from the use of administrative data sources (eg, rejected asylum seekers are missing) and of health surveillance information (eg, non-exhaustive information on COVID-19 tests, infections and hospitalisations).

⇒ The data will be updated during 4 yearly follow-ups for the entire study population.

INTRODUCTION

SARS-CoV-2 and the resulting outbreak of COVID-19 have heightened the need for a swift, efficient and comprehensive health data system at the population level.[1 2] Belgium installed active surveillance systems

at the beginning of the crisis to monitor the number of COVID-19 cases, hospitalisations and deaths in real time.[3–5] These data provide valuable insights into the direct health impact of COVID-19 in Belgium—with estimates reaching nearly 32 000 deaths, 126 000 hospitalised patients and over four million confirmed cases by June 2022.[6] A European comparison of excess mortality—a well-established proxy for the total COVID-19-associated mortality—shows that Belgium experienced substantial excess mortality during the first and second waves of the epidemic in 2020.[7–9]

Despite the wealth of information that is being collected on the current direct impact of COVID-19, two important knowledge gaps remain—that is, (1) What is the social patterning of COVID-19; and (2) What are the long-term direct health impacts of severe COVID-19 infections?

There is emerging evidence that COVID-19 has a socially patterned distribution, with higher risks for exposure, infection, severe illness and death among disadvantaged groups (eg, low-income groups, people with a migrant background) and certain occupations (eg, healthcare workers).[10–14] Currently, the available Belgian research on SE and SD differences builds on COVID-19 incidence on an aggregated level and on all-cause mortality during the first wave at the individual level.[10 13 15 16] There has been no detailed investigation, to our knowledge, on SE and SD differences among hospitalised COVID-19 patients in Belgium. It remains unclear how the health impact of COVID-19 is borne by different population subgroups throughout the pandemic. More detailed knowledge is crucial to understand the mechanisms underpinning the observed social health inequalities in COVID-19 health outcomes.

In addition, much uncertainty exists regarding the aftermath of COVID-19. Although information is still building up, prior outbreaks of SARS-CoV-1 already demonstrated long-term health effects in infected patients, including persistent lung impairment and reduced exercise capacity, years after hospitalisation.[17] Also in patients that endured a SARS-CoV-2 infection, these long-term health effects, that is, the post-COVID-19 condition, are becoming more recognised as an important public health issue.[18 19] Patients with persistent COVID-19 symptoms report the persistence of physical health issues (eg, fatigue, weakness, breathlessness) as well as psychological and social health impairments.[20 21] However, evidence from large longitudinal population cohorts remains imperative to understand the full impact of COVID-19 on population health, and to provide unbiased population estimates of persisting symptomatology.

To address these existing important knowledge gaps, we aim to fulfil two research objectives: (1) Identify SE and SD factors in COVID-19 testing, infection, hospitalisation and mortality; and (2) Describe the long-term direct health impact of COVID-19 by (2A) Comparing healthcare expenditure during and after a COVID-19 infection or hospitalisation; (2B) Investigating long-term health complications or premature mortality after a COVID-19 infection or hospitalisation; and (2C) Validating the COVID-19 nomenclature developed by the National Institute for Health and Disability Insurance for the purpose of improving pandemic preparedness for future outbreaks.

Belgium has rich administrative data collections to enhance the understanding of social health inequalities and the long-term health impact of COVID-19. However, these databases provide only fragmentary perspectives on the clinical health situation, the SE/SD context or the healthcare expenditure. The interoperability and reuse of these different clinical and administrative data sources pose various technical, administrative and data security challenges. The COVID-19 pandemic has clearly demonstrated the need for high-quality social and health data collections, as well as the importance of efficient, swift and cost-effective research tools. This paper describes the necessary data linkages between health records and administrative data sources to meet the study objectives, and thus provides a useful case study for the delicate undertaking of linking highly sensitive individual-level information for social health research. The expected result is a population-based data set with a case-cohort design covering inpandemic and postpandemic health information, as well as SE/SD characteristics, and healthcare use and related costs.

METHODS AND ANALYSIS

Study design, sample size and study period

The study builds on a case-cohort study design, in which 10% (ie, a sample of over one million participants) is randomly sampled from the entire Belgian population, constituting 11 492 641 inhabitants on 1 January 2020, to ensure a robust analysis and to obtain precise estimates.[22 23] Initial sampling will be performed based on the population structure on 1 January 2020 to evaluate the risk of SARS-CoV-2 infection and COVID-19 hospitalisation, and participants from the entire study population will be followed up for a period of 4 years to evaluate the potential long-term consequences of COVID-19. The study population will include (A) Individuals from the random sample that were not identified as a case (ie, controls); (B) Individuals from the random sample that were identified as a case; and (C) Individuals that were not sampled but were identified as a case using the Belgian COVID-19 surveillance data (figure 1).

