

EHDS

EU4H-2021-PJ2
EHDS2 Pilot ("HealthData@EU pilot")
101079839

D8.1

Recommendations on application of privacy enhancing technologies, data security and node compute capabilities

December 2024



**Co-funded by
the European Union**

Disclaimer: Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

1. Document information

1.1. Authors

Name	Affiliation
Ana M Martin-Moreno	ES
Anne Heidi Skogholt	HDIR
Camille Cloitre	HDH
Carlos Tellería-Orríols	IACS
Charles-Andrew Vande Catsyne	Sciensano
Christina Hilmarsen	NIPH
Eva García Álvarez	BBMRI-ERIC
Elvira Bräuner	Danish Medicines Agency (review phase)
Ionut Florin Sava	ECDC
Juan González-García	IACS
Luís Alves de Sousa	ECDC
Mélotie Bernaux	EC SANTE C1 (review phase)
Owe Langfeldt	EC SANTE C1 (review phase)
Petr Holub	BBMRI-ERIC
Romina Royo	BSC
Roxana Arzideh	HDL
Tala Haddad	Orphanet INSERM US-14
Tuomo Nieminen	THL (review phase)
Petteri Hovi	THL (review phase)
Zdenka Dudová	MU/BBMRI.cz

1.2. Document history

Date	Version	Editor	Change	Status
02/05/2024	0.1	Eva Garcia	First TOC draft	Draft
13/05/2024	0.1.1	Anne Heidi Skogholt Carlos Tellería Christina Hilmarsen Juan González García Zdenka Dudová + meeting attendees	PET TOC	Draft
23/05/2024	0.1.2		TOC	Draft
05/07/2024	1	All authors	First draft	Draft
12/08/2024	1.1	All authors	First review by WP8, WP9 and EC	Draft
12/2024	2	All authors	Integration of final steps of the use cases	Preliminary final version
03/2025	3	Eva Garcia, BBMRI Petr Holub, BBMRI Irene Schlünder, BBMRI Mélodie Bernaux, EC SANTE C1 Owe Langfeldt, EC SANTE C1 Guillaume Byk, EC SANTE C1 Jerome de Barros, EC SANTE C1	Document review	Final version

Accepted in Project Steering Group on *16 December 2024*

Copyright Notice

Copyright © 2022 EHDS2 Consortium Partners. All rights reserved. For more information on the project, please see <https://www.ehds2pilot.eu/>

2. TOC

1. Document information	2
1.1. Authors	2
1.2. Document history	3
2. TOC	4
3. Executive summary	6
4. Introduction	7
5. Context	8
5.1. HD@EU Pilot project and WP8	8
5.2. Use cases	8
5.2.1. Institutions and roles per use case	10
6. Privacy Enhancing Technologies (PETs)	12
6.1. Description of the available data per use case	13
6.2. Potential threats	16
6.2.1. Data leakage	17
6.2.2. Re-identification of individuals	17
6.2.3. Identification by statistical inference	17
6.3. Prevention measures	17
6.3.1. Anonymisation and Pseudonymisation	17
• K-anonymisation	18
6.3.2. Encryption	19
6.3.3. Differential Privacy	19
6.3.4. Data Erasure	19
6.3.5. Synthetic data	19
6.4. Mitigation measures	20
6.4.1. Auditing logs	20
6.4.2. Regular security audits	20
6.4.3. Incident response plan	20
6.5. PETs applied by the use cases	22
7. Data security	24
7.1. Data processing per use case	24
7.2. Potential threats	25
7.2.1. Data loss	25
7.2.2. STRIDE model	25
7.3. Prevention measures	26
7.3.1. Service/institution capacities	26
7.3.2. Encryption at rest	26
7.3.3. Limitation of storage	26

7.3.4. Regular backups	26
7.3.5. Access Control	26
7.3.6. Secure data communication	27
7.4. Mitigation measures	27
7.5. Data security measures per use case	28
8. Federated approach	31
8.1. Federated querying	31
8.1.1. Architecture description	31
8.1.2. Data security implications	33
8.1.3. Data protection implications	33
8.1.4. ECDC Use Case implementing Federated Querying	34
8.2. Federated analysis	35
8.2.1. Potential threats	37
8.2.1.1. Malicious software	37
8.2.1.2. Data disclosure after the analysis	37
8.2.2. Prevention measures	37
8.2.2.1. Source authentication and authorisation	37
8.2.2.2. Secure software review/audit	37
8.2.2.3. Test in an isolated environment	38
8.2.2.4. PETs	38
8.2.2.5. Data use/sharing agreements	38
8.2.2.6. Training	38
8.2.3. Mitigation measures	38
8.2.3.1. Software monitoring	38
8.2.4. Federated analysis by proxy analyst: How security is handled	39
8.2.5. Data visiting: How security is handled	41
9. Centralised approach	42
10. Compute capabilities	43
10.1. General observations	43
10.2. Compute capabilities needed by the use cases	44
11. Conclusions and recommendations	48
Relevant references	50

3. Executive summary

In this deliverable we present an overview of the Privacy Enhancing Technologies (PETs) and data security threats and measures deemed to be relevant for the future HealthData@EU (HD@EU) based on the use cases (UCs) of this pilot project. In addition, we have collected how the five UCs implemented those.

We also identify the different approaches for data access the use cases took. Briefly, all of them went for a federated analysis approach, where we distinguish “Data visiting” versus “Federated analysis by proxy analyst”. The UC led by the European Centre for Disease Prevention and Control (ECDC) piloted two more scenarios. Apart from the federated analysis decentralised approach, they also implemented a scenario in which data from the different participating countries are centralised for pooled analysis. Furthermore, they tested a technical method to access aggregated statistical results (as envisioned under the EHDS Article 69, “Health data request”). Specifically, they piloted a federated querying mechanism in collaboration with BBMRI.

The HD@EU infrastructure is expected to give access to many different data types and sizes. This, together with the fact that the EHDS envisions the data applicants to provide a description of the tools and computing resources needed when applying for data access, led us to collect those needs from the pilot UCs.

Finally, based on all the information collected in this report, we provide some recommendations to be considered in the preparation phase of the upcoming HD@EU.

4. Introduction

This document collects a general view and, most importantly, a view from the five HD@EU pilot use cases (UCs) perspective on the Privacy Enhancing Technologies (PETs), security measures, federated and centralised approaches, and compute capabilities needed when running their projects.

The data life cycle from TEHDAS¹ is used as the basis for giving such overview, following the strategy of the other two deliverables of this WP, based on data interoperability (D8.1) and quality & provenance (D8.2).

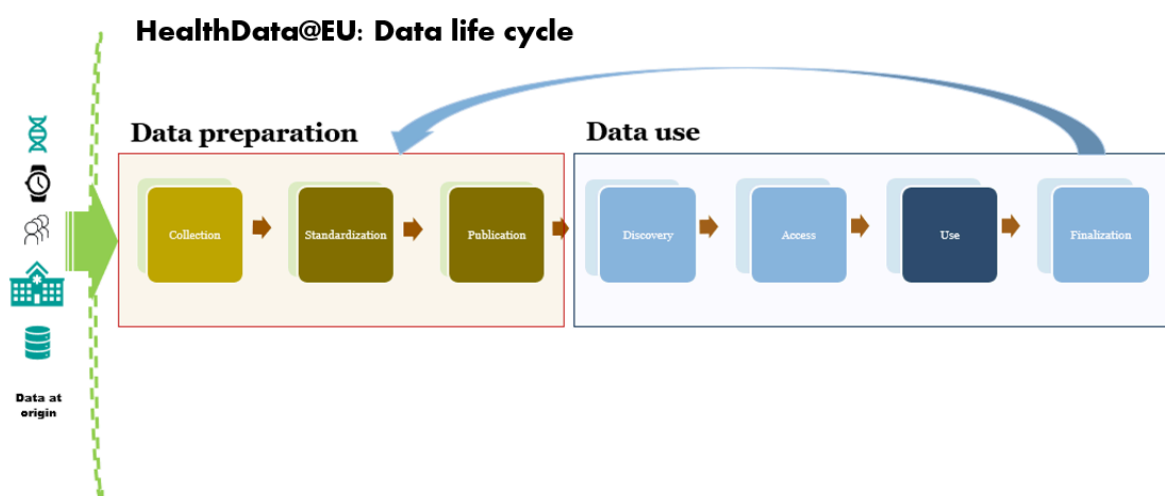


Figure 1. HealthData@EU Data life cycle from TEHDAS²

The chapters on PETs and data security are structured as follows:

- Present the situation in each use case
- Potential threats
- Prevention measures
- Mitigation measures
- Measures taken in each use case

The federated data access chapter is divided into querying and analysis. The former presents the architecture description, security implications and the implementation of the BBMRI solution, whilst the latter follows the structure presented above. In addition, as the use case led by ECDC is piloting a centralised approach, a short section is dedicated to it.

Compute capabilities are also part of this deliverable, giving some general observations and gathering the requirements from each use case, in terms of hardware and software.

Finally some conclusions and recommendations for the upcoming HD@EU are given.

¹ TEHDAS D6.2: [Recommendations to enhance interoperability within HealthData@EU](#)

² TEHDAS D6.2: [Recommendations to enhance interoperability within HealthData@EU](#)

5. Context

5.1. HD@EU Pilot project and WP8

The HealthData@EU Pilot³ is a two-year long project co-financed by the EU4Health programme. It builds a pilot version of the European Health Data Space (EHDS) infrastructure for the secondary use of health data ("HealthData@EU") which will serve research, innovation, policy-making and regulatory purposes, testing the EHDS⁴ Regulation for secondary use. The consortium collaborates closely with the European Commission and its team working on developing the central services for secondary use of health data. The project connects data platforms in a network infrastructure and develops services supporting the user journey for research projects using health data from various European Union (EU) Member States. Priority services include a metadata discovery service and a common health data access request.

This pilot project is also meant to provide guidelines for data standards, data quality and data security. For doing so, this work package (WP8) works as an observer, collecting the information, steps and issues that the use cases are experiencing during this journey. Thus, WP8 is not actively guiding them, but learning from them and only providing feedback upon request. However, it has an active role when piloting the federated querying, as explained later in this document.

5.2. Use cases

Surveillance of antimicrobial resistance use case led by ECDC

The European Centre for Disease Prevention and Control (ECDC) was invited to participate in Work Package (WP) 9 of the HealthData@EU Pilot, focusing on a specific use case to assess the feasibility of using the EHDS secondary use framework to support epidemiological surveillance of antimicrobial resistance (AMR). This UC primarily intends to assess the concordance between national AMR surveillance data (EARS-Net 2020 dataset) previously submitted to The European Surveillance System (TESSy) and data from the same participating countries and year through EHDS. ECDC is working with three participating countries (Belgium, Croatia, and Finland) in its UC.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

This UC involves the DARWIN EU network as well as four countries: Denmark, France, Croatia and Finland. All nodes have managed to access the data. It aims to address 5 research questions of growing complexity: estimate the incidence of venous and arterial thromboembolic events among 1/the general population; 2/patients with COVID-19; 3/patients with SARS-CoV-2 vaccination ; estimate 4/the impact of clinical risk factors and

³ [HealthData@EU Pilot](#)

⁴ https://www.europarl.europa.eu/doceo/document/TA-9-2024-0331_EN.html

prior SARS-CoV-2 vaccination on the incidence of venous and arterial thromboembolic events among patients with COVID-19 and worsening of COVID-19, as well as 5/the incidence rate ratios for such events among patients with COVID-19 during the period when Omicron was the dominant variant and people vaccinated against SARS-CoV-2, compared to background rates as estimated in objectives 1, 2 and 3. Based on the questions and variables defined in the protocol, the research teams aim to conduct the analysis both in the native and OMOP format, and compare the output of both analyses.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

This UC involves six countries: Belgium, Finland, France, Croatia, Denmark and Hungary, of which four have managed to access all or part of the requested data: Belgium, Finland, Croatia and Denmark. It aims to measure the uptake of tests, hospitalisation and vaccination in the general population and in vulnerable subpopulations, in order to get a European overview of the situation and to compare between Member States. Vulnerable subpopulations are defined through the use of socio-economic indicators (income, level of education, migratory background...).

