Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Data flow in the Luxembourg COVID-19 joint initiative: a blueprint for data-driven translational medicine

Irina Balaur

irina.balaur@lnds.lu

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
https://orcid.org/0000-0002-3671-895X

Piyapong Khumrin

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg; Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand
https://orcid.org/0000-0003-2098-9189

Soumyabrata Ghosh

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
https://orcid.org/0000-0003-0659-6733

Dmitry Bulaev

3. Competence Centre for Methodology and Statistics, Department of Medical Informatics, Luxembourg Institute of Health, Luxembourg https://orcid.org/0000-0002-0650-8871

Priyanka Mendon

3. Competence Centre for Methodology and Statistics, Department of Medical Informatics, Luxembourg Institute of Health, Luxembourg https://orcid.org/0000-0001-5759-4231

Kavita Rege

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
https://orcid.org/0009-0008-5992-3391

Milena Despotovic

4. Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Luxembourg https://orcid.org/0000-0001-7436-3398

Sarah Peter

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg

Rajesh Rawal

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
https://orcid.org/0000-0002-0151-9344

Jochen Ohnmacht

5. Transversal Translational Medicine, Luxembourg Institute of Health, Luxembourg
https://orcid.org/0000-0001-7109-6802

Joëlle V. Fritz

5. Transversal Translational Medicine, Luxembourg Institute of Health, Luxembourg
https://orcid.org/0000-0002-4694-3923

Marc Paul O'Sullivan

5. Transversal Translational Medicine, Luxembourg Institute of Health, Luxembourg
https://orcid.org/0000-0002-6184-3306

Carlo Haller

6. Inspection générale de la sécurité sociale, Ministère de la Santé et de la Sécurité sociale, Luxembourg

Vanessa Di Bartolomeo

6. Inspection générale de la sécurité sociale, Ministère de la Santé et de la Sécurité sociale, Luxembourg

Christophe Shinn

6. Inspection générale de la sécurité sociale, Ministère de la Santé et de la Sécurité sociale, Luxembourg

Rejko Krüger

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg; Transversal Translational Medicine, Luxembourg Institute of Health, Luxembourg; Department of Neurology and Parkinson Research Clinic, Centre Hospitalier de Luxembourg (CHL), Luxembourg
https://orcid.org/0000-0003-4258-6241

Guy Fagherazzi

4. Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Luxembourg https://orcid.org/0000-0001-5033-5966

Paul Wilmes

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg; Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Luxembourg https://orcid.org/0000-0002-6478-2924

Michel Vaillant

Competence Centre for Methodology and Statistics, Department of Medical Informatics, Luxembourg Institute of Health, Luxembourg https://orcid.org/0000-0003-4714-8128

Venkata Satagopam

venkata.satagopam@uni.lu

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
https://orcid.org/0000-0002-6532-5880

Research Article

Keywords: COVID-19, national initiative, data flow, biomedical data, socio-economical data, secure analytics environment

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7114519/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Data flow in the Luxembourg COVID-19 joint initiative: a blueprint for data-driven translational medicine

Irina Balaur[1*], Piyapong Khumrin[1,2], Soumyabrata Ghosh[1], Dmitry Bulaev[3], Priyanka Mendon[3], Kavita Rege[1], Milena Despotovic[4], Sarah Peter[1], Rajesh Rawal[1], Jochen Ohnmacht[5], Joëlle V. Fritz[5], Marc Paul O'Sullivan[5], Carlo Haller[6], Vanessa Di Bartolomeo[6], Christophe Shinn[6], Rejko Krüger[1,5,7], Guy Fagherazzi[4], Paul Wilmes[1,8], Michel Vaillant[3] and Venkata Satagopam[1*] on behalf of the CON-VINCE Consortium and the ORCHESTRA working group, the Predi-COVID Consortium and the CoVaLux Consortium

1. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
2. Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand
3. Competence Centre for Methodology and Statistics, Department of Medical Informatics, Luxembourg Institute of Health, Luxembourg
4. Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Luxembourg
5. Transversal Translational Medicine, Luxembourg Institute of Health, Luxembourg
6. Inspection générale de la sécurité sociale, Ministère de la Santé et de la Sécurité sociale, Luxembourg
7. Department of Neurology and Parkinson Research Clinic, Centre Hospitalier de Luxembourg (CHL), Luxembourg
8. Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Luxembourg

*Corresponding authors: irina.balaur@lnds.lu, venkata.satagopam@uni.lu

Abstract

CoVaLux is a unique national scientific consortium studying COVID-19, vaccination, and long-term health outcomes in Luxembourg. The program combines data from multiple research initiatives in a secure environment, establishing an effective framework for data management and analysis. Here, we delineate the development of the CoVaLux data flow, addressing critical requirements and showcasing how CoVaLux can serve as a model for similar medical research projects worldwide, particularly in pandemic preparation.

Main

The world-wide scientific efforts during the COVID-19 pandemic determined the initiation and conduction of a large number of scientific biomedical and clinical projects focused on various aspects of the COVID-19 disease, including understanding the epidemiology, transmission dynamics, immunological responses, and long-term consequences of SARS-CoV-2 infection, as

well as evaluating vaccines and treatments e.g.[1],[2],[3]. Launched in 2021, the COVID-19, vaccination and longer-term health consequences of COVID-19 in Luxembourg (CoVaLux) programme is a national programme that aims to contribute to a better understanding of the symptomatology, prevalence and socioeconomic factors of Long COVID and vaccination in Luxembourg[4]. CoVaLux involves close collaboration of different Luxembourg public research, clinical and government institutions, and integrates the national research resources generated during the pandemic, including the large-scale testing programme (LST)[5], Predi-COVID[6] and CON-VINCE/ORCHESTRA[7,8] studies. The LST was a population-wide large-scale screening programme aimed at suppressing epidemic dynamics which uncovered the importance of asymptomatic carriers in transmission[5]. The Predi-COVID[6] study is a longitudinal prospective cohort study including adults and children, positively tested for COVID-19 in Luxembourg, monitored remotely for health status and symptoms. The CON-VINCE/ORCHESTRA[7,8] program is a longitudinal cohort representative of the Luxembourgish population between 18 and 75 years, established to assess the viral spread and the immune status related to SARS-CoV-2 since April 2020. The CoVaLux programme also incorporates administrative, social and demographic data of the Luxembourgish population, centralised nationally by the General Inspectorate of Social Security[9] (Inspection générale de la sécurité sociale: IGSS). By incorporating socio-economic data from IGSS, CoVaLux enables comprehensive analyses that encompass not only health outcomes but also socio-economic factors, providing a holistic view of Long COVID's impact. Details on the CoVaLux programme and on these studies are provided in the Online Methods.