Three case definitions are determined. Tested cases are defined as individuals who had a COVID-19 PCR or antigen test. Confirmed cases are individuals who had a positive COVID-19 PCR or antigen test. Hospitalised cases are individuals who were hospitalised with a COVID-19 infection as diagnosed by a positive COVID-19 PCR or antigen test. The time frame in which cases can be identified starts 1 July 2020. Case identification stops (A) When the legal mandate for federal COVID-19 health surveillance ends or (B) At the final update in January 2026. As per the design, all cases will be added to the final study

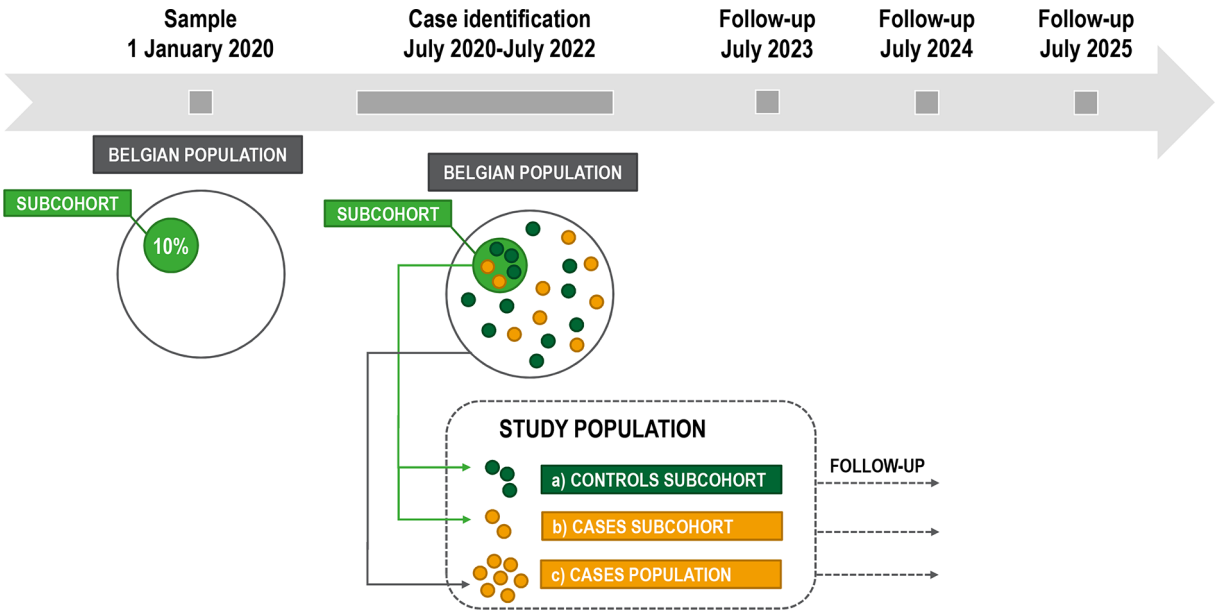

Figure 1 Schematic overview of the case-cohort study design.

population. To allow replication of the results from the different analyses that are planned in this study, a copy of the final analysis set used for the analysis will be saved on the same server.

This case-cohort design, which is typically used when studying rare diseases, has two major advantages over nested case-control studies. First of all, control persons from the sampled subcohort can be used as the comparison group for multiple study outcomes rather than identifying a new set of controls for different outcomes.[24] Second, this design might be particularly useful in an era of rising amounts of privacy-sensitive health data.[25]

Data sources

The current study builds on existing databases from three main data holders, that is, healthdata.be (Healthdata), Statistics Belgium (Statbel) and the InterMutualistic Agency (IMA).

Healthdata collects health data from primary data providers.[26] Two COVID-19-related databases are available: (1) The COVID-19 test database containing information on persons tested for COVID-19; and (2) The COVID-19 clinical hospital database containing information on hospitalised patients with a suspected or confirmed COVID-19 diagnosis. The COVID-19 test data originate from laboratories, pharmacies and physicians (general practitioners, physicians at hospital/collectivities).[27] Since May 2020, test data from the different data sources are reported in real time to Sciensano in a standardised procedure via an electronic platform developed by Healthdata. The databases include a unique individual identifier allowing us to distinguish individuals with multiple positive COVID-19 PCR or antigen test using a deduplication procedure.[5 28] The COVID-19 clinical database includes information on hospital admitted confirmed or suspected COVID-19 cases provided by all hospitals in Belgium via an admission and discharge questionnaire. The database includes information on comorbidities, treatment provided and complications. The major limitation of these data sources is their coverage. The COVID-19 test database only includes identification numbers necessary for linkage from 1 July 2020 onwards. During this period, the testing strategy was broadened to all symptomatic patients, high-risk contacts and inbound travellers from high-risk countries.[5] The database is established using different test criteria over time and does not cover asymptotic or other cases that did not present themselves for testing. To date, there is no information on the bias in coverage of the COVID-19 test database, although this is currently being investigated by Sciensano. Similarly, the COVID-19 clinical database is not exhaustive, since the registration of patients is not enforced by the authorities. A comparison with the mandatory data collection on hospital bed occupancy indicates that 55% of all COVID-19 hospitalised persons in Belgium are represented in the COVID-19 clinical database. The major strength of this database, however, is the real-time data and follow-up of patients during the pandemic.