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

This UC involves five countries: Finland, France, Denmark, Hungary, Norway, of which four have managed to access all or part of the requested data (all except Hungary).

This UC aims to address two questions:

- Are health trajectories leading to cardiometabolic diseases (i.e. cardiovascular diseases, type 2 diabetes and stroke) comparable across countries?
- Can we leverage longitudinal disease trajectories to forecast risk for cardiometabolic diseases?

To address these questions, the UC set up two pipelines: one for calculating incidence rate, and one for machine learning. The first requires stringent data harmonisations. The second target is building a prediction model for the 5-years risk of developing selected main cardiometabolic outcomes, based on age, sex, place of residence, diagnoses, operation and medication.

Genomic data linked to health data, with a focus on cancer, led by ELIXIR

This UC involves the Research Infrastructure BBMRI-ERIC, as well as four countries: Belgium, Norway, Denmark and Hungary, of which two have managed to access all or part of the requested data (Belgium and BBMRI). The UC draws on genomic data to interpret complex mutational patterns and medical trajectories of metastatic colorectal cancer patients in the context of associated clinical data. The objective is to confirm known gene signatures and unveil new ones beyond other confounding factors like origin and socio-economics factors as well as confirming new data-driven hypotheses on having different signatures depending on disparate factors like tumour localization or age.

Although not all HDABs and Research/Public Health teams have access to the data, the UC

has adapted the analysis protocols to align with the available data and deliver results in a timely manner within the project's timeframe.

5.2.1. Institutions and roles per use case

The paragraphs below aim to give a description of the institutions participating in each use case, mainly focusing on those countries and Research Infrastructures (RIs) that have access to the data needed to reach the goal of each UC (listed below).

Surveillance of antimicrobial resistance use case led by ECDC

All the institutions included in ECDC's UC had access to the historical 2020 AMR dataset that was previously reported to ECDC's TESSy in 2021.

- Belgium: Sciensano acted both as data holder and Health Data Access Body for the historical AMR dataset and as public Health team of the UC Node.
- Croatia: the University Hospital for Infectious Diseases Dr. Fran Mihaljevic (BFM) acted as data holder and provided the historical dataset to the public Health team in the UC Node, the Croatian Institute of Public Health (HZJZ).
- Finland: the Finnish Institute for Health and Welfare (THL) acted as data holder and provided the historical dataset to the public Health team in the UC Node, Finnish Social and Health Data Permit Authority, Findata.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

- DARWIN EU network: acts both as HDAB and as research team for this Use case.
- Finland: the Finnish Institute for Health and Welfare (THL) acts both as HDAB and as research team for this Use case.
- France: the Health Data Hub acts as HDAB, on behalf of EMA, which formally plays the role of research team; in practice, most of the analyses are done by national experts identified by the HDH but still working on behalf of EMA.
- Croatia: the Croatian Institute of Public Health (CIPH) acts both as HDAB and as research team for this Use case.
- Denmark Health Data Authority (DHDA - Denmark) acts as HDAB and Danish Medicines Agency (DKMA) acts as the research team.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

- Belgium: Sciensano acts as the use case leader & the research team for this use case. Data used for the use case are hosted by the HealthData organisation that is located inside Sciensano but not directly managed by Sciensano. Finally, the official HDAB is the recently created Health Data Agency (HDA) (<https://www.hda.belgium.be/fr>).
- Denmark: the Danish Health Data Authority (DHDA) acts as HDAB while the Central Denmark Region provides the research team.
- Finnish Institute for Health and Welfare (THL - Finland) acts both as HDAB and as research team for this Use case.

- Croatia: the Croatian Institute of Public Health (CIPH) acts both as HDAB and as research team for this Use case.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

- Finland: THL acts as HDAB for this Use case, while the research team (and Use case lead) is from the University of Helsinki.
- France: the HDH acts as HDAB and the University of Bordeaux provides the research team
- Denmark: the Danish Health Data Authority (DHDA) acts as HDAB while the University of Copenhagen provides the research team.
- Norway: since January 1st 2024, the Norwegian Institute of Public Health (NIPH) acts as HDAB and research team. Due to political decisions, the HDAB was relocated from a directorate for the NIPH.

Genomic data linked to health data, with a focus on cancer, led by ELIXIR

- Belgium: Sciensano acts both as HDAB and as the research team.
- European Research Infrastructure: BBMRI acts both as HDAB and as the research team.

6. Privacy Enhancing Technologies (PETs)

Processing personal data may affect the patients' right of preserving their privacy, therefore, in line with the data minimisation principle in the GDPR⁵, measures should be taken to limit the processing of personal data to what is necessary for achieving the intended purpose and to limit the impact of such processing. When dealing with personal data, as we try to preserve the privacy of data subjects by deleting or obfuscating identifying data, we usually reduce the utility of this data, because we are reducing the amount of information contained in the dataset. This situation forces us to strike a balance between privacy and utility.

The principle of "data minimisation" means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specific purpose. They should also retain the data only for as long as is necessary to fulfil that purpose. In other words, data controllers should collect only the personal data they really need, and should keep it only for as long as they need it.

The data minimisation principle is expressed in Article 5(1)(c) of the GDPR and Article 4(1)(c) of Regulation (EU) 2018/1725, which provide that personal data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"⁶. It is notable that processing and storage time may need to be extended, e.g., when processing data for scientific contributions with relevant reasons including repeated analyses.

Privacy Enhancing Technologies (PET) are a continuously improving set of technologies and procedures that help us to increase the privacy of data subjects with a minimum loss of information. The correct application of one or several PETs on a given dataset, depending on the use case, will allow us to optimise the ratio between privacy and utility of the dataset for a specific research question.

In order to set a common ground for reading this document, it is key to have a look at the definitions of anonymous and pseudonymous data. In this deliverable, we are sticking to those definitions given by GDPR:

- Anonymised data (Recital 26): "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."
- Pseudonymised data (Article 4): "means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

Anonymisation and pseudonymisation are the techniques to generate such data and they are further explained below in this document.

It is worth noting that anonymous and pseudonymous data are subject to different legal requirements. While pseudonymous data always remains personal data and its use is

⁵ [General Data Protection Regulation \(GDPR\)](#)

⁶ https://www.edps.europa.eu/data-protection_en

therefore subject to the various requirements of the GDPR, anonymous data is not subject to the GDPR as long as the type of use does not lead to a re-identification risk. In legal terms, it is not always easy to distinguish between personal and anonymous data. The GDPR takes a risk-based approach, so there is no black and white solution.

Even 8 years after the adoption of the GDPR, there are still unresolved legal issues in connection with the term anonymisation. In particular, it is still disputed whether the term is to be understood in absolute or relative terms. The better reasons are likely to speak in favour of the latter, but the European Court of Justice (ECJ) will have the final say, as the question in question is currently pending before it.⁷

The following statements are therefore subject to the proviso that changes may still arise as a result of the case law of the European Court of Justice, in particular with regard to the distinction between anonymous and pseudonymous.

6.1. Description of the available data per use case

Knowledge about the data is key to decide which PETs and security measures should be applied. Thus, here a brief description of the datasets used in each UC is provided. Of note, all data is included, those that are available for running the UC and those that were not accessible when the reason for the inaccessibility is related to security measures.

Surveillance of antimicrobial resistance use case led by ECDC

Participating countries and institutions (Belgium, Croatia, and Finland), from the original data holders' point of view, were able to locate the complete historical 2020 AMR dataset that was previously reported to ECDC TESSy in 2021, and provided country-specific node access to it. The dataset follows the Antimicrobial resistance (AMR) reporting protocol 2021, according to the European Antimicrobial Resistance Surveillance Network (EARS-Net) surveillance data for 2020⁸.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

- Finnish Institute for Health and Welfare (THL - Finland): The main data sources are nation-wide Care Registers collected and maintained by THL. The data are analysed within THL. Socioeconomic and medication data that are delivered from other institutes. THL has strict rules and processes for data protection. Protocols applying data minimisation from GDPR include direct access to any direct identifiers for only those who need them and securing the pseudonymised id-code person-id mapping.
- Croatian Institute of Public Health (CIPH - Croatia): The main data sources are

7

<https://curia.europa.eu/juris/document/document.jsf?text=&docid=295078&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=28925561>; <https://curia.europa.eu/juris/liste.jsf?language=en&td=ALL&num=T-557/20>; https://edpl.lexxion.eu/data/article/16563/pdf/edpl_2020_04-007.pdf; <https://www.lexology.com/library/detail.aspx?g=87099344-3ffd-4f9f-8376-e51f32683726>;

⁸ [Antimicrobial resistance \(AMR\) reporting protocol 2021](#)

nation-wide health care Registers collected and maintained by CIPH. The data is analyzed within CIPH. Socioeconomic data is held by other data holders outside CIMP and include the Pension Insurance Fund, the National Bureau of Statistics and the Tax Administration. Those data holders are not willing or allowed by law to share this kind of information at the moment of requesting data during this project. CIPH puts efforts in data minimisation and anonymisation to ensure privacy.

- Denmark Health Data Authority (DHDA - Denmark): The main data sources are nationwide health care Registers collected and maintained by DHDA. The data are analysed within DKMA. DHDA has strict rules and processes for data protection. Protocols applying data minimisation from GDPR include direct access to any direct identifiers for only those who need them and securing the pseudonymised id-code person-id mapping. All data has been converted to the OMOP CDM and DKMA has onboarded Darwin EU.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

The data required has been listed in a common data model dedicated to this use case. All variables needed are: Sex & age of the person, date of death, country of origin (First Nationality), country of residence, area of residence (NUTS 3), education level, income category, migration background, household type, COVID-19 hospitalisation event (determined by a positive test within 14 days before the date of hospitalisation), total number of all & positive COVID-19 PCR tests during the study period, brand and date of the first 3 doses of the COVID-19 vaccine, number of vaccine doses received during the study period, COVID-19 vaccination status & date of being considered fully vaccinated.