We describe the development of the CoVaLux data flow, which includes the data management steps of data harmonisation, access request, transfer, multi-source integration (from cohorts and IGSS administrative data including from the LST) as well as sharing of the aggregated results within the CoVaLux consortium. Initially, each cohort, Predi-COVID and CON-VINCE/ORCHESTRA, had an independent data access model and separate data management processes, which created redundancy in administrative processes for handling data access requests. Moreover, it was anticipated that the integration of the cohort data with the IGSS data (which has a distinct data access management) increases the complexity of the data access. In CoVaLux, we aimed to develop a data flow process to minimise the administrative overhead for the integration of data from cohorts and/or IGSS, and to enable user-friendly data analysis platforms. All steps of the CoVaLux data flow were performed accordingly to the data security and protection regulation and under strict coordination of the Data Protection Officers (DPOs) from the CoVaLux consortium to ensure conforming to the General Data Protection Regulation (GDPR) of the European Union.

The CoVaLux programme is centred on Research Questions (RQs) and corresponding hypotheses. At the core of the CoVaLux data flow, we apply the data minimisation for data sharing. The CoVaLux data flow (Figure 1) aims to address a set of RQs and corresponding hypotheses, which were proposed, refined and approved by the CoVaLux consortium, as follows.

Data harmonisation: The mapping of variables of Predi-COVID, CON-VINCE/ORCHESTRA and IGSS were performed. The data harmonisation was a key step that helped in understanding the common as well as similar variables across the datasets (details given in the Online Methods). Since the 3 databases had variations in the variable definitions, coding schemes and phrasing of the questions, data variable mapping across the 3 data dictionaries required manual

harmonization. While common demographic variables like “age” and “sex” were relatively straightforward to align (hence, having a 1-to-1 mapping), conceptually similar variables, such as those related to COVID symptoms and quality of life questions, required a manual, keyword-based approach to identify and match corresponding entries across data dictionaries.

Preparation of the RQs: The CoVaLux consortium proposed a set of RQs and corresponding hypotheses. Researchers consulted the harmonised set of variables and selected variables of interest for each RQ specifically. The selection of RQ-specific variables of interest represented a core step of the RQ management, directing the next steps of i) feasibility assessment, ii) consortium-level approval and iii) data access handling for every RQ. Following the variable selection, each RQ was checked for feasibility by performing a dedicated power calculation and checking availability of participants data and samples.

Approved RQs: After the feasibility assessment, the CoVaLux Management Board (composed of project coordinators and work package leaders) approved a final set of RQs and corresponding hypotheses, which were assigned for coordination and investigation to the working team that made the initial proposal. All further decisions and potential changes on data access were performed under the supervision and with the approval of the CoVaLux Management Board and the updates communicated to the CoVaLux consortium. The list of approved RQs includes:

1. Specific clinical, epidemiological, digital and sociodemographic factors are associated with Long-COVID in adults and children
2. Specific biological markers can predict Long COVID severity
3. Other (respiratory) pathogens, specific microorganisms and co-infections increase the risk and severity of Long COVID
4. Vaccination mandate lessens the burden on the healthcare system in subsequent waves of infection
5. Vaccination lessens the risk and severity of long COVID
6. Vaccination reduces the risk and severity of re-infection by future variants

The approved RQs were classified in two major categories:

1. *Group 1*: RQs that required accessing, integrating, and working with both cohort data collected in a research environment (referred here as cohort data) and data relating to social protection and the labor market collected in a real-world setting (referred here as IGSS data).
2. *Group 2*: RQs that required accessing, integrating, and working with cohort/research data only.

Data access requests: Having the selected list of variables required from each data source, the RQ working teams make formal data access requests. For data from the Predi-COVID and CON-VINCE/ORCHESTRA cohorts, access requests are submitted to the respective cohort-specific Data Access Committees (DACs), consisting of researchers representing different domains, while for IGSS data, access requests are made through the IGSS MicroData Platform (https://igss.gouvernement.lu/fr/microdata-platform.html).

Data minimisation: We followed this GDPR principle, and every RQ working team received access only to the variables (from Predi-COVID and CON-VINCE/ORCHESTRA) that are specifically

needed to address the respective RQs, upon the review and approval of the respective cohort DACs.

Management of sensitive and pseudonymised data within each cohort: In the CoVaLux cohorts, the personal identifiable information (PII, sensitive data) was accessed only by a specific clinical team of the corresponding cohort owner (named as Cohort Team in Figure 1), which was distinct of the technical team that had access to the pseudonymised data within the cohort database (named as Data Manager, with access to the COHORT Database in Figure 1). This clear and strict separation was one of the key measures to ensure the protection of sensitive data.

Mapping between CoVaLux data sources: For the Group 1 of RQs, requiring integration of cohort and IGSS data, the mapping between the cohort and IGSS data is envisaged to be implemented per each such RQ as follows:

1. The Data Manager of the COHORT Database (Figure 1) prepares and sends a pseudonymised identifier (RECORD ID in Figure 1) to the cohort-specific clinical team (Cohort Team in Figure 1).
2. The cohort-specific clinical team securely transfers date of birth and postal code together with the RECORD ID to IGSS (the Transfer I step in Figure 1).
3. IGSS checks the mapping between the cohort data and IGSS data, following the Transfer I (above), and notifies the RQ working team on the status. If the information shared via the Transfer I resolves without any ambiguity, successfully linking the cohort and IGSS data, there is no need to perform Transfer II. Otherwise, the IGSS team will share securely only the RECORD IDs that need further mapping to the cohort data. Then, the cohort-specific clinical team performs Transfer II to IGSS securely, using the first and last names, only for those unmapped RECORD IDs from Transfer I.
4. Having the data access request approved, IGSS transfers securely pseudonymised data for selected variables to an IGSS Virtual Desktop (VD), which is a secure virtual machine (VM) i.e. Secure Processing Environment (SPE) within the IGSS MicroData Platform. These IGSS VDs are located within the Government IT Centre[10] (Centre des technologies de l'information de l'État, CTIE) and were assigned to specific RQs. Similarly, the Data Manager transfers securely pseudonymised cohort-based data for selected variables together with the RECORD IDs to the corresponding IGSS VD.
5. Once the data transfer is finished by both the IGSS and the Data Manager of the cohort(s) to the same VD, this can be accessed by the RQ working team using a digital certificate (issued by the Government, and analogous to a digital signature). The pseudonymised data can then be linked by the Data Analyst with the help of the corresponding RECORD IDs.
6. Both the input to and output from VDs are assessed for quality and are checked for reidentification issues by the IGSS team to ensure that the variables that are imported correspond with the ones taken into consideration in the assessment of the data request and to ensure that only aggregated results are exported out of the VDs.