Statbel collects, produces and disseminates statistics on the economy, society and territory in Belgium.[29] More specifically, Statbel data provide insights into the vital status, household situation, educational attainment, employment status, sector of activity, income (personal and household), and migration background at the individual level.[30] This data source covers the total population officially living in Belgium, as identified by the National Register. The major limitation of this data set is the time delay for some variables (eg, causes of death; tax information) of approximately 2 years.[31] The major strength of this database is its coverage, since all registered Belgian inhabitants are included, and its ability to include precise SE/SD and geographical information.

IMA integrates data from all seven health insurance agencies in Belgium, and contains information on the reimbursement of health services provided by general practitioners, hospitals, health specialists and other medical professionals, as well as information on the reimbursement of medication in public pharmacies.[32] The major limitation of this data set is its absence of diagnostic information and its lack of specificity for certain variables. For example, information on rehabilitation is available in terms of the number of sessions, but no information is available regarding the content of these sessions. The major strength of this data set is its coverage, and representativeness for the Belgian population. Since membership in a health insurance agency is mandatory in Belgium, the IMA data reaches almost full population coverage with 99% of all legal residents.[33] Missing data in the IMA data set is mainly attributable to a delay in timing of the registration or healthcare costs that have not been officially declared for reimbursement. Furthermore, the data provided by IMA is a well-established data set, which contains up-to-date and time-specific information for health-related events (eg, the date of hospitalisation).

Clinical COVID-19 data

Information on COVID-19 is provided by Healthdata, and is subdivided into COVID-19 test and COVID-19 clinical databases. Test-related information includes the following information: onset of symptoms, type of test, test result and the timing of the test in epidemiological weeks. Both PCR and antigen tests are included and are specified under the type of test. The hospital-related information entails three main categories: data related to hospital admission, hospital discharge and intensive care unit (ICU). Admission variables include the timing of the admission (epidemiological week), reason for admission (eg, COVID-19 or non-COVID-19 related), symptoms at admission (eg, fever, gastrointestinal), comorbidities at admission (eg, diabetes, cardiovascular disease), smoking behaviour and diagnostic methods used. Discharge (ie, discharged alive or deceased) variables include complications (eg, multiple organ failure, acute respiratory distress syndrome, sepsis), lab values (eg, $Pa0_2$, C reactive protein), received treatments and in-hospital mortality. ICU-related variables include information on the use of respiratory support (invasive vs non-invasive), and the use of extracorporeal membrane oxygenation. All hospital-related data contain information on the hospital length of stay and in-hospital length of respiratory support.

Socioeconomic and sociodemographic data

SE/SD information is provided by Statbel, and includes age, sex, civil status, household size, presence of schoolchildren in the household. In addition, the data also contain household type operationalised using the LIPRO typology (eg, single-person household, collective household).[34 35] Statbel also provides constructed variables on the individual migration background: the country of birth (eg, Belgian, other European) and the country of descent (eg, Belgian, other European), the latter also capturing parental country of birth. Different indicators for the SE context are included. Education level is based on the 2011 International Standard Classification of Education classification with eight levels (early childhood education up to doctoral education).[36] Occupational information includes working status (eg, employed, pensioned); self-employment and sector of employment operationalised using the Statistical Classification of Economic Activities in the European Community. Information on net taxable income is available on individual and household level expressed as deciles. Geographical information includes the most detailed territorial unit, the statistical sector, which can be used to link data that are available on a geographical level. Examples of data on the statistical sector include environmental data (eg, air pollution, noise exposure, green spaces) and data on indices of multiple deprivation. Lastly, mortality information is available on date of death (week of the year) and cause of death (International Statistical Classification of Diseases Revision-10 four-character subcategories).

Healthcare data

Health-related data are provided by IMA and include information on reimbursements of medications and other medical-related costs, healthcare usage, non-COVID hospitalisations, rehabilitation, GP visits, and the usage of specific COVID-19 codes.

Data linkage procedure

The data linkage process, as depicted in figure 2, entails considerable challenges in terms of data security and record linkage techniques. Multiple steps of pseudonymisation and de-pseudonymisation are needed to ensure that none of the involved parties could have access to both the sensitive research data and the national identification numbers of the Social Security (NISS), and that no data holder would be able to enrich his/her own database with data from one of the other data holders. Only the researchers are entitled to have access to the linked, pseudonymised database. Different parties are therefore involved in the (de)pseudonymisation processes:

- The electronic platform eHealth is a federal public institution aimed at improving data transfers between public health actors while protecting personal privacy and information security.[37] The main role of eHealth will be to act as Trusted Third Party (TTP) for the conversion between the NISS of all persons included in this study and their pseudonyms used in the linked data.
- The Single Point of Contact (SPOC) is the National InterMutualistic College (NIC), which is the highest consultative body of the Belgian health insurance agencies. The SPOC will be responsible for the conversion between the NISS of all members of the health insurance agencies included in the study and the pseudonyms used at the level of the insurance agencies.