- Sciensano (Belgium, research team): The research team managed to have access to both clinical & socio-economic data that comes from the LINK-VACC project (Linking of registers for COVID-19 vaccine surveillance). This database links selected variables from existing registries for COVID-19 vaccine surveillance, in order to ensure the monitoring of COVID-19 vaccines in the phase following their marketing authorization (post-authorization surveillance). This includes the measurement of uptake and coverage of the vaccination, the estimation of vaccine effectiveness, and continuous monitoring of the vaccine's safety. For these purposes, existing pseudonymised data on COVID-19 laboratory test results, hospitalised COVID-19 patients, COVID-19 vaccinations, underlying health problems, socio-demographic and -economic factors, and healthcare worker status are linked.
- Central Denmark Region (research team) and Danish Health Data Authority (HDAB): The research team managed to have access to both types of sensitive data.
- Finnish Institute for Health and Welfare (THL - Finland): Data is gathered and analysed within THL. Socioeconomic and medication data that are to be delivered from other institutes are not available for the researcher group yet. THL has strict rules and processes for data protection. Minimisation includes direct access to any direct identifiers for only those who need them and securing the pseudonymised id-code person-id mapping. Some of the data from Statistics Finland cannot be given at person level outside their secure environment, Fiona(R). For analyses involving those variables, the analysis data gathered from elsewhere is sent to Fiona and after linkage it is analysed there, but remotely by the THL researcher team.

- Croatian Institute of Public Health (CIPH): The research team has managed to have access to clinical data (testing, vaccination & hospitalisation) but not to socio-economic data which is managed by the Tax administration, National Bureau of Statistics and Pension Insurance Fund. Multiple requests to receive that data have been filed, but unsuccessfully due to national law which does not allow the exchange of such datasets due to privacy concerns. The research team has created a dataset based on the common data model in order to generate a final report with both types of sensitive data in order to run scripts successfully, but the data should not be used as evidence due to poor reliability and granulation.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

Data necessary in the use case is population level socioeconomic data, administrative data, and medical data. As clinical data in terms of measures and results are not part of all countries' registries, such data are not included in the use case.

Taking Norway as an example, the Norwegian Institute of Public Health has been granted access to relevant and necessary data from:

- The Norwegian population database (Norwegian population live by 01.01.2010)
- The Norwegian Patient Registry (every contact with hospitals 2008-2022)
- The Norwegian Prescription Database (all medication delivered from prescriptions 2004-2022)
- The Norwegian Cardiovascular Registry (cardiovascular events 2012-2022)
- Statistics Norway (education, work, habitation 2010-2022)

The Norwegian, Danish and Finnish teams analysed their respective nationwide population data. The French team analysed an excerpt of ca. 17 % of the national population data, i.e., 12 million individuals rather than 68 million.

Genomic data linked to health data, with a focus on cancer, led by ELIXIR

Data necessary in the use case is data on individual patient levels of which there exist genomic data. The population is identified on the basis of eligible patients where the relevant data could be detected.

Two data types were needed for this use case: clinical and genomics data.

In parallel to data access applications, a data model for clinical data was built based on the 1+MG and the BBMRI CRC cohort ones, choosing those variables that were common to both data models or those that were deemed necessary for running the use case, even if they were not present in both models.

When it comes to genomics data, ideally, whole genome sequences (WGS) in VCF format (Variant Call Format) were needed.

Data from both types have to be linked at individual level, in a way that the clinical variables and the genomic sequences are available for each patient included in the use case.

The data available for this use case at the moment are the following:

- 5000 samples sequenced using Gene panel ([NGS convention gene list](#)) in VCF.

- 128 samples sequenced with the FMI NGS gene panel, which are not in VCF.
- 59 samples sequenced using the NGS panel of 523 genes (Illumina's TSO500 NGS panel) in VCF format.
- 425 WGS samples in VCF format.

Table 1. Summary of data description for each use case.

UC lead	Countries	Data
ECDC	Belgium Croatia Finland	Complete historical 2020 AMR dataset that was previously reported to ECDC TESSy in 2021.
EMA	Darwin EU Finland France Croatia Denmark	Data from nation-wide Care Registers.
Sciensano	Belgium Denmark Finland Croatia	All variables needed and included in the data model are: Sex & age of the person, date of death, country of origin (First Nationality), country of residence, area of residence (NUTS 3), education level, income category, migration background, household type, COVID-19 hospitalisation event (determined by a positive test within 14 days before the date of hospitalisation), total number of all & positive COVID-19 PCR tests during the study period, brand and date of the first 3 doses of the COVID-19 vaccine, number of vaccine doses received during the study period, COVID-19 vaccination status & date of being considered fully vaccinated.
HDH/University of Helsinki	Norway Denmark Finland	Population level socioeconomic data, administrative data, and medical data.
Elixir	Belgium BBMRI	Clinical and genomics data on individual patient levels of which there exist genomic data.

6.2. Potential threats

Regarding the preservation of patients' privacy in the context of data analysis within the UCs and taking into account all phases of the process, the main risk is the possibility for anyone to access personal information to which he or she should not have access, with an absolute or high probability of attribution of that data to a natural person. A natural person may be identified from direct personal identifiers, or from a combination of indirect identifiers.

Identification via non-direct identifiers may also occur from aggregated statistical results data. This access to personal data can be done by the researcher or analyst who has been granted access to the pseudonymised data, or may be the result of a data leakage, or from publishing or sharing aggregated statistical data not properly reviewed for privacy concerns. The identification of potential threats would suggest the selection and use of PETs, alone or combined, to prevent or avoid the impact in the case of threat materialisation.

6.2.1. Data leakage

In the case of personal non-anonymised data, the leakage of all or part of the data can expose to unauthorised eyes the personal information contained in the dataset. Avoiding such leakage requires security measures, but complementing these measures with PETs can mitigate the impact of the leakage or even eliminate the possibility of unauthorised access to personal data. Data exposition or unauthorised data copy can be done in any of the phases of data processing: data collection, data storage, data transfer and data computation.

6.2.2. Re-identification of individuals

Re-identification is any process by which information is attributed to data in order to identify the individual to whom the data is related. This threat is especially relevant when data are pseudonymised, and the pseudonymisation key can be accessed by analysts or is subject to leak.

6.2.3. Identification by statistical inference

Sometimes, even if pseudonymisation keys are not accessible but it is possible to access some public personal information about natural persons, for example from social networks, it may be possible to infer the identity of some people through statistical correlation between public events and pseudonymised health data.

6.3. Prevention measures

Ideally, the threats above should not materialise. In order to reach that goal, prevention measures should be set up.

6.3.1. Anonymisation and Pseudonymisation

Anonymisation⁹ and pseudonymisation¹⁰ are data protection techniques used to safeguard personal information. Anonymisation involves transforming data in such a way that individuals cannot be identified, either directly or indirectly, making re-identification virtually impossible (as a residual risk of re-identification is present, see also legal remarks under 5.). In evaluating whether data is sufficiently processed in order to be called anonymous, we also have to take into account possible techniques and available information that, if added

⁹ [AEPD-EDPS joint paper on 10 misunderstandings related to anonymisation](#)

¹⁰ [Data Protection Commission. \(2019\). Guidance on Anonymisation and Pseudonymisation.](#)

to the statistics, could identify one individual or an identifiable group of people in a given context.

On the other hand, pseudonymisation replaces private identifiers with new identifiers or pseudonyms, allowing data to be linked to the same individual across different data sets without revealing their actual identity. However, very often each institution uses its own pseudonym, usually different ones for each project, due to data security reasons. That's why linkage is not always possible. Unlike anonymisation, pseudonymisation is designed to be reversible as the pseudonyms can be linked back to the original data with additional information stored separately. Thus, pseudonymisation allows privacy preserving record linkage techniques.

Both, pseudo- and anonymisation, are crucial for compliance with data protection regulations such as GDPR, **balancing the need for data utility and privacy**. For specific features such as the reporting of incidental findings or going back to the patients for other reasons (consents, follow up with the study and so on), pseudonymisation is the only approach.

As pseudonymous data remain personal data, a sound legal basis for processing these data must be demonstrated at any time and step of the processing. A legal authorisation for use may consist in the consent of the data subject (research participant), but also in special regulations under national or EU law. In view of the unresolved legal issues in connection with the term anonymisation under the GDPR, it is not surprising that the interpretation of this term differs among the Member States.

Of note, aggregated data is not a legal term and it does not imply that the data are anonymous, it can still be personal data and anonymisation measures may be needed (e.g., k-anonymisation, see below). In Norway, anonymous data have to be aggregated in groups of individuals to make them accessible, and the risk of cross-referencing the table(s) to other available data has to be deemed acceptable (i.e. no risk using the GDPR's 'means reasonably likely to be used' test). As compared to possibilities available when analysing the data in its original format, anonymised data surely offers less, but may prove to be useful.

A uniform interpretation and application of the GDPR is needed. Clear guidance from the European Data Protection Board is needed, and ECJ case law may provide further clarification, to ensure uniform application as soon as possible. However, as long as the legal uncertainties persist, it is advisable to apply the strictest standard to be on the safe side.

- **K-anonymisation**

K-anonymisation is a technique that ensures each individual in a dataset is indistinguishable from at least $k-1$ others based on certain identifying attributes. This is achieved by generalising or suppressing parts of the data until each record is identical to at least $k-1$ other records in their pseudo-identifiers, so there is a probability of $1/k$ of a given record to correspond to a natural person whose additional external information could lead to his or her identification in the dataset. As in other anonymisation techniques, the fact of it might involve suppressing part of the data, can lead to an information loss, especially when studying in low incidence conditions. There exist some variants in the k-anonymisation algorithms¹¹ that allow a minimum loss of information by selecting the dimensions to modify

¹¹ Gionis, A., Tassa, T. (2007). k-Anonymization with Minimal Loss of Information. In: Arge, L., Hoffmann,

or generalise, depending on the specific research question.

6.3.2. Encryption

Encryption is a technique used to protect data from being stolen, changed or compromised. The technique works by scrambling data into codes that can only be unlocked/reset with a unique digital key. Encryption could be used during data storage, transfer, and querying.

Some algorithms use a public/private asymmetric infrastructure. In these cases, a different key is used for encryption and decryption. This system is adequate for transferring data, as it not only guarantees the privacy and not disclosure of data, but also integrity, authenticity and non-repudiation.

Nevertheless, the advances in computing and algorithmic, especially in the field of quantum computing, can compromise the security of encrypted data in some time, and the information that we consider secure and protected, may be easily exposed in a few years.

6.3.3. Differential Privacy

Differential privacy, intuitively, captures the increased risk to one's privacy incurred by participating in a database¹². It's a rigorous mathematical framework that establishes the risk of inferring the presence of a natural person's data in a given dataset, even if it is anonymised, and defines a way of avoiding this inference by adding a calculated amount of noise to the dataset, without reducing its utility.

6.3.4. Data Erasure

In light of Secure Processing Environments (SPEs) described in the EHDS Regulation (Article 50) and GDPR's principle of storage limitation (Article 5(1)(e)) it is necessary to have technologies for data deletion after the data is analysed. Data erasure is a software-based process of securely overwriting digitally stored information with random binary data according to a specified standard, then verifying and certifying that the erasure has been successful¹³. Data erasure might be employed as well when decommissioning old hardware or when data has reached the end of its retention period as per data protection regulations¹⁴.

6.3.5. Synthetic data

Synthetic data¹⁵ is data that is developed on the basis of existing data and modelled to reproduce the characteristics and structure of those data. Synthetic and real world data should deliver very similar results when undergoing the same statistical analysis. Some HDABs involved in this pilot project are concerned because users (outside of this project) are requesting real data for creating synthetic data for further use. Discussions are ongoing

M., Welzl, E. (eds) Algorithms – ESA 2007. ESA 2007. Lecture Notes in Computer Science, vol 4698. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75520-3_40

¹² https://link.springer.com/chapter/10.1007/11787006_1

¹³ <https://www.blancoco.com/resources/article-what-is-data-erasure/>

¹⁴ [Guide on good data protection practice in research \(European University Institute\)](#)

¹⁵ [European Data Protection Supervisor \(EDPS\): Synthetic data](#)

in those HDABs (particularly in Norway), as the further use of synthetic data needs to be considered when deciding on granting access to the data. In addition, what measures and restrictions should be applied to the synthetic data generated as a result is another point of discussion. This should be considered case by case and use by use when handling synthetic data generated from real world data, to make sure that there is no risk of re-identification, considering the means reasonably likely to be used.