A key step in the CoVaLux data flow, involving multiple interdisciplinary meetings and workshops, was finding strategies for mapping records between the cohort and IGSS data. In Luxembourg, every resident has a unique national identifier, which is connected to all administrative, medical and tax information of the individual. While such an identifier would do a correct mapping for individuals, the research organisations such as UNILU and LIH were not allowed to collect it. .

Thus, for mapping the cohort data to IGSS data, we implemented the strategy for performing two separate secured transfers (steps 2 and 3, above). The first transfer includes the date of birth and postal code, and the second one is optional to solve unclear matches and includes the initials of the first and family names, respectively.

For the Group 2 of RQs, requiring the cohort data only, the cohort Data Manager collects and transfers securely pseudonymised data for selected variables to secure VMs, hosted and managed by the University of Luxembourg (UNILU). While for the RQs requiring both IGSS and cohort data (Group 1), the data access will be handled by the IGSS MicroData Platform, for the RQ needing only the cohort data (Group 2), the CoVaLux consortium implemented secure remote data access via high-performance VMs at UNILU. For this, we followed the examples of the IGSS MicroData Platform and of the infrastructure developed by the COVID-19 Task Force of Research Luxembourg[11]. The data access requests include authentication and authorisation functionalities, following the approvals from the DACs for the working teams for each RQ. In this way, the working team receives access to their RQ-specific data. Finally, once the data transfer is complete, the RQ working team has access to these secure high-performance VMs, equipped with data analytics software, to perform the data analysis. Analysis output is checked to make sure the results are non-sensitive aggregated data. Details on the CoVaLux secure environment at UNILU are given in Online Methods.

For both RQ groups, only the aggregated results of the analysis are made available to the RQ working team. The plan is to perform cross-validation of the results from different RQs as well as to strengthen collaboration between RQ working teams.

The development of the CoVaLux data flow (Figure 1) is a multi-step process involving communication and collaboration across the CoVaLux members from multiple institutes and enables controlled-access, multi-source integration, and secure analytical environment for CoVaLux data. Specifically, the harmonised multilayer data consisting of clinical phenotype data (demographics, medical history, vaccination, acute and long-term symptoms severity, comorbidities etc.), molecular data, medico-administrative data and mobile app based real world data, from Predi-COVID**,** CON-VINCE/ORCHESTRA and IGSS, were integrated into a secure platform, using secure virtual machines (called VDs at IGSS or VMs at UNILU), and made available to the CoVaLux members to perform further data analysis. All the CoVaLux data flow steps, including communications with the IGSS MicroData Platform, were secured and access-controlled with a multi-factor authentication layer. As reported in Table 1, the CoVaLux data flow is consistent with the “Safes framework” of the Secure Process Environment (SPE) and Trusted Research Environment (TRE) recommended by the European Health Data Space (EHDS)[12] and the Second Joint Action Towards the European Health Data Space (TEHDAS2)[13], which focuses on providing access to a secure analytics environment, where the algorithms are brought by researchers to the data for analysis rather than the data is transferred across research institutes [14,15].

Table 1 – Alignment of the “Safes framework” of the SPE and TRE and the CoVaLux data flow

Safes	**Brief summary**	**Implementation in the CoVaLux data flow**
Safe People	Only trained and approved researchers can access the data	RQ-specific analytics team(s) within the CoVaLux consortium

Safe Projects	Only approved projects can use the data. The research should deliver clear public benefit.	Approved RQs on COVID-19 and Long COVID and data access requests procedures to IGSS and to cohort DACs
Safe Settings	Only controlled and secured settings/ environment (both physical and digital) for data access	Secure virtual machines (VDs at IGSS and VMs at UNILU) Data flows are encrypted, secured and controlled.
Safe Data	Only de-identified data is accessible to researchers	Curated, pseudonymised and RQ-specific data accessible to the RQ-specific analytic team, following the data minimisation
Safe Outputs	Only outputs checked and ensuring that no individual can be identified	Manual output checks and only aggerated results shared as output

The CoVaLux programme offers a valuable example of national-level Long COVID research by combining clinical, epidemiological, and administrative data within an integrated healthcare research system. Through population-based cohorts such as Predi-COVID and CON-VINCE, linked with socio-economic data from the IGSS, CoVaLux enables detailed, longitudinal analysis of the long-term impacts of SARS-CoV-2 infection. Luxembourg's smaller size together with the centralized infrastructure allow for more streamlined data integration, whereas larger countries, despite having coordinated national initiatives, often face additional complexities in achieving comprehensive data harmonization.

We believe that CoVaLux’s strategies can be applied in other projects managing biomedical and administrative data in similar national or international settings. The CoVaLux data flow serves as a model for future epidemiological studies requiring linkage of cohort data to socio-economic information from IGSS. Moreover, CoVaLux is setting a stage for successful implementation of SPE following the GDPR, EHDS, TEHDAS recommendations. As such, CoVaLux holds international relevance for guiding data-driven public health policy and the development of integrated surveillance systems.