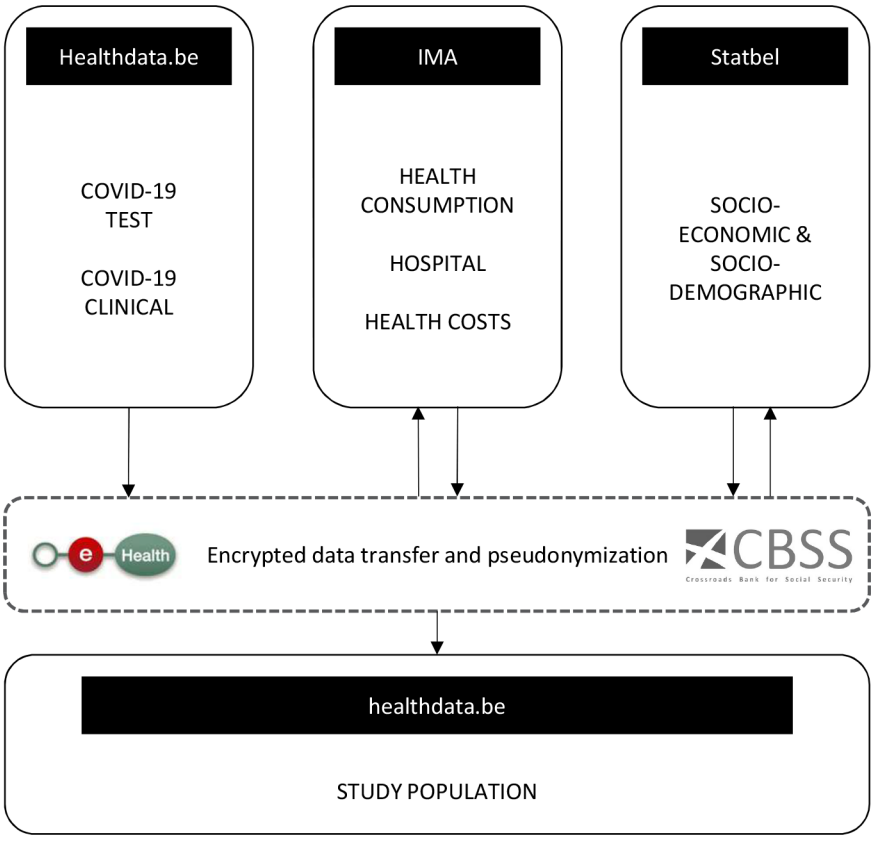

Figure 2 Schematic overview of the linkage process and involved trusted third parties. CBSS, Crossroads Bank of Social Security; IMA, InterMutualistic Agency.

- The Crossroads Bank of Social Security (CBSS), acting as TTP of the health insurance agencies. The CBSS is a federal public institution that collects, validates and manages data from a network of over 3000 social sector actors.[28] This TTP will be responsible for the conversion between the pseudonyms used at the level of the insurance agencies and the pseudonyms used in the IMA data warehouse.

The involved parties have also elaborated in accordance with art. 4.5 of the General Data Protection Regulation (GDPR) all necessary secured transfer processes that guarantee the transfer of the data of each data holder directly to Healthdata, using technical identifiers instead of the original identifiers in the data sources. The TTP eHealth plays a pivotal role in these processes by making sure that every actor in the linkage procedure has the right encrypted key to select the necessary population or variables at the right time, and by transferring the necessary conversion tables with the technical identifiers and the final pseudonyms to Healthdata where the actual linkage will take place. In doing so, the identity of study participants is protected throughout the procedure and the data flows only contain the essential information for the safe, encrypted transfer of requested variables. eHealth will also securely store the conversion key between the NISS and the final pseudonyms during the full duration of the project in order to enable the yearly follow-up data linkage. A more detailed overview of the data linkage procedure can be found in online supplemental appendix 1.

As shown in figure 3, the data linkage procedure can be divided into three major linkage steps. Because the three data holders organise their data using their specific identifiers, several (de)pseudonymisation procedures are performed by the TTPs in these three steps. The starting point for the data linkage is the random 10%-sample from the total Belgian population drawn by Statbel. This list of Statbel identifiers from the subcohort is sent to Healthdata via eHealth using a pseudonymisation procedure. Healthdata sends back a list of Healthdata identifiers for the subcohort and the unsampled cases to Statbel via eHealth using the same pseudonymisation procedure. In a second step, Statbel sends all of the available SE/SD data for the subcohort and the unsampled cases to Healthdata using a new technical identifier. The key for the technical identifier is sent securely to eHealth, where the Statbel identifier is translated to the Healthdata identifier. A list with the Healthdata and technical identifier is sent to Healthdata securely.