6.4. Mitigation measures

In the event of a materialised threat in a data infrastructure, mitigation measures are needed to reduce its impact.

6.4.1. Auditing logs

Regular auditing logs are important for monitoring and reviewing activities within the system or database. It is tightly connected with access policies and management which may differ at each institution. By keeping track of who accesses the system and what changes they make, organisations can identify potential security incidents early. Regular log audits help in understanding the patterns of normal behaviour and thus, can quickly flag any deviation that might indicate a security threat.

6.4.2. Regular security audits

Regularly reviewing systems for vulnerabilities. This proactive approach can help in identifying potential threats before they can be exploited. Indeed, Data Protection Impact Assessments are a requirement under the GDPR in certain situations (Article 35).

6.4.3. Incident response plan

An incident response plan is a predefined set of instructions and procedures to detect, respond to, and recover from information security incidents. This plan outlines the roles and responsibilities of the incident response team, steps for addressing the breach, and strategies for communicating with stakeholders. The goal is to minimise damage, recover any compromised data, and prevent future incidents.

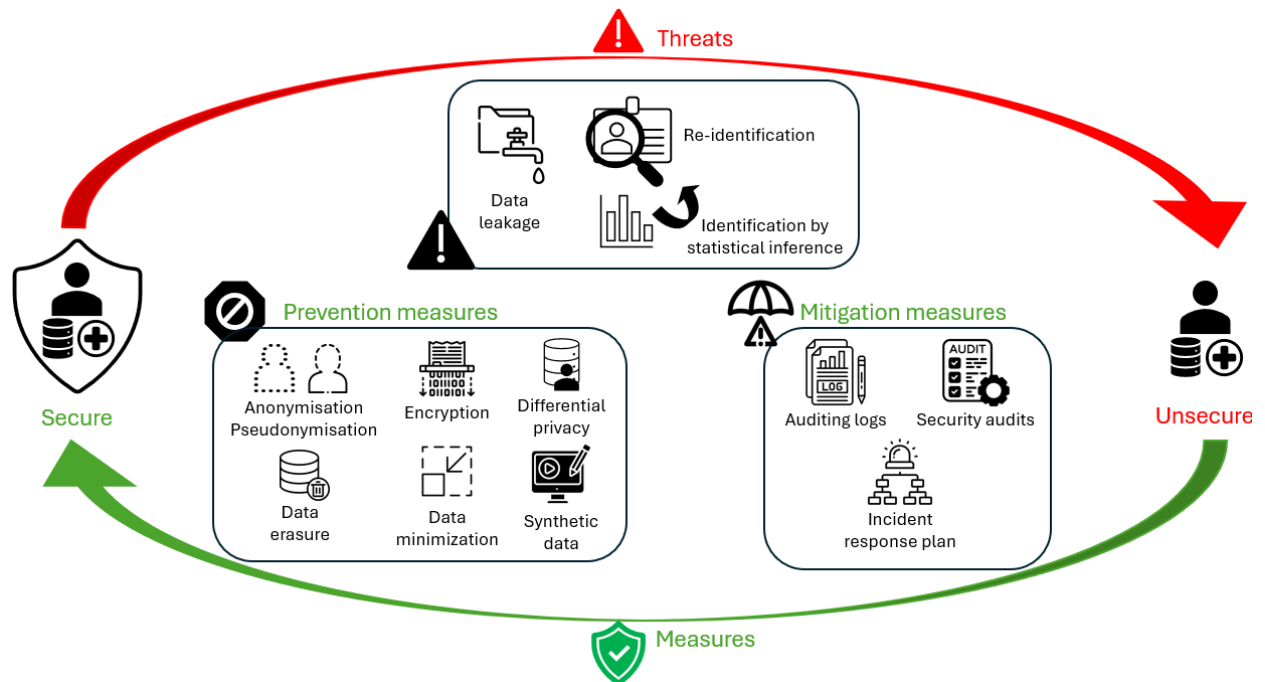


Figure 2. Graphical representation of the PETs explained above.

Looking at Figure 1, most of the above threats, prevention and mitigation measures are relevant throughout the entire data life cycle:

- Collection:
 - **Data leakage** is especially sensitive here, as using the pseudonymisation keys may sometimes be needed for performing some preparation steps such as linkage. Because of the same reason, **re-identification of individuals** is also a main threat during this phase.
 - **Anonymisation and Pseudonymisation** are relevant since the very beginning of the cycle until the end, as depending on the level of access and the dissemination of such data Anonymisation and Pseudonymisation strategies must be followed, as it is reflected in the use cases (see below).
 - **Encryption**: collecting data could mean moving them from one location to another. Thus, encryption when transferring data plays a major role.
 - When building a dataset for a project, **differential privacy** must be considered.
 - Only those data relevant for a given project must be collected, according to the **data minimisation** principle.
- Publication and Discovery:
 - **Identification by statistical inference** is related to these phases belonging to Data preparation and Data use, as metadata and/or aggregated data are published for discoverability reasons.
- Access and Use:
 - Similar to the Collection step, data Access and Use may happen together with data move and, hence, **encryption** should be applied as a prevention measure.
 - In order to accelerate the development of a project or for testing purposes, **synthetic data** could be used by the participant in a given project.

- Finalisation:
 - **Data erasure** must be applied some time (described in EHDS Regulation) after the project is over.

Regarding mitigation measures, they should be present throughout the entire cycle, as they are relevant and might be needed in all phases in case of an incident.

6.5. PETs applied by the use cases

Surveillance of antimicrobial resistance use case led by ECDC

The historical 2020 AMR dataset in ECDC's UC is pseudonymised by default. The data content in this dataset is not encrypted. Principles of data erasure, including secure processing environment deletion, at each node followed country-level legal and technical requirements.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

For this use case, the data minimisation principle is involved and THL utilises pseudonymised data, as the direct identifiers are not distributed to the researcher teams. For this use case DHDA and DKMA use pseudonymised data and the data minimisation principle is involved in line with that applied at THL.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

For this use case, an analytical pipeline consisting of multiple scripts is distributed across all nodes. The initial segment of this pipeline is designed to run within a secure processing environment on individual data to generate aggregated results. To prevent the identification of individuals in small groups, a threshold ($n < 10$) is applied to the aggregated data, ensuring anonymity and protecting individual privacy. The output is a file containing only aggregated data that can be transferred outside the secure processing environment.

When applying for data hosted by HealthData in Belgium, it is necessary to specify the variables needed together with a justification of this need. Health data are pseudonymised by default based on the unique personal ID number. This variable can be used to link several tables within the LINK-VACC project. Exporting aggregated data out of the secure processing environment requires it to be accepted by the HealthData organisation through a structured process.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

Norwegian data from the eligible data sources are pseudonymised based on the unique personal ID number (Personal ID). Statistics Norway gave their decision, drew the defined

national population, and established and kept hold of the linkage key consisting the Personal ID and a project specific pseudonym ID. Statistics Norway then shared the linkage key with the other identified sources via encrypted file transfer. Considerable data minimisation took place during case management at the Norwegian HDAB, forcing the joint research team to discuss necessities and level of detail in respect to the purpose. The decision included pseudonymisation of timestamps, demanding Statistics Norway to modify their data extract as a consequence. Following the final decision, each respective source extracted approved data and transferred the pseudonymised data in encrypted form to the project's dedicated SPE. The project linked the data using the pseudonym ID for each registered person.

Genomic data linked to health data, with a focus on cancer, led by ELIXIR






Applying for access to the BBMRI CRC-Cohort, requires to specify the variables needed for running the use case together with a justification of this need. Clinical data are pseudonymised, using specific pseudonyms only for this use case. Regarding genomics data, only somatic variants are made available.

However, when talking about pseudonymisation of genomic data, Recital 92 of the EHDS Regulation must be considered, as it might be considered as particularly sensitive data. There are several discussions around the anonymity of genomics data in a similar way that happens with Whole Slide Imaging data: *"according to the Art. 29 Working Party (predecessor of European Data Protection Board under the GDPR) even organisational measures can influence the status of anonymity (Art. 29 WP opinion 136, concept of personal data, p. 17), since identifiability depends on the background knowledge of potential attackers: the same data might be anonymous in one setting and personal data in another. Examples of such organizational measures include: access control and contractual obligation to not re-identify research participants, to not share the data with third parties, and to make them internally accessible only under confidentiality obligations, provided that the contractual party is reliable and able to fulfill those obligations¹⁶".*

The analysis carried out in the use case is based on a federated approach, ensuring that data remains at each individual node, thereby mitigating potential risks associated with data transfer. Following the analysis, only anonymised and aggregated data are shared, protecting individual privacy and maintaining data integrity while still allowing for collaborative outcomes.

¹⁶ Holub, P., Müller, H., Bíl, T. et al. Privacy risks of whole-slide image sharing in digital pathology. Nat Commun 14, 2577 (2023). <https://doi.org/10.1038/s41467-023-37991-y>

Table 2. Summary of the PETs applied by each use case.

UC lead	PETs applied	Graphical representation
ECDC	Pseudonymisation Data erasure	
EMA	Pseudonymisation Data minimisation	
Sciensano	Pseudonymisation Results: Aggregated data threshold (n<10) Data minimisation	
HDH/University of Helsinki	Pseudonymisation Encryption Data minimisation	
ELIXIR	Pseudonymisation Data minimisation	

7. Data security

Data integrity and availability on storage are the focuses of data security. In a similar way that when it comes to privacy, there are potential threats to consider, as well as prevention and mitigation measures.

7.1. Data processing per use case

The HDABs and/or research teams that are storing the data used by each use case are listed below.

Surveillance of antimicrobial resistance use case led by ECDC

All three institutions included in ECDC's UC and representing the UC nodes held a copy of the historical 2020 AMR dataset that was previously reported to ECDC TESSy in 2021, for use in their corresponding SPE, for the duration of the UC.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

Each participating research team/HDAB accesses the data in their respective secure processing environment. Only carefully reviewed aggregated statistical data are taken outside the SPE.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

Each participating research team/HDAB accesses the data in their respective secure processing environment. Only carefully reviewed aggregated statistical data are taken outside the SPE.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

Each participating national research team accesses the data in their respective SPE.

Genomic data linked to health data, with a focus on cancer, led by ELIXIR

BBMRI is holding the clinical and genomics data that are going to be analysed within this use case. The RI provides a SPE to which the use case leads would access and perform the analyses needed to carry out the project.

Other nodes (BE and DK) would process their own data in their secure environment and supercomputer premises (GenomeDK), respectively. The Norwegian data would be stored at the University of Oslo's Service for Sensitive Data (TSD).

7.2. Potential threats

7.2.1. Data loss

Risks include accidental or intentional deletion, inability to access or corruption of data. This loss can be caused by technical issues (storage fail, disk overflow...) or human actions (hacking and data hijacking).