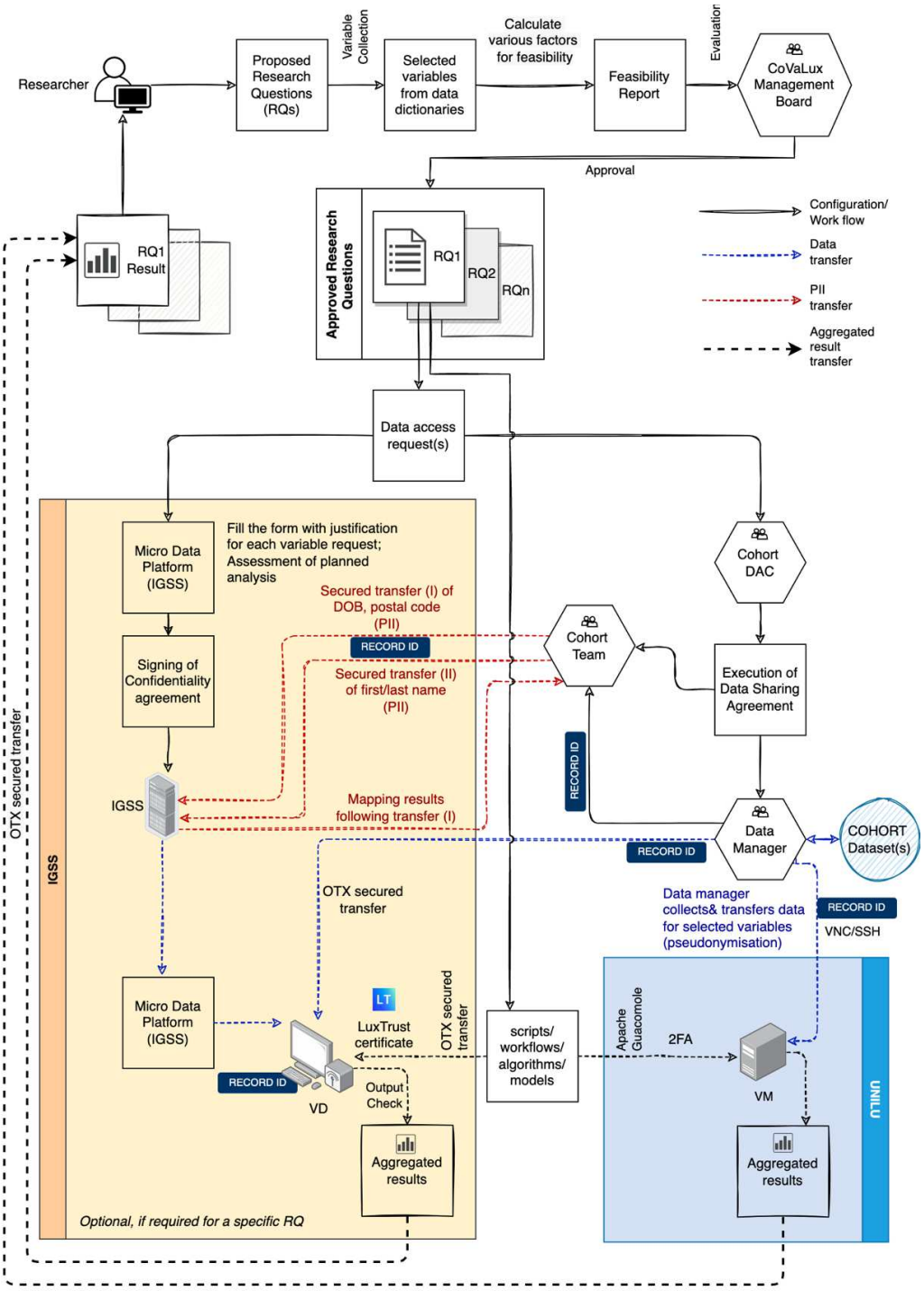


Figure 1: *The CoVaLux Data flow is Research Question (RQ) and corresponding hypothesis centric. Selected variables from approved RQs are communicated to data providers for approvals from respective cohort-specific Data Access Committee (DAC) and/or via the IGSS MicroData Platform. Upon cohort-specific DAC and IGSS approvals, the necessary cohort data to address a RQ is pseudonymised and securely made available to IGSS Virtual Desktop (VD) or Virtual Machine (VM) at UNILU, based on the data sources and computational needs. If mapping between IGSS and cohort data is needed, this is performed iteratively, in a secured, encrypted and controlled environment. After data analysis and/or running a model, only aggregated results are made available to the respective RQ research team. Data flows are secured and encrypted.*

Online Methods

The CoVaLux Consortium and goals: The CoVaLux programme involves close collaboration among Luxembourg's public research institutions, including Luxembourg Institute of Health (LIH), University of Luxembourg (UNILU), Luxembourg Institute of Science and Technology (LIST), National Health Laboratory (LNS) and Luxembourg Institute of Socio-Economic Research (LISER), clinical partners such as the Centre Hospitalier de Luxembourg (CHL) and the Centre Hospitalier Neuro-Psychiatrique (CHNP), and governmental institutes such as General Inspectorate of Social Security (Inspection générale de la sécurité sociale, IGSS) and Luxembourg Ministry of Health (MoH). Thus, the CoVaLux consortium leverages expertise in medicine, virology, immunology, public health, psychology, environmental microbiology and data science to tackle key research questions pertaining to the longer-term health consequences of COVID-19 and vaccination in Luxembourg.

The CoVaLux programme aims to enhance prevention, diagnosis, prognosis, support and treatment of the COVID-19 in Luxembourg. The CoVaLux programme also targets investigation of the short- to mid-term impact of vaccination in relation to the evolution of immunity and breakthrough infections by variants, and how this knowledge can be used to inform the future vaccination strategy in Luxembourg.

Data sets integrated in the CoVaLux programme: PrediCOVID, CON-VINCE/ ORCHESTRA, IGSS: Within the CoVaLux framework, time series data from Predi-COVID, CON-VINCE/ORCHESTRA, and the extensions of Predi-COVID & ORCHESTRA as part of CoVaLux studies were collected by the teams from LIH and UNILU-Luxembourg Centre For Systems Biomedicine (LCSB).