In the third step, IMA will transfer all available data for the sampled cases, the unsampled cases and the sampled controls via CBSS to Healthdata using a second new technical identifier. CBSS sends the key for the technical identifier to eHealth. Again, using a pseudonymisation procedure, eHealth translates the CBSS identifier to the Healthdata identifier. Using the technical identifiers,

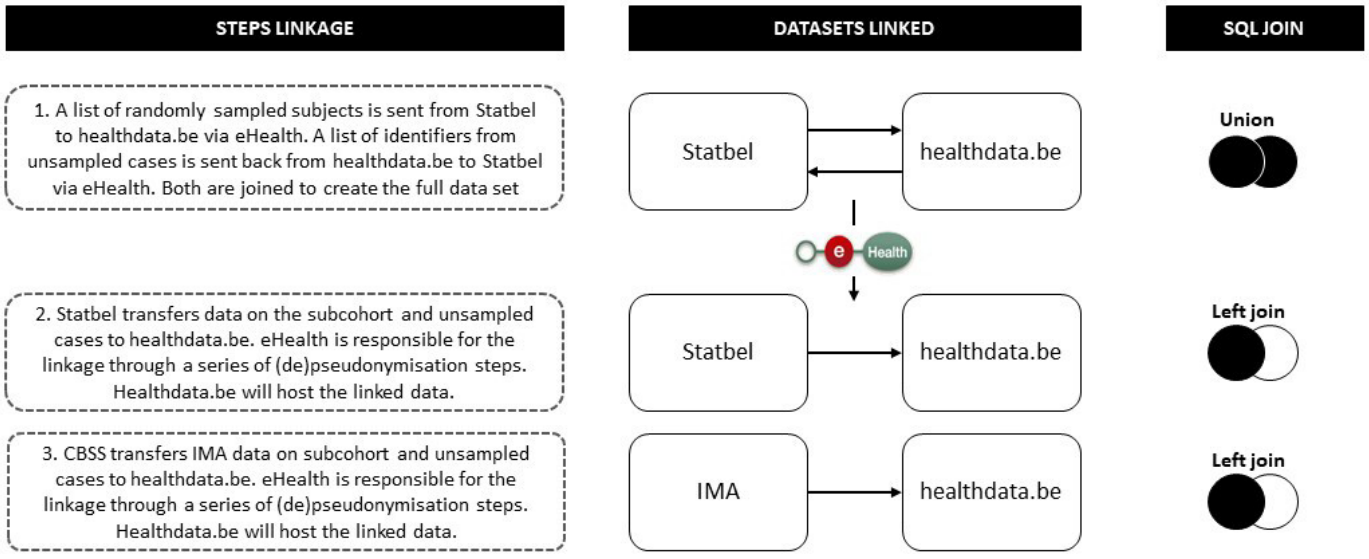

Figure 3 Data linkage of COVID-19, SE/SD and healthcare data sets. CBSS, Crossroads Bank of Social Security; IMA, InterMutualistic Agency.

Healthdata will link the received Statbel and IMA data to the COVID-19 test database and the COVID-19 clinical database. Healthdata will delete the technical identifiers after the linkage. The Healthdata identifier will be pseudonymised and a new data set specific identifier will be created. The final data set for analysis with only this new data set specific identifier will be hosted within the secure Healthdata analysis environment. To obtain up-to-date information, the acquired data set will be updated yearly following the same procedure during the follow-up period of 4 years. The keys for the pseudonymisation procedure will be stored securely at eHealth to allow individual-level follow-up. To mitigate the risk of re-identification, small-cell risk analyses will be performed by an independent partner after the total linkage procedure and after each data linkage update.

Data management

Data management tasks entail data cleaning and constructing new variables (eg, educational classification scheme) to ensure cohesion across the different analysis tasks. We have scheduled data maintenance tasks after each yearly update to check for data anomalies. Table 1 provides a summary of the data analysis plan with the planned analysis tasks per objective and per COVID-19 health outcome. Online supplemental appendix 2 provides a detailed table with the envisaged analyses per research objective. Primary outcome measures are (A) Having had a COVID-19 PCR or antigen test; (B) Having had a positive COVID-19 PCR or antigen test; and (C) Having been hospitalised with a PCR or antigen-confirmed COVID-19 diagnosis.

Statistical analysis

The statistical analysis will first describe baseline characteristics of the subcohort and identified cases using means, medians and proportions. For each of the main objectives, as listed in table 1, an appropriate statistical model was selected. To account for the possible sample selection bias, preliminary analyses will investigate the missingness mechanisms following the method of Peskoe and colleagues.[38] To account for the potential over-representation of cases compared with controls, weighting procedures will be applied to the Cox proportional hazards and logistic regression models. These weights will be calculated based on the representation of cases and controls in the overall data set compared with the population data.[22 24 39] To this end, there are packages available in a ready-to-deploy format within R, such as *coxphw*.[40] If weighting procedures are not directly available, inversed probability weighting will be applied to the models.[41] Information regarding missing data will be reported by variable. Causal diagrams and frameworks will be applied where causal relations are implied.[42] Each objective will be evaluated using two-tailed statistical tests at the 5% significance level. The analysis will mainly be conducted in the latest available version of R.[43]

Cohort characteristics

The entire Belgian cohort includes 11 492 641 inhabitants on 1 January 2020, divided over three regions with 1218 255 inhabitants in the Brussels Capital Region, 6629 143 in the Flemish Region and 3645 243 in the Walloon Region. In total, the Belgian population includes 5832 577 women (50.8%) and 5660 064 men (49.2%). An important limitation of the administrative study cohort is that some vulnerable population groups (eg, rejected asylum seekers and migrants without a valid residence permit) are not covered. In addition, the data on COVID-19 tests, infections and hospitalisations is not exhaustive possibly creating a sample selection bias.