7.2.2. STRIDE model

STRIDE¹⁷ is a methodology for identifying potential security threats in software applications or systems. Some threats from this model apply here:

- Spoofing: Pose as something or somebody else.
- Tampering: Malicious modification of data or code, e.g., by man-in-the middle attack possible because of weak message or channel integrity checks.
- Information disclosure (Data leakage): Exposure of data to unauthorised persons, e.g., by man-in-the-middle because of lack of confidentiality for the channel.

¹⁷ M. Howard and S. Lipner. The security development lifecycle: SDL-A process for developing demonstrably more secure software . 2006.

- Elevation of privilege: A user gains unauthorised access to resources.

7.3. Prevention measures

7.3.1. Service/institution capacities

From the perspective of size of the data and function (maintenance, human capacities) ensuring that the infrastructure can handle the management of such data.

7.3.2. Encryption at rest

Protecting stored data from unauthorised access. The way or technology used may differ within each institution or use case. Broadly, there exist encryption algorithms that use a single key for both encryption and decryption. These algorithms are adequate for storing and retrieving information in a secure manner. Encrypting keys should be stored in secure key management systems¹⁸ to prevent unauthorized access and ensure the integrity of the encrypted data.

7.3.3. Limitation of storage

This topic has a two-fold approach:

It may be the physical or technical constraints of the storage system. It involves considerations such as the total volume of data that can be stored, the maintenance requirements of the storage system, and the human resources needed to manage the data. Such capacity limitations can impact the quality of service, as overloading the system can lead to slower data retrieval times and increased risk of system failures.

It can be linked to the principle of data minimisation. Those institutions storing data for secondary use should store only those data that are relevant and necessary for defined purposes according to Article 5(1)(c) and (e) of the GDPR and Article 4(1)(c) of Regulation (EU) 2018/1725.

7.3.4. Regular backups

It is necessary to maintain copies of data in multiple physical locations for recovery, in case one storage location is damaged or destroyed.

7.3.5. Access Control

There should be clear policies on who can access the data. Granularity of the rights should be distributed within given roles, according to a role-based access control (RBAC) model. No person should be a super user having access to everything - for example to mapping tables with personal identifiers and given pseudonyms and at the same time to pseudonymised data. Two aspects are key when controlling access:

- Authentication and authorisation system

¹⁸ For example Vaultwarden <https://github.com/dani-garcia/vaultwarden> or Bitwarden <https://bitwarden.com/download/>

- Authentication checks the identity of the users, typically through methods like passwords or multi-factor authentication (MFA).
- Authorisation determines what an authenticated user is allowed to do, by assigning specific permissions based on their role. Thus it is key to define the role, rights and permissions of each user, based on the tasks they have to perform.
- Configuration management
 - Firewalls to control the traffic in the secure network.
 - VPNs providing a secure connection over the Internet.

7.3.6. Secure data communication

Solutions such as TLS (Transport Level Security) are widely used to ensure secure communications. In particular, TLS is a protocol that encrypts data transfer, ensuring data integrity.

7.4. Mitigation measures

The mitigation measures to be applied in case of an incident are the same as the ones described in the PET section above.

Looking at the different threats and measures from the data life cycle perspective (Figure 1), three phases are especially relevant when it comes to data security:

- Collection
 - **Data loss** is critical at the time the data are collected and stored, that's why **regular backups** are such an important prevention measure.
 - Considering the available **service/institution capacities** is key to make sure the data management can be done in an efficient manner.
 - When data is collected and stored, **encryption at rest** prevents unauthorised access.
 - Considerations regarding **limitation of storage** need to be addressed in this phase.
 - As several institutions might be involved **secure data communication** must be ensured here and also in the two next phases.
- Access and Use:
 - **Access control.**
 - The **STRIDE model** mainly applies to these two phases of the TEHDAS data life cycle.
 - As in the Collection phase, **service/institution capacities** play a key role in the "Use" phase.
 - Risk assessment is a critical process used to evaluate potential threats, estimate their likelihood and impact, and prioritise mitigation measures. It helps organisations to identify vulnerabilities and assess the severity of potential disruptions to operations and assets. By understanding these risks in advance,

organisations can allocate resources effectively, strengthen critical systems, and enhance overall resilience to security threats.

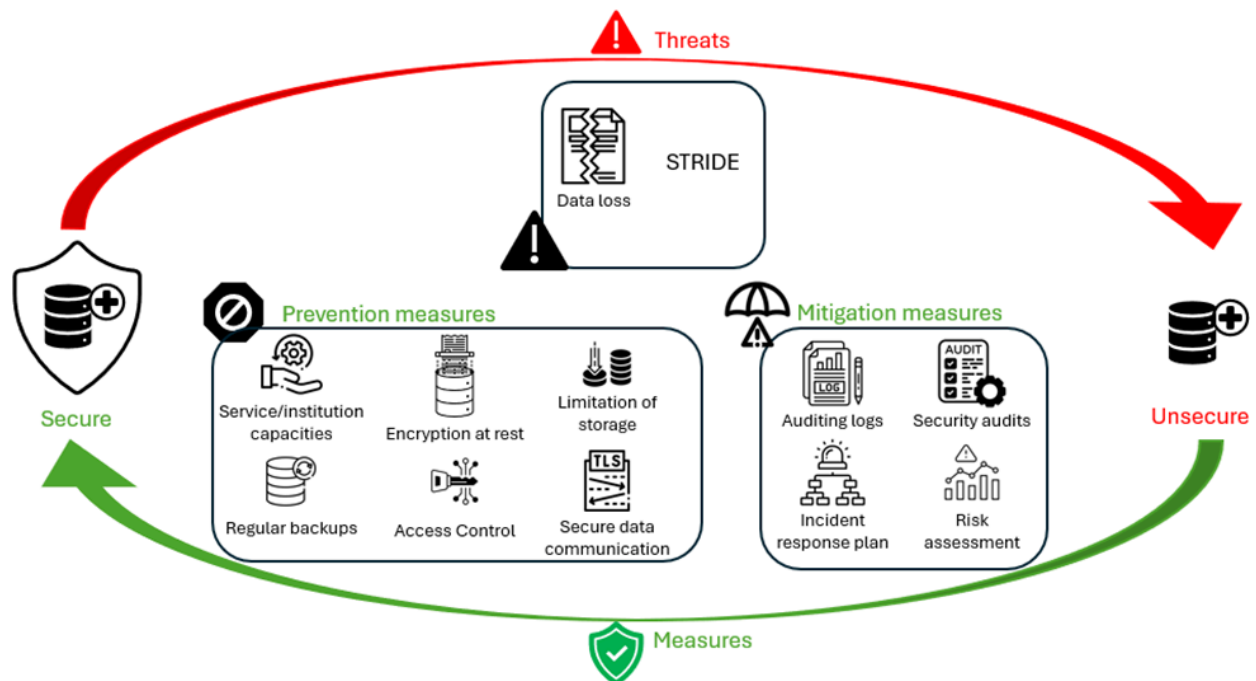


Figure 3. Graphical representation of the data security measures explained above.

7.5. Data security measures per use case

For each use case, a description of the security measures that they have in place for storing the data is provided below.

Surveillance of antimicrobial resistance use case led by ECDC

All three institutions included in ECDC's UC as public health teams have created a dedicated SPE, albeit with different technological approaches. In one participating country (Finland), the SPE was used as a service, according to the national legislation^{19,20} determining SPE's specifications and requirements. Access to this SPE was carried out with multi-factor authentication and restricted to authorised staff. For the other two countries, SPE access control (two-factor authentication for one country and multi-factor authentication for one country) and security rules for use were dependent on the nationally-defined requirements of each institution. Access to these SPE's was also restricted to authorised staff.

¹⁹ [Findata: Regulation on secure operating environments](#)

²⁰ [Findata: Annex 1: Requirements for a Secure Operating Environment](#)

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

At THL Finland, database administrators have indirect access to all of the data in a single database, but pseudonymised and non-pseudonymised data are stored in different databases. In general, access to data is restricted to only those necessary to the task of analysing the data. By convention, analyses of personal data are performed in dedicated servers/machines instead of personal computers. Only anonymous statistics and anonymous results are exported from the analysis environment. The situation at DHDA Denmark is the same, only accessing pseudonymised data.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

In Belgium, access to data is restricted to only authorised projects with a list of authorised persons. Login to the secure processing environment is a two-factor authentication process. Exporting aggregated data out of the secure processing environment requires it to be accepted by the HealthData organisation through a structured process. All data transfers (import/export) are managed by HealthData.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki






The Norwegian Institute of Public Health research team uses the service for sensitive data (TSD) at the University of Oslo²¹ for this use case. Access is restricted to only those necessary to the task of analysing the data. Only anonymous statistics and anonymous results of the scripts run can be downloaded and shared.

Genomic data linked to health data, with a focus on cancer, led by ELIXIR

The BBMRI team set up a SPE to pilot its use. It has restricted access and only authorised individuals have access to it and its data. Regular backups of the data stored in the SPE are performed.

²¹ [Services for sensitive data \(TSD\). University of Oslo.](#)

Table 3. Data security measures applied by each use case.

UC lead	Data security measures	Graphical representation
ECDC	Restricted access	
EMA	Restricted access	
Sciensano	Restricted access	
HDH/University of Helsinki	Restricted access	
ELIXIR	Restricted access Regular backups	

8. Federated approach

As part of the EHDS2 Pilot ("HealthData@EU pilot"), the use cases piloted federated querying and analysis approaches to test their proof-of-concept usability within the EU health data ecosystem.

8.1. Federated querying

The Regulation envisions the existence of a Health data request (Art. 69), from which the user would obtain "a response only in anonymised statistical format". Thus, in order to pilot such queries aiming to get statistical results, a federated querying approach was piloted as part of the activities of WP8 using the BBMRI-ERIC Federated Search platform in collaboration with the UC led by ECDC.

In this project, we are piloting a technical mechanism for executing federated queries. However, it is important to note that we are not following the full Health data request process, as according to the regulation, "The health data applicant may submit a health data request for the purposes referred to in Article 53 with the aim of obtaining a response only in an anonymised statistical format. A health data access body shall not provide a response to a health data request in any other format and the health data user shall have no access to the electronic health data used to provide that response". In this pilot, we are not going through the entire process, we are testing a technical method to demonstrate how federated predefined queries can be performed across multiple data sources, ensuring that the data remains in place and that only aggregated results are provided.

From a Data life cycle perspective, this federated querying approach would take place after the Access phase (Figure 1) and, in a way, forms part of the Data Use phase, as some level of analysis can be performed based on the query results. However, since this method relies on structured queries, it does not replace the need for SPEs; rather, it serves as a complementary approach for accessing data in the form of aggregated query outputs.

8.1.1. Architecture description

The local data is made available to queries via a central search interface. Local and central components together enable findability, accessibility and interoperability of datasets. The architecture does not differ if the access is from the public internet or allowed only for authorised personnel, according to setup and use case requirement. Local component (called "Bridgehead"²²) is run within the local secure network in the institution and the data flow is in the direction from the local to the central component via open APIs. The administrators of the central component are responsible for further processing of the aggregated information received from the local components. Figure 4 depicts the overall architecture where the local components contain Data Storage with harmonised data extracted and transformed from the Local Database. This means that the data in the Local Component is structured and is already ready to use in a given standard. In the case of the ECDC UC in the HealthData@EU Pilot implementation, HL7 FHIR standard was used.