Predi-COVID[1] has 35 visits including clinical questionnaires, self-reporting questionnaires, sampling visits, long COVID questionnaires. Participants were actively followed for 15 days from the time of confirmation of diagnosis, whether they were at a hospital or at home in isolation or quarantine. Short evaluations were also performed at week 3 and week 4, and then monthly for a period up to 11 months to assess potential long-term consequences of COVID-19. The last monthly questionnaires were more detailed to focus on long-COVID symptoms evaluation and were completed at M12, M15 and M24. All participants were followed digitally and a selected cohort participated in sample collection (blood, stool, saliva, sputum, hair, swab) at different stages of the study. The number of participants attending the baseline visit of Predi-COVID study was 1259 (excluding all participants who withdrew consent), resulting in 21142 records (clinical questionnaires, self-reporting questionnaires, sampling visits, long COVID questionnaires) captured in REDCap (Van Der Bilt University) and Ennov Clinical (Ennov) databases. There were 1337 variables recorded over all visits.

CON-VINCE[2] has seven follow-up visits and ORCHESTRA continued follow-up with most participants from CON-VINCE for four more visits over two years. The participants were tested for SARS-CoV-2 by rRT-PCR and for anti-SARS-CoV-2 antibodies presence at all sampling visits in CON-VINCE, and the collection of stool samples was optional at each sampling visit. For ORCHESTRA, dried blood spots (DBS) and stool was sampled. Anti-SARS-CoV-2 antibody presence was checked for all DBS returned. There were 1865 individuals enrolled into CON-VINCE/ORCHESTRA study, follow-up over 11 study visits with various attendance rates, totalling

in 16080 questionnaires completed (fully or partially), and recorded in REDCap database. The number of variables recorded over all visits was 5238.

The General Inspectorate of Social Security (IGSS) (https://igss.gouvernement.lu) is placed under the authority of the minister responsible for health and social security. The IGSS helps to draft legislative and regulatory measures on social security. One important task of IGSS is to supervise social security institutions which, under the laws and regulations, is carried out by the government or a member of the government. The IGSS also assists social security institutions in the process of implementing good governance by referring to methods and techniques generally accepted on the basis of recognised international standards. Moreover, IGSS carries out analyses and studies in order to assess and plan social welfare schemes both at national and international level. The IGSS MicroData Platform on Labour and Social Protection is used to make pseudonymised microdata available for research in the areas of employment and social protection. This is administrative data, managed centrally by IGSS as part of its mandate, and collected for the most part by Luxembourg's social security institutions. For researchers, the IGSS facilitates access to nationally collected COVID-19 data from the Luxembourg Health Directorate. Within the Covalux framework, data analysis on the MicroData Platform integrates deidentified administrative data on social protection and employment alongside COVID-19 data.

In Luxembourg, both the specific COVID-19 legislation and the general law on infectious diseases provide the legal basis for the Health Directorate's collection of COVID-19 data. Firstly, the specific Covid law grants to the Health Directorate the authority to collect and process personal data related to COVID-19 for purposes such as epidemiological monitoring and contact tracing. Secondly, the law on the mandatory reporting of infectious diseases requires laboratories to report test results for certain diseases, including COVID-19, to the Health Directorate. The specific Covid law permits IGSS to be the recipient of the processed data, which it pseudonymizes for scientific research purposes, historical purposes, and statistical purposes.

Development of the CoVaLux harmonised dataset: We developed an integrated umbrella cohort across the Predi-COVID, CON-VINCE/ORCHESTRA, and IGSS dedicated to the CoVaLux study to facilitate data control and management. The data dictionaries were developed for each cohort. The Long COVID Glossary was built for the CoVaLux programme using international standard ontologies, controlled terminologies such as defined in the eTRIKS Starter Pack[16] and the IMI FAIRplus project (https://faircookbook.elixir-europe.org/[17,18]) on top of the International Severe Acute Respiratory and emerging Infection Consortium (ISARIC: https://isaric.org/) database definition for the both CON-VINCE/ORCHESTRA and Predi-COVID data. Further, the pseudonymized data was curated and mapped to Long COVID Glossary to facilitate the data interoperability as well as cross-cohort data analysis and result validation.

Cohort harmonisation for Long COVID RQs: During CoVaLux, the two data cohorts were homogenised based on time elapsed since SARS-CoV-2 infection. The aim was to select a population of interest for addressing RQs on Long COVID, which is one of the major focuses of the CoVaLux programme. The dates of infection for CON-VINCE/ORCHESTRA participants were derived, allowing for filtering of participants in acute stage or matching by the number of months since infection with Predi-COVID participants.

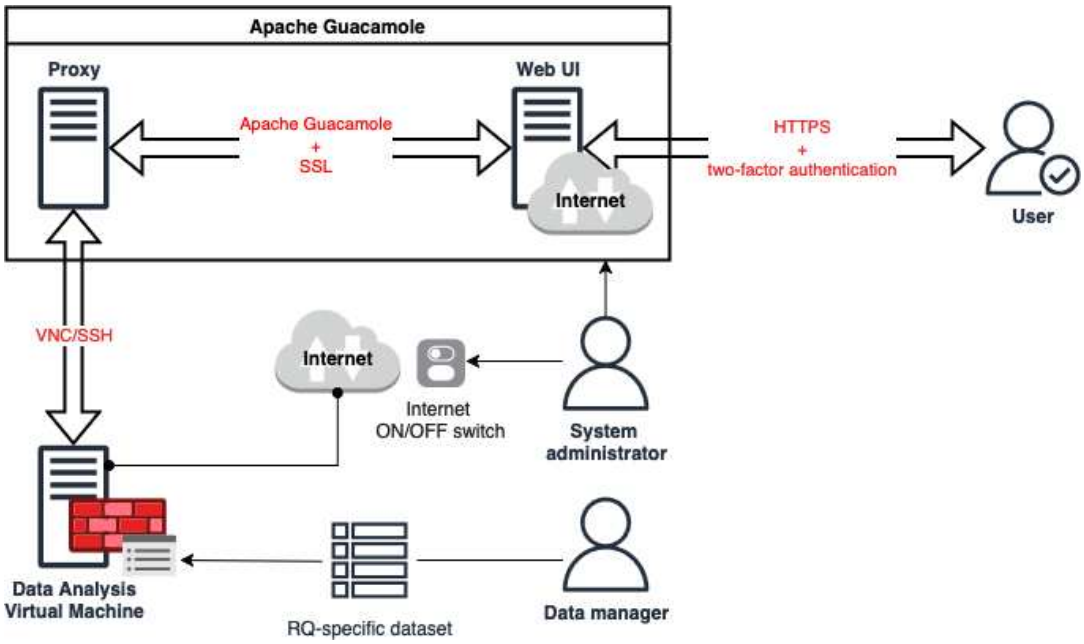
Secured and encrypted system of data and metadata management (storage, access, analysis) in the Predi-COVID and CON-VINCE/ ORCHESTRA cohorts:

Data and metadata of each cohort are managed by secured and encrypted storage systems. The CON-VINCE/ORCHESTRA data are stored on servers hosted at UNILU and the Predi-COVID data are stored on servers hosted at LIH. The servers were secured through a three-layer security framework: user authentication at the application level, a firewall at the operating system level, and a reverse proxy at the network level. Only encrypted communications were allowed between data users and the server. IP range restriction and/or two-factor-authentication was implemented for servers hosting sensitive information, wherever possible. Data backup was set up to prevent data loss.