Table 1 Data analysis plan

	Test	Infect.	Hosp.
0 Assess the data quality of the linkage by investigating the missingness mechanisms			
Modularisation of the data provenance	x	x	x
Little's test of missing completely at random			x
1 Identify SE/SD risk factors			
Absolute age-adjusted probabilities	x	x	x
Logistic binomial regression models	x	x	x
2a Compare healthcare expenditure during and after a COVID-19 outcome			
Generalized Linear Models negative binomial family distributions		x	x
2b Investigate health complications after a COVID-19 outcome			
Weighted Schoenfeld residuals test		x	x
Cox proportional hazards models		x	x
2b Investigate premature mortality after a COVID-19 outcome			
Weighted Schoenfeld residuals test			x
Cox proportional hazards models			x
2c Validate the COVID nomenclature			
Cohen's kappa and other concordance measures			x
Logistic regression models			x
Classification trees			x

'Test' refers to having had a COVID-19-related PCR or antigen test, 'Infect.' refers to having had a positive COVID-19 PCR or antigen test, and 'Hosp.' refers to having been hospitalised with COVID-19 as diagnosed with a positive COVID-19 PCR or antigen test.

Patient and public involvement

The current study is organised within the scope of the HELICON project aimed at understanding the long-term and indirect of the COVID-19 crisis on Belgian population health. A multidisciplinary team of researchers from Sciensano, Katholieke Universiteit Leuven, Université Catholique de Louvain, Universiteit Gent, Université Libre de Bruxelles and Vrije Universiteit Brussel will investigate the multidimensional impact of the crisis and social patterns therein in this 4-year project. As such, this study benefits from the active involvement of representatives of public services and patient groups in the project. To inform the general public, a dedicated project website was created, which can be consulted at https://www.brain-helicon.be/.

ETHICS AND DISSEMINATION

Because of the sensitive nature of the linked data sources, the ethical and legal basis for the construction and use of this new linked data source is extensively outlined in the data application.[44] The data linkage procedure builds on the legislation of Healthdata, Statbel, IMA, eHealth and CBSS. The legal basis for the COVID-19-related data collections was established on 25 August 2020, right before the start of the second wave in Belgium. As a result, no health information from the first wave and interwave period can be linked. The project was granted approval by the Belgian Information Security Committee, the official national body that preventively investigates compliancy with the principles of the GDPR and grants deliberations with a binding scope on the communication of, for example, personal health data to use these personal and medical data within a clearly defined framework.[44] In addition, the Ghent University Hospital ethics committee has assessed the project's consideration of ethical and legal issues and has granted their approval (B.U.N. 1432020000371). The project partners' institutions, Sciensano and Université Libre de Bruxelles, are responsible for the processing of the data. The involved researchers have access to the linked pseudonymised data. The pseudonymised data will be stored for 10 years, following the approval request to the Information Security Committee.

The processing is based on the grounds of public interest (art. 6.1 (e) of the GDPR) and, in particular, for data concerning health, for reasons of scientific research (art. 9.2 (j), of the GDPR). The GDPR gives persons whose data have been processed a right of access, rectification, deletion, restriction and objection. Since this project links pseudonymised data the risks of re-identification of persons are minimised. The involved project partners would therefore need additional information from the applicant. More information on the implementation of GDPR and the Belgian legal framework is available in the data information sheet of the project website.[45]

Preliminary results will be shared with the project's follow-up committee with representatives from the data holders, regional and federal government agencies,

public health institutions, academics, and patient groups. Their feedback will be incorporated throughout the project. Reports and scientific publications with the results of the HELICON project will be shared with partners, stakeholders, and federal and regional ministries of public health. We will present our results to diverse audiences in (inter)national conferences, policy groups, traditional media outlets, (non)academic publications, the project website and the project webinar series. All reports and presentations will be checked to prevent disclosure of confidential information. Only aggregated data (in text, tables and graphs) that guarantee the anonymity of study participants will be approved for publication.