²² <https://github.com/samplify/bridgehead>

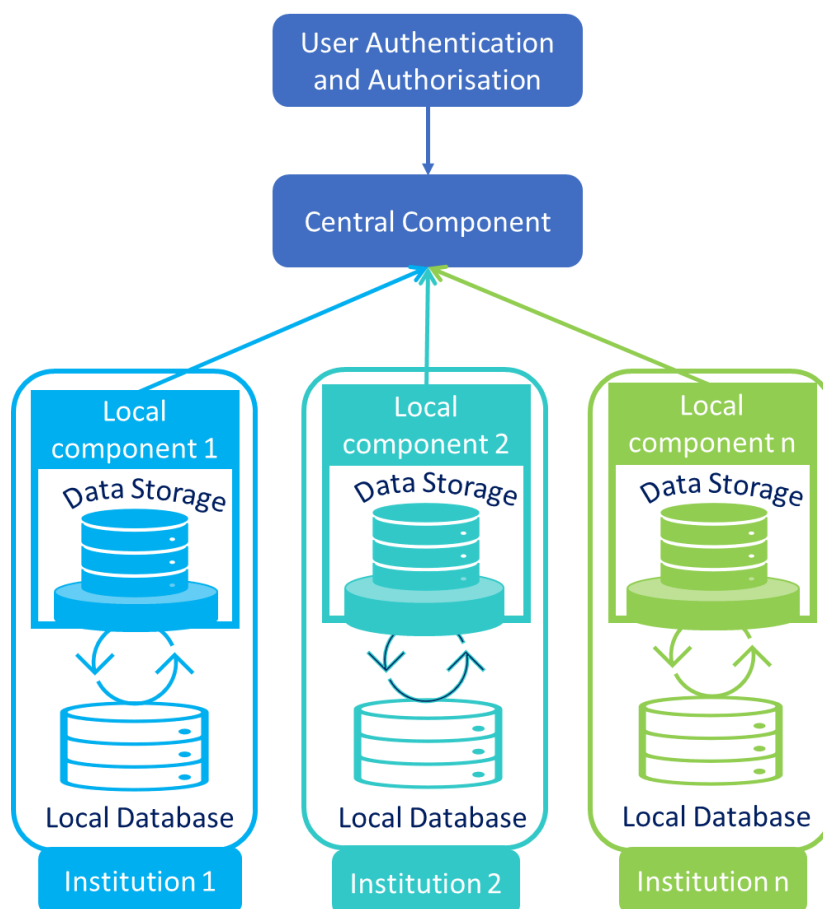


Figure 4. Overview of the architecture of the federated querying approach piloted in this project.

The query created by the user in the central component (Explorer/Locator) is firstly stored in it. The local components in the institutions periodically retrieve new queries from the central components using the Beam infrastructure (see 7.1.2). They then run the query internally against the data in the local data warehouse in the local component and determine which records match the search criteria. Aggregated information about records matching the search criteria is returned to the central component and is visible to the querying user. The queries enabled from the central component (search tree is built according to the data available in the local components) are predefined so the data owner is aware of the information shared with the data user.

Network security implemented

The components are federated and communicate via the public Internet. The confidentiality of the communication is ensured by the following measures (Communication security):

- Communication between individual components of the federated IT infrastructure is always over encrypted connections (HTTPS). Keys and certificates used for this purpose are created in such a way that they meet the currently accepted requirements (e.g. key

length, algorithm), and are issued by certification authorities meeting the Mozilla Root Store Policy²³.

- Firewalls ensure that the servers running the central components are accessible only through the protocols and ports required to communicate with users or other components (generally HTTPS connections). Administrative access is restricted to the people nominated by the operator.
- All ingress communication towards the sites' Bridgeheads is mediated through the Beam component of the Bridgehead. Thanks to Beam's architecture, Bridgeheads can therefore operate behind firewalls and proxy servers without being accessible via a public web address on the Internet.
- There is no direct communication between local components in different institutions.
- Patient and sample counts are obfuscated locally before being sent to the central infrastructure. This is done by introducing some randomness into the count and then rounding it to the nearest multiple of ten.

8.1.2. Data security implications

The federated system with the above described architecture was developed to enable secure search through existing data sources from one central point. Samplify.Beam²⁴ is a distributed task broker designed for efficient communication across stringent network environments. It provides the most commonly used communication patterns across strict network boundaries, end-to-end encryption and signatures, as well as certificate management and validation. It is used for the communication between central and local components, and allows institutions to register with the search infrastructure in a secure manner.

8.1.3. Data protection implications

Technically, the sites where the local components together with the harmonised data warehouses run randomise the aggregate patient/sample counts using methods of "statistical disclosure control". Specifically, a random value drawn from a Laplace distribution with mean 0 and a standard deviation of 5 is added to the patient/sample count. Since this random value can be positive or negative, an upward or downward deviation from the real patient/sample count is possible. This procedure strongly resembles the concept known as "differential privacy"²⁵, although not all of its rigorous mathematical statements can be achieved in practice. Therefore, a further obfuscation step is performed in the form of rounding to the nearest 10th digit. The greatest risk of re-identification in differential privacy is posed by performing multiple queries, therefore a) randomization is saved with respect to the query result and b) user control and restriction measures, described in the data protection concept²⁶, are implemented.

²³ <https://www.mozilla.org/en-US/about/governance/policies/security-group/certs/policy/>

²⁴ <https://github.com/samplify/beam>

²⁵ https://link.springer.com/chapter/10.1007/11787006_1

²⁶

https://www.bbmri.de/fileadmin/user_upload/PDFs/Datenschutzkonzept/2023-Datenschutzkonzept_GBA_v1.2_1.pdf

8.1.4. ECDC Use Case implementing Federated Querying

Surveillance of antimicrobial resistance use case led by ECDC

The current implementation of federated querying for ECDC's Use case (see Figure 5) is based on the BBMRI architecture. It can present several advantages as a complementary approach for collecting and reporting public health surveillance data. Two country nodes (Belgium and Croatia) were able to set up local Bridgeheads (BH), testing a clean data version of the reference AMR surveillance dataset. For the pilot purposes with this UC the FHIR profiles were created²⁷ and a tool for data extraction from source data set, transformation into target standard and load into target storage (ETL) from the TESSy files was programmed.

A Locator instance was installed at ECDC and connected to the aforementioned countries BHs. An application for handling query requests and node responses was also installed, allowing for a predefined set of queries to be performed upon the data in each of the nodes. This application is also responsible for managing proper statistical disclosure control and/or obfuscation of aggregated query results (see above). Hence, the federated querying approach under piloting could improve the timeliness and frequency of indicator-level surveillance data reporting, while operating within decentralised governance frameworks, subject to distinct administrative and legislative data processing requirements. Of note, as explained above, the full "Health data request" process (Art. 69) was not piloted in this project. Instead, the focus was on demonstrating the available technical mechanisms rather than the complete procedural framework required within the EHDS.

²⁷ <https://simplifier.net/HD-EU-ECDC-AMR-UC/>

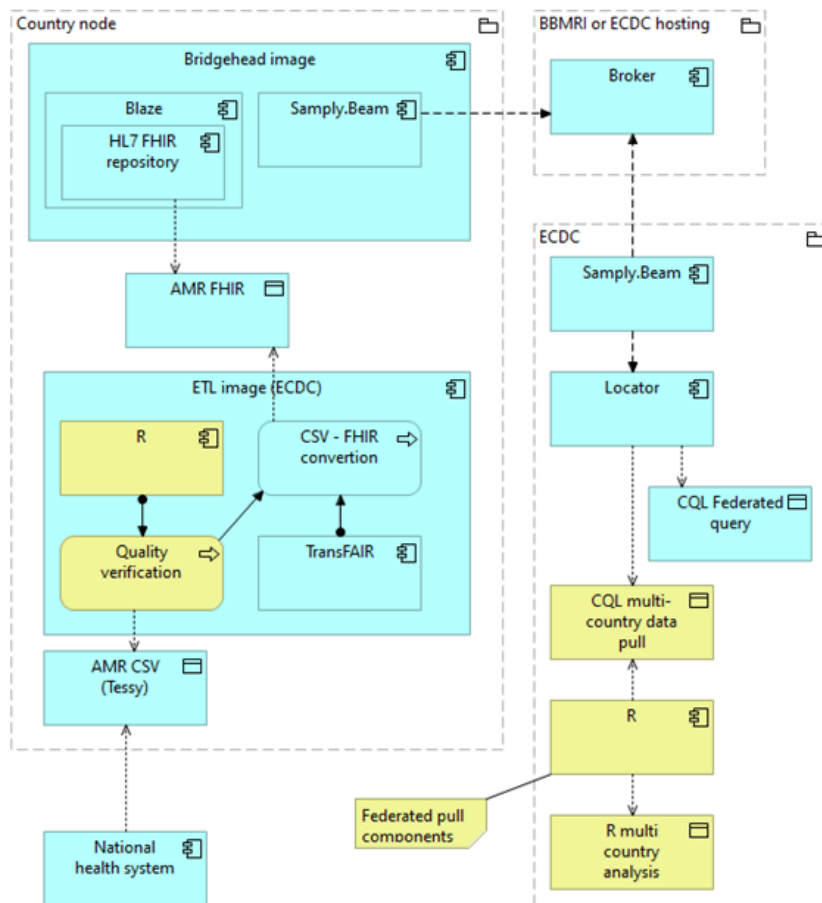


Figure 5. Federated querying architecture in ECDC Use Case using the BBMRI Locator customised software.

8.2. Federated analysis

The five use cases are piloting a federated analysis approach. However, two modalities of this approach can be distinguished. On one hand, in some use cases, the leading team prepares the algorithms and sends them to the research/public health teams. The teams that are storing the data execute the analyses in each server and send the aggregated results back to the leading team. On the other hand, the use case leading team requests access to the SPE where the data are, performing the analyses needed there.