The RECORD ID mapping (Figure 1) was kept in the Generic Pseudonym Administration Service (gPAS) hosted by LIH for Predi-COVID and by UNILU for CON-VINCE/ORCHESTRA cohorts. The gPAS applies privacy-enhancing techniques supporting privacy-by-design and ensuring non-attribution and allows storing directly person identifying data separately and securely from medical data and supports the data controller to meet the GDPR's data security requirements (Art. 32 lit. 1 EU GDPR).

In addition to the cohort data, data-governance-related metadata to fulfil regulatory compliance were stored in a central metadata repository hosted at the UNILU site. The data dictionary containing metadata and variable description are managed in an online system with a domain access control and login authentication.

For every RQ from Group 2 (where only cohort data is requested), a dedicated secure VM at UNILU was set up with software requirements such as Stata or IBM-SPSS, as requested by the RQ working team. The data stewards set up and provide instructions for the use of an encrypted transfer channel using IBM-Aspera tool to allow secure data transfer by the cohort data provider. Once the data is transferred, the Data Manager moves the data to the VM using internal secured shell (ssh) access. Only the specific RQ working team is authorised to access the VM remotely via a web platform, which allows one user connection at a time. Apache Guacamole, a clientless remote desktop gateway, enables access to these VMs via a web browser interface. Access to this gateway is secured by two-factor authentication. Logged in users can access the graphical desktop interface of VMs for which they have authorised access to and use the preinstalled tools to analyse the data. The VMs remain disconnected from the internet access, so any further data transfers from or to the VMs is blocked. Supplementary Figure1 indicates a graphical description of the processes managed within the RQ-specific secure VMs at UNILU.



Supplementary Figure1– *Schematic representation of the processes managed within the RQ-specific secure VM at UNILU. First, a System administrator sets-up a dedicated secure virtual machine (VM) for a RQ, following requirements from a RQ working team. The Data manager moves the RQ-specific dataset to the VM, allowing for further analytics. Once the data transfer is finished, the VM is decoupled from the internet, and only authorised users from the RQ working team can access remotely the VM, via the Apache Guacamole system and following two-factor authentication. The research team can perform data analysis for the specific RQ.*

CoVaLux Consortium (alphabetically listed by first name)

Achilleas Pexaras[3], Alessia Fiorelli[5], Alexander Hundt[3], Alexander Skupin[2], Alexey Kolodkin[3], Alexia Mendibide[3], Ali Znati[3], Aljosa Celebic[3], Amir Sharify[3], Amy Parrish[5], Ana Carolina Teixeira Santos[1], Anja Leist[1], Antonio Cosma[3], Armin Rauschenberger[3], Arnaud D'Agostini[3], Atte Aalto[2], Aurélie Fischer[3], Bianca Dragomir[3], Brian De Witt[3], Brice Appenzeller[3], Carlos Gamio[3], Carolina Coimbra[3], Cécile Walczak[4], Christelle Bahlawane[3], Christophe Shinn[8], Diana Pauseiro[3], Diego Garcia[3], Dmitry Bulaev[3], Enzo Bieber[5], Estelle Henry[3], Eve Herkenne[3], Feng Q. Hefeng[3], François Ancien[2], Françoise Neu[3], Frank Glod[3], Gaël Hamot[3], Ganesh Manoharan[3], Geeta Acharya[3], Gessica Contesotto[3], Guy Fagherazzi[3], Henry-Michel Cauchie[4], Irina-Afrodita Balaur[2], Isabelle Ernens[3], Jasmin Schulz[3], Jérôme Graas[3], Jéssica Fontes Estrella[5], Jesús Fuentes[2], Joao Araujo[5], Jochen Ohnmacht[3], Joëlle Fritz[3], Jorge Goncalves[2], Katy Beaumont[3], Kavita Rege[2], Laetitia Schramm[2], Laura Georges[3], Laurent Heirendt[2], Laurent Prévotat[3], Leslie Ogorzaly[4], Lucie Remark[3], Manon Gantenbein[3], Marc O'Sullivan[3], Marc Suhrcke[6], Markus Ollert[3], Matteo Pinna Pintor[1], Maud Theresine[3], Mélanie Vausort[3], Michaël Schnell[3], Michel Vaillant[3], Milena Despotovic[3], Myriam Alexandre[3], Olga Kofanova[3], Olivia Domingues[3], Olivia Roland[3], Paul Wilmes[1,2], Pauline Lambert[3], Piyapong Khumrin[2,9], Priyanka Mendon[3], Rejko Krüger[1,3,7], Sabrina Deroo[5], Sabrina Saracino[3], Saïda Mtimet[3], Sandrine Munoz[1], Soumyabrata Ghosh[2], Sylvianne Poncelet[3], Tainà Marques[2], Therese Staub[7], Thi Van Trang Nguyen[3], Torsten Bohn[3], Tuesday Lowdnes[2], Ute Aurbach[5], Valentin Groues[2], Valerie Voorsluijs[2], Vanessa Di Bartolomeo[8], Venkata Pardhasaradhi Satagopam[2], Victoria Lorentz[3], Wim Ammerlaan[3], Xin He[2], Yvan Devaux[3], Zainab Zafar[2]