Author affiliations
[1]Department of Epidemiology and Public Health, Sciensano, Brussel, Belgium
[2]Department of Rehabilitation Sciences, Ghent University, Gent, Belgium
[3]Interface Demography, Vrije Universiteit Brussel (VUB), Brussels, Belgium
[4]Statistics Belgium, Brussels, Belgium
[5]Research Institute of Health and Society (IRSS), Catholic University of Louvain, Louvain-la-Neuve, Belgium
[6]Department of Public Health, Ghent University, Gent, Belgium
[7]Intermutualistic Agency, Brussels, Belgium
[8]School of Public Health, Universite Libre de Bruxelles - Campus Erasme, Bruxelles, Belgium
[9]Groupe de Recherche sur les Relations Ethniques, les Migrations et l'Egalité, Université Libre de Bruxelles, Bruxelles, Belgium
[10]Department of Translational Physiology, Infectiology and Public health, Ghent University, Merelbeke, Belgium

Contributors RDP, LVdB and BD designed the study. RDP and LVdB wrote the first draft of the manuscript. DJ co-wrote the Methods and Analysis section. RDP, LVdB, YB, LC, SG, JG, DDS, DJ, YK, PL, NS, JR, AR, DVC, SV, KV and BD contributed to the conception of the work, reviewed the first draft and provided feedback. RDP and LVdB prepared the final version of the manuscript, which was approved by YB, LC, SG, JG, DDS, DJ, YK, PL, NS, JR, AR, DVC, SV, KV and BD. The corresponding authors had final responsibility for the decision to submit for publication. The corresponding authors attest that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding This work was supported by Belgian Science Policy Office (BELSPO) within the BRAIN-be 2.0 framework supporting pillar 3 Federal societal challenges (grant number B2/202/P3/HELICON).

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Data sharing is prohibited by the Belgian Information Security Committee.

ORCID iDs
Laura Van den Borre http://orcid.org/0000-0001-5709-3533
Judith Racape http://orcid.org/0000-0002-1266-0191

REFERENCES

1 Budd J, Miller BS, Manning EM, *et al*. Digital technologies in the public-health response to COVID-19. *Nat Med* 2020;26:1183–92.
2 Barnard S, Chiavenna C, Fox S, *et al*. Methods for modelling excess mortality across England during the COVID-19 pandemic. *Stat Methods Med Res* 2022;31:1790–802.
3 Van Goethem N, Vilain A, Wyndham-Thomas C, *et al*. Rapid establishment of a national surveillance of COVID-19 hospitalizations in Belgium. *Arch Public Health* 2020;78:121.
4 Renard F, Scohy A, Van der Heyden J, *et al*. Establishing an ad hoc COVID-19 mortality surveillance during the first epidemic wave in Belgium. *Euro Surveill* 2021;26:2001402.
5 Meurisse M, Lajot A, Dupont Y, *et al*. One year of laboratory-based COVID-19 surveillance system in Belgium: main indicators and performance of the laboratories. *Arch Public Health* 2021;79:188.
6 Sciensano. COVID-19 weekly epidemiological report (3 June 2022). Brussels, Belgium Sciensano; 2022. Available: https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_Weekly_report_NL.pdf [Accessed 10 Aug 2022].
7 Vestergaard LS, Nielsen J, Richter L, *et al*. Excess all-cause mortality during the COVID-19 pandemic in Europe – preliminary pooled estimates from the Euromomo network, March to April 2020. *Euro Surveill* 2020;25:2001214.
8 Nørgaard SK, Vestergaard LS, Nielsen J, *et al*. Real-time monitoring shows substantial excess all-cause mortality during second wave of COVID-19 in Europe, October to December 2020. *Euro Surveill* 2021;26:2002023.
9 Molenberghs G, Faes C, Verbeeck J, *et al*. COVID-19 mortality, excess mortality, deaths per million and infection fatality ratio, Belgium, 9 March 2020 to 28 June 2020. *Euro Surveill* 2022;27:2002060.
10 Vanthomme K, Gadeyne S, Lusyne P, *et al*. A population-based study on mortality among Belgian immigrants during the first COVID-19 wave in Belgium. Can demographic and socioeconomic indicators explain differential mortality? *SSM Popul Health* 2021;14:100797.
11 Decoster A, Leuven K, Minten T, *et al*. The income gradient in mortality during the COVID-19 crisis: evidence from Belgium. Leuven, 2020.
12 Nguyen LH, Drew DA, Graham MS, *et al*. Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *Lancet Public Health* 2020;5:e475–83.
13 Gadeyne S, Rodriguez-Loureiro L, Surkyn J, *et al*. Are we really all in this together? The social patterning of mortality during the first wave of the COVID-19 pandemic in Belgium. *Int J Equity Health* 2021;20:258.
14 Khalatbari-Soltani S, Cumming RC, Delpierre C, *et al*. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J Epidemiol Community Health* 2020;74:620–3.
15 Meurisse M, Lajot A, Devleesschauwer B, *et al*. The association between area deprivation and COVID-19 incidence: a municipality-level Spatio-temporal study in Belgium, 2020–2021. *Arch Public Health* 2022;80:1–10.
16 Decoster A, Minten T, Spinnewijn J. The income gradient in mortality during the COVID-19 crisis: evidence from Belgium. *J Econ Inequal* 2021;19:551–70.
17 Ngai JC, Ko FW, Ng SS, *et al*. The long-term impact of severe acute respiratory syndrome on pulmonary function, exercise capacity and health status. *Respirology* 2010;15:543–50.
18 Yelin D, Wirtheim E, Vetter P, *et al*. Long-term consequences of COVID-19: research needs. *Lancet Infect Dis* 2020;20:1115–7.
19 Soriano JB, Murthy S, Marshall JC, *et al*. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect Dis* 2022;22:e102–7.
20 Michelen M, Manoharan L, Elkheir N, *et al*. Characterising long term COVID-19: a living systematic review. *BMJ Glob Health* 2021;6:e005427.
21 Castanares-Zapatero D, Laurence K, Marie D, *et al*. Long COVID: pathophysiology – epidemiology and patient needs. Brussels Belgian Health Care Knowledge Centre (KCE); 2021.