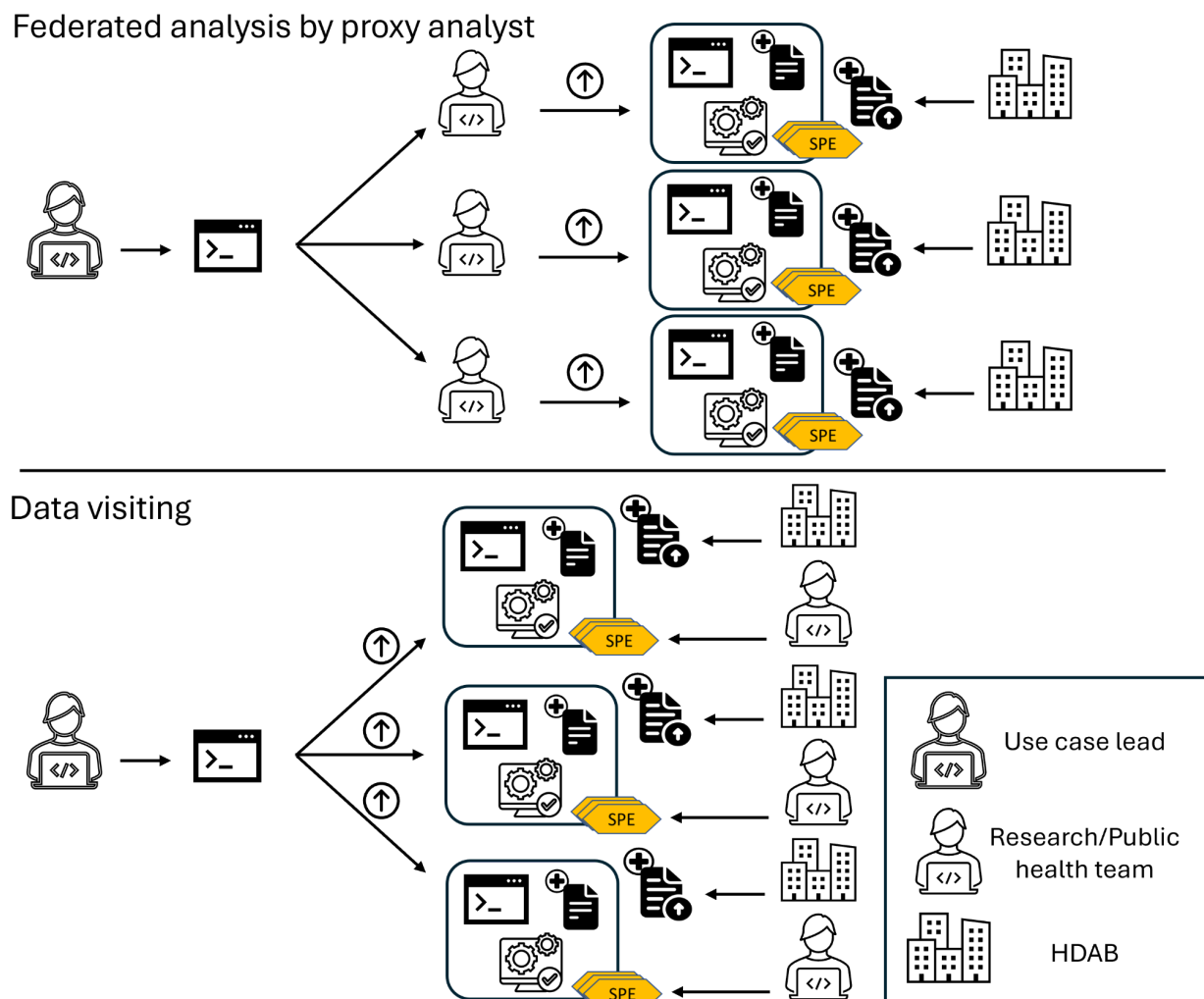


Figure 6. Federated approaches: Federated analysis by proxy analyst versus Data visiting.

Table 4. Classification of the use cases based on the federated approach modality they are piloting.

Use case	Data visiting	Federated analysis by proxy analyst
Surveillance of antimicrobial resistance use case led by ECDC	X	X
Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA		X
Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano		X
Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki		X
Genomic data linked to health data, with a focus on cancer, led by ELIXIR	X	

All the considerations stated above regarding PET and Data security must be taken into account while running an algorithm-to-data kind of analysis. In addition, there are some specific to this approach:

8.2.1. Potential threats

8.2.1.1. Malicious software

The algorithms or code sent for being executed could be designed or can accidentally harm, exploit or misuse the data, network or computer system. Indeed, in projects with federated approach including running code that is created elsewhere this code could include harmful parts including such causing data leakage. This can include leaking the data as it is or exporting results that are not anonymised enough.

8.2.1.2. Data disclosure after the analysis

Release of sensitive data when reporting the results.

8.2.2. Prevention measures

8.2.2.1. Source authentication and authorisation

Verify the software source (or submitter) and make sure that they are trusted and authorised to submit the software for a given analysis.

8.2.2.2. Secure software review/audit

Check the algorithms and code sent before the execution, to make sure it will not compromise the data and system. This can be done manually or programmatically.

8.2.2.3. Test in an isolated environment

Prior to running the software in the entire dataset, they are run in an isolated environment (or sandbox) using a subset of the data or even synthetic data if available.

8.2.2.4. PETs

Apply the PETs explained above to the results in order to prevent data leakage when reporting the results.

8.2.2.5. Data use/sharing agreements

Even if data does not move from one country to the other, they are still either used or shared with external users. Therefore the necessary measures need to be in place and mentioned in the corresponding agreements, together with specific requirements for a given use case. For instance, In Norway, the method "Visiting data" would require standard contract clauses under Article 28 (7) of Regulation (EU) 2016/679 and Article 29 (7) of Regulation (EU) 2018/1725, whilst the method "Algorithm to data" does not.

Possibilities and options for sharing aggregated data should be evaluated prior to choice of SPE, as sharing and discussing results is a crucial part of any federated approach.

8.2.2.6. Training

The users preparing the analysis or analysing the data, have to be aware of the protocols and good practices when managing health data. Thus, they should receive the appropriate training for that purpose and if required, also confiding to the necessary dispensation from privacy.

Data analyses are carefully planned by participating nodes in order to take into account special characteristics of data in each node.

8.2.3. Mitigation measures

8.2.3.1. Software monitoring

Monitor the activity of the algorithm while it is running to quickly detect any suspicious activity.

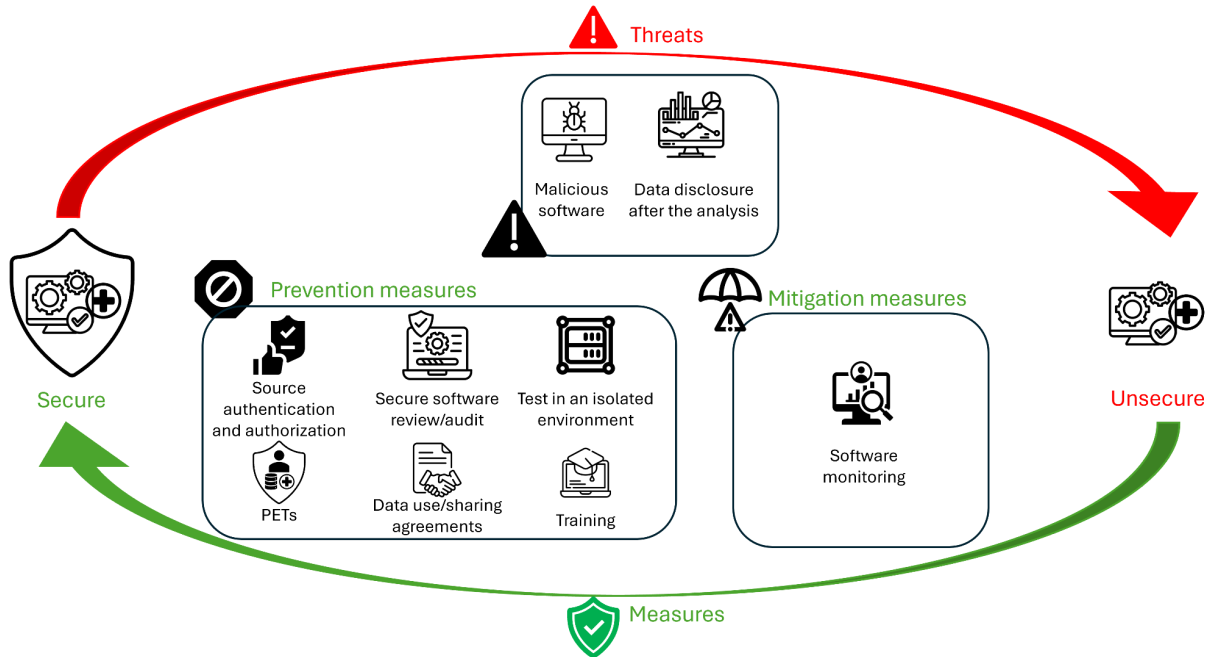


Figure 7. Graphical representation of the security measure for federated analysis explained above.

8.2.4. Federated analysis by proxy analyst: How security is handled

Surveillance of antimicrobial resistance use case led by ECDC

In addressing the security protocols for the ECDC AMR UC, ECDC implemented a robust framework that prioritises data integrity and controlled access across different scenarios. For the federated analysis, security is enhanced through a secured processing environment (SPE) that is strictly accessible to specifically nominated users via personal credentials. This selective accessibility ensures that only authorised personnel can engage with the data, significantly reducing the risk of unauthorised access. Within this setup, the data remains on the SPE, and dedicated R scripts are employed for analytical operations. For all SPEs in the three participating countries, forensics evaluation of analytical source code was carried out before its execution in the SPE, and exfiltration of aggregate indicator-level anonymised tables from the SPE was carried out according to national legal requirements.

Importantly, only anonymised results are permitted to be extracted from the system, safeguarding the confidentiality of the data. For doing so, ECDC provides access credentials to the ECDC SFTP server where the aggregated data are transferred. The aggregated output is transferred to ECDC by the public health team granted access (SFTP is considered the minimum security requirement for the data transfer), and the files once in the SFTP are transferred to a

restricted network storage inside ECDC infrastructure. When the analysis of the aggregated results is done, the data that were stored temporarily on the ECDC SFTP server are removed and the results are made public only after checking that no personal identification is possible.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

A federated approach was chosen and the algorithm (script) was run locally in each country. Before running the algorithm it was reviewed by an analyst. All analysis data was pseudonymised and pseudonymisation keys were stored in a separate database. Only authorised personnel accessed the data and minimisation was utilised. Database logs were in use.

The analysis codes implemented outside THL are reviewed before execution to identify data leakage risks. Aggregated statistical results are reviewed to identify privacy concerns before sharing them outside THL. The same methods are applied in Denmark.

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

A federated approach was chosen as the use case target whole population data and data transfer between countries was not envisageable.

For all use cases, access to sensitive data is controlled, granted only on necessity after approval, and limited to few authorised personnel.

In most SPE's of this use case, it is not allowed to import an external Docker image containing the analysis pipeline (scripts) inside the SPE to generate the aggregated output. Importing external files (scripts) into the SPE needs specific approval, which can take some time depending on the HDAB. For the Belgium node, the research team decided that it would be faster to manually recreate the analysis script inside the SPE infrastructure than asking for importing it.

All SPEs in use in the use case authorise only aggregated results and reports to be exported. For the purpose of this use case, only CSV files containing only non sensitive aggregated data have been exported out of the SPE.

In Finland, Statistics Finland evaluates all outputs from their secure environment aiming at preventing any possibilities of revealing anybody's identity. When exporting results from the THL environment the aim is the same but the method of making the decisions may be less technical and is for example not always bound to any rule of thumb of $k=5$.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

Integrated in the application for accessing data, there are descriptions on where and how data is to be analysed, including security measures.

Data access authorities in both Finland and Denmark, solely accept access to their own national

data on approved processing environments within the respective country. Both countries would however have accepted transfer of other nations' data to be gathered with their own. Consequential, federated analysis of all data was conceived as the favourable solution.

The respective country's data were made available in respective SPE's. In Denmark, France and Finland, the SPE's chosen are provided by the respective HDAB. In Norway, an SPE provided by the University of Oslo was chosen: the TSD. Choosing TSD was a question of convenience and prior experiences, rather than a thorough risk analysis. The TSD relies on encrypted information during transit, changelog, multi-factor authentication, access restriction and logging, and download restrictions. In an ongoing Direct Grant project, the TSD and two more university SPE's are being assessed with the intention of establishing guidelines for functioning as approved SPE²⁸. Also in Norway, the HDAB is granted legal authority to decide where data is to be processed.