1 University of Luxembourg, Esch-sur-Alzette, Luxembourg

2 Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-Belval, Luxembourg

3 Luxembourg Institute of Health, Strassen, Luxembourg

4 Luxembourg Institute of Science and Technology, Luxembourg, Luxembourg

5 Laboratoire National de Santé, Dudelange, Luxembourg

6 Luxembourg Institute of Socio-Economic Research, Esch-sur-Alzette, Luxembourg

7 Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg

8 General Inspectorate of Social Security (IGSS), Luxembourg, Luxembourg

9 Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

CON-VINCE CONSORTIUM and ORCHESTRA working group

Tamir Abdelrahman[1], Geeta Acharya[2], Gloria Aguayo[2], Pinar Alper[3], Wim Ammerlaan[2], François Ancien[3], Ariane Assele-Kama[2], Christelle Bahlawane[2], Katy Beaumont[2], Nadia Beaupain[2], Lucrèce Beckers[2], Camille Bellora[2], Fay Betsou[2], Luc Biver[4], Sandie Boly[2], Dirk Brenner[2], Henry-Michel Cauchie[5], Eleftheria Charalambous[2], Emilie Charpentier[2], Estelle Coibion[1], Sylvie Coito[6], Delphine Collart[5], Manuel Counson[2], Brian De Witt[2], Antonelle Di Pasquale[4], Olivia Domingues[2], Claire Dording[2], Jean-Luc Dourson[7], Bianca Dragomir[2], Tessy Fautsch[2], Jean-Yves Ferrand[2], Thibault Ferrandon[7], Ana Festas Lopes[2], Guillaume Fournier[1], Joëlle V. Fritz[2], Manon Gantenbein[2], Piotr Gawron[3], Laura Georges[2], Soumyabrata Ghosh3, Stéphane Gidenne[6], Enrico Glaab[3], Clarissa P. C. Gomes[3], Borja Gomez Ramos[3], Vyron Gorgogietas[3], Jérôme Graas[2], Valentin Groues[3], Wei Gu[3], Gael Hamot[2], Anne-Marie Hanff[2], Maxime Hansen[2], Linda Hansen[8], Lisa Hefele[2], Laurent Heirendt[3], Ahmed Hemedan[3], Estelle Henry[2], Margaux Henry[2], Eve Herkenne[2], Sascha Herzinger[3], Christiane Hilger[2], Laetitia Huiart[2], Alexander Hundt[2], Judith Hübschen[2], Gilles Iserentant[2], Philipp Jägi[9], Anne Kaysen[3], Piyapong Khurmin[3], Fédéric Klein[7], Tommy Klein[4], Stéphanie Kler[2], Alexey Kolodkin[2], Rejko Krüger[2], Pauline Lambert[2], Jacek Jaroslaw Lebioda[3], Sabine Lehmann[2], Marie Leick[1], Anja Leist[10], Morgane Lemaire[2], Andrew Lumley[2], Annika Lutz[10], João Manuel Loureiro[8], Monica Marchese[2], Tainà Marques[3], François Massart[3], Patrick May[3], Maura Minelli[2], Alessandra Mousel[2], Maeva Munsch[2], Sophie Mériaux[2], Friedrich Mühlschlegel[1], Mareike Neumann[2], Trang Nguyen[2], Beatrice Nicolai[8], Marc Paul O'Sullivan[2], Leslie Ogorzaly[5], Jochen Ohnmacht[2], Christiane Olesky[3], Markus Ollert[2], Claire Pauly[3], Laure Pauly[3], Lukas Pavelka[3], Christian Penny[5], Magali Perquin[2], Achilleas Pexaras[2], Palma di Pinto[6], Marie France Pirard[1], Jean-Marc Plesseria[2], Guilherme Ramos Meyers[3], Armin Rauschenberger[3], Lucie Remark[2], Antonio Rodriguez[4], Basile Rommes[3], Kirsten Rump[3], Estelle Sandt[2], Bruno Santos[2], Venkata P. Satagopam[3], Aurélie Sausy[2], Margaux Schmitt[2], Christiane Schmitt[9], Reinhard Schneider[3], Valerie Schröder[3], Serge Schumacher[4], Alexandra Schweicher[8], Sneeha Seal[2], Jean-Yves Servais[2], Florian Simon[2], Amna Skrozic[3], Chantal Snoeck[2], Kate Sokolowska[2], Lara Stute[3], Hermann Thien[2], Stéphane Toll[9], Noua Toukourou[3], Christophe

Trefois[3], Johanna Trouet[2], Nguyen Trung[1], Jonathan Turner[2], Michel Vaillant[2], Daniela Valoura Esteves[2], Carlos Vega Moreno[3], Charlène Verschueren[2], Maharshi Vyas[3], Claus Vögele[10], Cécile Walczak[5], Xinhui Wang[3], Femke Wauters[8], Bernard Weber[9], Emilie Weibel[7], Tania Zamboni[2]

1: Laboratoire National de Santé, Dudelange, Luxembourg

2: Luxembourg Institute of Health, Strassen, Luxembourg

3: Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-Belval, Luxembourg

4: TNS-ILRES, Bertrange, Luxembourg

5: Luxembourg Institute of Science and Technology, Luxembourg, Luxembourg

6: Ketterthill, Belvaux, Luxembourg

7: BioNeXt, Leudelange, Luxembourg

8: Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg

9: Laboratoire Réunis, Junglinster, Luxembourg

10: University of Luxembourg, Esch-sur-Alzette, Luxembourg

Acknowledgement

We would like to give special thanks to all participants of the Predi-COVID and the CON-VINCE studies and to each volunteer that helped in sample processing or sample kit preparation. We thank the participants of the ORCHESTRA study in Luxembourg for their important contribution to research.

We also acknowledge the involvement of the interdisciplinary and inter-institutional study team that contributed to Predi-COVID. The full list of the Predi-COVID team can be found here: https://www.lih.lu/en/predi-covid-project-team/.

We acknowledge the joint effort of the CON-VINCE team involved in sample and data collection.

We would like to thank the Research Luxembourg COVID-19 Task Force (Ulf Nehrbass, Paul Wilmes, Frank Glod, Lars Geffers, Jasmin Schulz, Henry-Michel Cauchie, Rudi Balling) for their overall support of the CON-VINCE and Predi-COVID studies.

We are very grateful for the financial support by the Fonds National de la Recherche and the André Losch Foundation, which enabled us to carry out the project. We acknowledge the joint effort of the CoVaLux Consortium members from the partner institutions contributing to the CoVaLux programme. We thank IGSS for support on the CoVaLux programme and on providing feedback on various aspects of the data flow. We also would like to thank the Ministry of Health of Luxembourg as well as the Directorate of Health of Luxembourg for their support.