22 Onland-Moret NC, van der A DL, van der Schouw YT, *et al*. Analysis of case-cohort data: a comparison of different methods. *J Clin Epidemiol* 2007;60:350–5.
23 Duran X, Vanroelen C, Deboosere P, *et al*. Social security status and mortality in Belgian and Spanish male workers. *Gac Sanit* 2016;30:293–5.
24 Sharp SJ, Poulaliou M, Thompson SG, *et al*. A review of published analyses of case-cohort studies and recommendations for future reporting. *PLoS ONE* 2014;9:e101176.
25 Casali PG, Vyas M. Data protection and research in the European Union: a major step forward, with a step back. *Ann Oncol* 2021;32:15–9.
26 Healthdata.be. About Healthdata.be. 2022. Available: https://healthdata.sciensano.be/en/about-healthdatabe [Accessed 31 Aug 2022].
27 Healthdata.be. Database COVID-19 testresults. 2021. Available: https://covid19lab.healthdata.be/quality-feedback [Accessed 30 Jun 2021].
28 Proesmans K, Hancart S, Braeye T, *et al*. COVID-19 contact tracing in Belgium: main indicators and performance, January–September 2021. *Arch Public Health* 2022;80:118.
29 Statbel. About Statbel. Available: 2022.https://statbel.fgov.be/en/about-statbel [Accessed 31 Aug 2022].
30 Statistics Belgium. Gegevenscatalogus. Lijst Van Beschikbare Variabelen en Modaliteiten (in Opbouw). 2021. Available: https://statbel.fgov.be/nl/over-statbel/privacy/microdata-voor-onderzoek/gegevenscatalogus [Accessed 27 Aug 2021].
31 Maetens A, De Schreye R, Faes K, *et al*. Using linked administrative and disease-specific databases to study end-of-life care on a population level. *BMC Palliat Care* 2016;15:86.
32 Intermutualistic Agency. Wie Zijn we? 2022. Available: https://www.ima-aim.be/-Wie-zijn-we- [Accessed 31 Aug 2022].
33 Devos C, Cordon A, Lefèvre M, *et al*. Performance of the Belgian health system. Brussels, Belgium Belgian Health Care Knowledge Centre (KCE); 2019. Available: https://kce.fgov.be/en/performance-of-the-belgian-health-system-–-report-2019 [Accessed 01 Jun 2021].
34 Van Imhoff E, Keilman N. *LIPRO 2.0: an application of a dynamic demographic projection model to household structure in the Netherlands*. Amsterdam: Swets & Zeitlinger, 1991.
35 Lodewijck E, Deboosere P. LIPRO: Een Classificatie Van Huishoudens. 2008. Available: https://publicaties.vlaanderen.be/view-file/5158
36 UNESCO Institute for Statistics. International standard classification of education: ISCED 2011. In: *Comparative Social Research*. 2012: 30.
37 eHealth. Wie Zijn Wij? 2021. Available: https://www.ehealth.fgov.be/ehealthplatform/nl/organisatie [Accessed 07 Jul 2021].
38 Peskoe SB, Arterburn D, Coleman KJ, *et al*. Adjusting for selection bias due to missing data in electronic health records-based research. *Stat Methods Med Res* 2021;30:2221–38.
39 Kulathinal S, Karvanen J, Saarela O, *et al*. Case-cohort design in practice–experiences from the MORGAM project. *Epidemiol Perspect Innov* 2007;4:1–17.
40 Dunkler D, Ploner M, Schemper M, *et al*. Weighted Cox regression using the R package Coxphw. *J Stat Softw* 2018;84:1–26.
41 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656–64.
42 Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
43 R Core Team. R: a language and environment for statistical computing. Vienna, Austria R Foundation for Statistical Computing; 2020. Available: https://www.R-project.org/
44 Information Security Committee. Beraadslaging NR. 22/014 Van 11 Januari 2022 met Betrekking tot de Mededeling Van Persoonsgegevens die de Gezondheid Betreffen Van de Healthdata COVID-19 clinic Databank, Statbel en Het Intermutualistisch Agentschap Aan Sciensano en L'Université Libre de Bruxelles met Het Oog op Het Bestuderen Van Covid -19 Hospitalisaties in Het Kader Van Het HELICON –project. 2022. Available: https://www.ehealth.fgov.be/ehealthplatform/file/view/AX54CWc4Fbc33iE1rY5a?filename=22-014-n034-HELICON-project.pdf [Accessed 10 Aug 2022].
45 HELICON. Data protection in the context of the HELICON project. 2021. Available: https://www.brain-helicon.be/docs/helicon-infosheet-en.pdf [Accessed 08 Jul 2021].