A special thanks also to the data protection officers Sandrine Munoz and Laurent Prévotat.

Furthermore, we would like to acknowledge the whole Communication teams involved within CON-VINCE, especially Sabine Schmitz, Arnaud D'Agostini, Didier Gossens, Hélène Jacuszin for their excellent work and support during the implementation and execution of CON-VINCE.

We would like to thank Philippe Lamesch for important and successful fundraising for research on COVID-19 in Luxembourg.

Funding

The Predi-COVID study was supported by the Luxembourg National Research Fund (FNR) (Predi-COVID, grant number 14716273), the André Losch Foundation and by European Regional Development Fund (FEDER, convention 2018-04-026-21).

The CON-VINCE Study was funded by the Luxembourg National Research Fund (FNR 14716281/CON-VINCE/Kruger) and the André Losch Foundation (Luxembourg). The funders had no role in the design and conduct of the study, nor in the decision to prepare and submit the manuscript for publication. The ORCHESTRA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167. The views expressed in this paper are the sole responsibility of the authors and the Commission is not responsible for any use that may be made of the information it contains.

The work was further supported by the Luxembourg Government through the CoVaLux programme (16954531).

DECLARATIONS

Ethics approval and consent to participate:

The CoVaLux programme was approved by the national research ethics committee (Comité National d'Ethique de Recherche, CNER No. 202209/06) and the Ethics Review Panel of the University of Luxembourg (ERP 24-022 CoVaLux).

The Predi-COVID study was approved by the national research ethics committee (CNER No. 202003/07). The Predi-COVID study was conducted in full conformance to the principles of the 1964 'Declaration of Helsinki' and was registered as Clinicaltrials.gov NCT04380987 (date of registration: 2020-05-07).

The CON-VINCE collection was conducted according to the Ethical Principles for Medical Research Involving Human Subjects, as stated in the 2013 revised version of the 1964 World Medical Association Declaration of Helsinki and registered as Clinicaltrials.gov NCT04379297 (data of registration 2020-04-10). The national research ethics committee (Comité National d'Ethique de Recherche, CNER) reviewed and the Luxembourgish Ministry of Health approved the study (CNER No. 202004/01 and 831x6ce0d). The ORCHESTRA study was reviewed by CNER and approved by the Luxembourgish Ministry of Health (CNER No. 202106/03 and 83ax20074, respectively).

All participants completed an electronic Informed Consent Form and had the right to withdraw from the study at any time.

Author contribution

IB and PK contributed to the development of the methodology and prepared the initial draft of the manuscript. SG, DB and PM contributed to data curation and methodological development. DB also contributed to the preparation of the original draft. KR, SP, RR, JO and CH contributed to the development of the methodology and to reviewing and editing the manuscript. SG, IB, DB and SP also contributed to the development of Figures 1 and Supplementary Figure1. SP, CH and CS contributed to software development, and VDB provided supervision. MD was responsible for project administration and contributed to drafting and revising the manuscript. JVF and MPO contributed to the investigation, methodology development and provision of resources, and supervised aspects of the work while contributing to manuscript revisions.RK, GH, PW, MV and VS conceptualized the study, acquired funding, and contributed to the investigation, methodology development and provision of resources, while supervising the project and reviewing and editing the manuscript. MV additionally contributed to the preparation of the original draft. All authors read and approved the final manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

Reference

1. Narahari, A. K. *et al.* Analysis of National Institutes of Health Funding for the COVID-19 Pandemic. *Open Forum Infectious Diseases* **11**, ofae064 (2024).
2. Coronavirus research and innovation. https://research-and-innovation.ec.europa.eu/research-area/health/coronavirus_en.
3. WHO Coronavirus Network. https://www.who.int/groups/who-coronavirus-network.
4. CoVaLux project. *researchluxembourg* https://www.researchluxembourg.org/en/covid-19-task-force/projects/covalux/.
5. Wilmes, P. *et al.* SARS-CoV-2 transmission risk from asymptomatic carriers: Results from a mass screening programme in Luxembourg. *The Lancet Regional Health – Europe* **4**, (2021).
6. Fagherazzi, G. *et al.* Protocol for a prospective, longitudinal cohort of people with COVID-19 and their household members to study factors associated with disease severity: the Predi-COVID study. *BMJ Open* **10**, e041834 (2020).
7. Tsurkalenko, O. *et al.* Creation of a pandemic memory by tracing COVID-19 infections and immunity in Luxembourg (CON-VINCE). *BMC Infectious Diseases* **24**, 179 (2024).

8. Novelli, S. *et al.* Waning of anti-SARS-CoV-2 antibodies after the first wave of the COVID-19 pandemic in 2020: A 12-month-evaluation in three population-based European studies. *PLOS ONE* **20**, e0320196 (2025).
9. General Inspectorate of Social Security - The Luxembourg Government. http://igss.gouvernement.lu/en.html (2024).
10. Government IT Centre - The Luxembourg Government. http://ctie.gouvernement.lu/en.html (2024).
11. Covid-19 Task Force. *researchluxembourg* https://www.researchluxembourg.org/en/covid-19-task-force/.
12. The European Health Data Space (EHDS). https://www.european-health-data-space.com/.
13. Second Joint Action Towards the European Health Data Space – TEHDAS2. *Tehdas* https://tehdas.eu/.
14. Hubbard, T., Reilly, G., Varma, S. & Seymour, D. *Trusted Research Environments (TRE) Green Paper*. https://zenodo.org/records/4594704 (2020) doi:10.5281/zenodo.4594704.
15. Alliance, U. H. D. R. & NHSX. *Building Trusted Research Environments - Principles and Best Practices; Towards TRE Ecosystems*. https://zenodo.org/records/5767586 (2021) doi:10.5281/zenodo.5767586.
16. eTRIKS Standards Starter Pack Release 1.1 April 2016. https://zenodo.org/records/50825.
17. Rocca-Serra, P. *et al.* The FAIR Cookbook - the essential resource for and by FAIR doers. *Sci Data* **10**, 292 (2023).
18. Rocca-Serra, P. *et al.* D2.1 FAIR Cookbook. https://zenodo.org/records/6783564 (2022).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure1.png