Access to data is controlled, granted only on necessity, and limited to few personnel. Access to the Finnish data has been performed under a different project, and there is an ongoing process of securing re-use of these data in the use case.

All SPE's in use in the use case prohibit downloading or exporting data from the SPE. Only aggregated results and reports may be shared. During the process, challenges related to exporting and sharing aggregated results surfaced. In Finland, every result needs to be reviewed by Findata prior to export for sharing. In Denmark, every export requires a fee. In Norway, the chosen SPE allows solely the principal investigator to export results.

8.2.5. Data visiting: How security is handled

Even though the ECDC use case also piloted the data visiting by accessing SPEs to analyse the data, here we focus on the example of the Cancer Genomics use case, as it was not described before.






Genomic data linked to health data, with a focus on cancer, led by ELIXIR

In Belgium, connection to the virtual server relies on a VMware Horizon Client, or subsequently through a VMWare Horizon HTML Access. Any sessions are terminated at ten hours run time. The system reboot weekly updating the unique project as well as Microsoft security measures.

In BBMRI, a virtual machine where remote management of the operating system is done via ssh and the public keypair with only selected users is set up. In addition, password authentication has been turned off and it is continuously monitored. However, it is important to keep in mind that this is under development and further measures may be added when the analysis is performed.

²⁸ [SPUHiN: FAIR Secure Procurement and Use of Health data in Norway](#)

Table 5. Security measures applied by each use case during federated analysis.

UC lead	Data security measures	Graphical representation
ECDC	Authorised personnel Forensics evaluation of analytical source code	
EMA	Algorithm it was reviewed by an analyst Pseudonymisation Restricted access Data minimisation	
Sciensano	Restricted access	
HDH/University of Helsinki	Restricted access	
ELIXIR	Pseudonymisation Data minimisation	

9. Centralised approach

The ECDC use has piloted a centralised approach, where they have aggregated multinational data and performed joint analysis by ECDC on a centralised SPE. In this case, multiple nodes can transfer encrypted national data through secured channels to a centralised SPE (owned by EC/ECDC). For doing so, ECDC creates a public/private key pair to support the data encryption and sends the public key to the Public health teams to use for encrypting the data subject for analysis according to ECDC guidelines for data encryption. Thus, the Public health team encrypts the data using the public key provided by ECDC, connects to ECDC SFTP server and uploads the encrypted data on a commonly agreed date and time. Then, ECDC decrypts and analyses the data directly in the ECDC SPE and when the processing is done the original data are removed. Finally, ECDC validates the output and performs the reporting only after checking that no personal identification is possible. If the validation fails, the process starts again from the data upload step.

A stringent security model is applied, similar to the one explained for the federated approach. Data from various national SPEs are aggregated into the central ECDC SPE, where access is tightly regulated and limited to designated individuals. This controlled access ensures that only qualified personnel can perform analyses on the aggregated data, thereby maintaining a high level of data security and integrity. After the completion of the analysis, the raw pseudonymised data are promptly deleted (as described in the data erasure section above), and only the anonymised analytical results are preserved. This procedure not only minimises the exposure of sensitive data but also aligns with best practices for data protection and compliance with

regulatory standards. These strategic implementations underscore ECDC's commitment to maintaining the highest standards of data security and user confidentiality across all processing environments and analysis scenarios within the ECDC AMR UC framework.

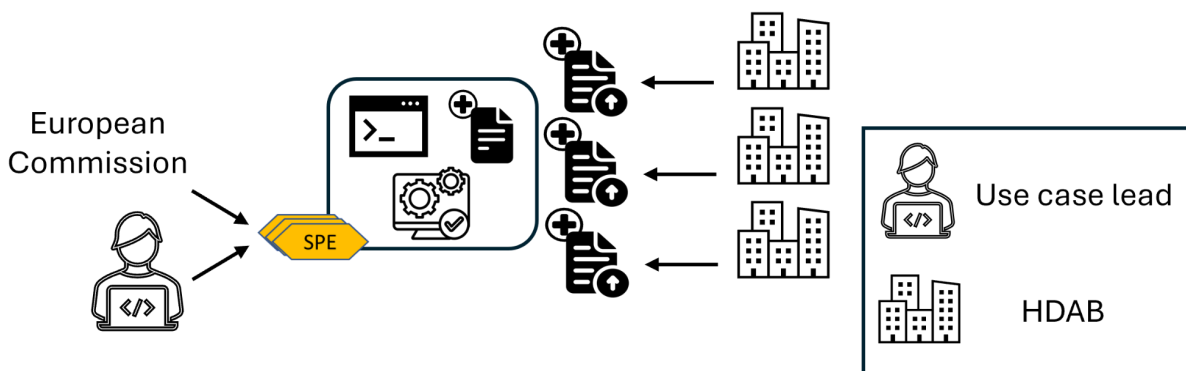


Figure 8. Centralised approach.

10. Compute capabilities

10.1. General observations

The compute capabilities expected to be provisioned are very variable. As reflected in the use cases selected in the present project, the expected EHDS' analysis workloads will be highly heterogeneous, in terms of the data to be used and in the selected techniques to carry on the analyses.

Types and volumes of data may vary from a limited amount of highly structured data required, for example, to perform a survival analysis over a cohort of patients with a rare disease, which will require few gigabytes of storage and memory, and few processors (less than 4); major genome-wide association studies (GWAS) for a high-incidence disease such as diabetes, where the use of high-performance computing (HPC) facilities to speed-up the discoveries is now state-of-art; or, training of AI-based computer vision models to support the radiologist decision making, where there is also a high data storage demand and highly efficient hardware, such as GPU or specialised AI-accelerators (e.g. Google's TPU²⁹ or Cerebras' Wafer Scale Engine³⁰), are widely used to generate AI models in a reasonable time.

These approaches usually refer to single-system solutions (a single server for survival analysis, a HPC system for the GWAS or a GPU-based cluster for the AI imaging). The complexity of the computing capabilities is increased when considering a possible federated approach in the analyses, where multiple systems need to be interconnected by a secure network to coordinate a distributed computation of the results. This is implicit in the HPC solutions –an HPC system is actually a highly-integrated distributed system–, but requires an

²⁹ <https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus>

³⁰ <https://www.cerebras.net/product-chip/>

extra orchestrating layer in the software approaches.

Reproducing the analysis environment and software used to implement and test the federated analysis algorithm may be difficult. This can lead to a situation where the analysis algorithm does not work on site without modifications. The algorithm may also be inefficient, either in general or in the context of a specific database system. With large databases, inefficiency may lead to errors due to memory or network overloads, which also means that the algorithm will not work. There is a need to develop processes to resolve these situations. The algorithms should be developed and tested iteratively and jointly to ensure that the algorithm works in different computing systems and databases.

Finally, in terms of the software, there is now a common trend of pursuing the containerisation of the software solutions (e.g., using Docker or Singularity). The containers facilitate the replication of the environments and thus, better reproducibility of the analyses, which is key in the research domain. Additionally, the use of analysis notebooks (e.g., Jupyter Notebooks or Apache Zeppelin) is a common approach in the data science domain to perform regular statistical analyses and much more complex machine learning or AI algorithms. Notebooks store data in the notebook files and therefore using them in the context of version control and sensitive data requires extreme care.

10.2. Compute capabilities needed by the use cases

According to Article 67 of the EHDS Regulation, data access applications shall include, among others, a description of the tools and computing resources needed for a secure environment. Thus, researchers should consider the data volume prior to applying for data, as this might impact eligibility of available SPE's. In addition, the infrastructure should be ready to provide the computational requirements. Those needed for the five pilot UCs are listed below.

Surveillance of antimicrobial resistance use case led by ECDC

- Hardware
 - Secure processing environment (SPE) was used by each participating node, albeit with different hardware setups across nodes (two countries opted for virtual machines, while one country used a dedicated SPE service).
- Specific software
 - No containers were used for deploying the required analytical scripts, albeit it was considered as a valid option.
 - One of the SPEs required access via a specific VPN network. For the other two SPEs, multi-factor authentication was required in order to access.
 - The analytical bundle was provided as an R project package. It was deployed in each node SPE using RStudio and R dedicated package libraries. Both software components were previously installed in the designated SPEs.

Natural history of coagulopathy (blood clotting) related events in COVID-19 patients and risk factors, led by EMA

In the Finnish THL node, setting up the exact required software for the analyses was not possible, most likely because the required R package versions were very old. A working R package configuration was formed by using the latest R packages versions or versions

from another Darwin EU project. The OMOP database is very large and not all parts of the analyses ran successfully. The analysis involved joins using temporary tables, which is inefficient in Postgres, especially without an explicit analyses-procedure first. These inefficient joins probably caused the program to freeze.

The Danish team has completed all diagnostics and analyses on the OMOP DK-DHR database. We were able to use the R Shiny App.

- Hardware
 - THL specs used:
 - R process: Physical server, OS: Xubuntu 22.04, CPU: Xeon w3-2435 3.10 GHz, RAM: 256GB
 - Database backend: Postgres

Population uptake metrics: COVID-19 test positivity, vaccination and hospitalisation, led by Sciensano

- Hardware

For the Belgium node, the SPE used to analyse data is a virtual machine with the following resources : 4CPU & 64GB RAM. The minimum amount of RAM actually needed is around 16GB. This minimum specification of RAM may be higher for other nodes as it depends on the population size which can vary a lot depending on the node. No special GPU is needed as GPU accelerated tasks like AI processing for the use case were not needed.
- Specific software

R and R Studio are needed for this use case and supported inside the SPEs of the different Research teams to analyse the data. Regarding specific R libraries, most of the ones required for the use case were supported. If needed, more can be installed after an internal approval.

Comparing nationwide health trajectories to evaluate European Health Data interoperability: an application to cardiometabolic diseases, led by HDH in partnership with the University of Helsinki

- Hardware

Norwegian Institute of Public Health: Secure processing environment used to analyse data is a 1 Linux RHEL-based virtual with 1 TiB storage/backup machine, upgraded to 8 CPUs/GPUs, and 32 GiB RAM due to the volume of data.

University of Helsinki: CSC Sensitive Data (SD) Desktop is a service for analysing sensitive research data. OS: Ubuntu 22.04.3 LTS, processor: AMD Epyc CPU with 32 cores, 116GB of RAM, 1.2TB volume.

France: Secure Processing Environment was provided by HDH. OS: Ubuntu 24.10, Processor: 13th Gen Intel(R) Core(TM) i5-1335U.

Denmark: Public Health database servers hosted at Danmarks Statistik. OS: Windows Server 2012 R2 Standard 64-bit. Processor; Intel Xeon Gold 6130 @2.19GHz with 64

cores, 1024GB RAM, 23,5TB (Resources shared among multiple researchers in the department of Public Health).

- Specific software

Norwegian Institute of Public Health: R and RStudio and additional R libraries were needed installed on the virtual machine to run the analytic pipeline.

University of Helsinki: R and RStudio and additional R libraries were needed installed on the virtual machine to run the analytic pipeline.

France: R and RStudio, along with additional R libraries, were installed on the virtual machine to run the analytic pipeline. The dataset, consisting of 12 million records from France, required distributed processing capabilities. Initial data preprocessing was done in Python using the Jupyter Notebook environment, where Apache Spark was used for handling and processing the large dataset. Descriptive analysis, feature engineering, and the machine learning pipeline were then executed in RStudio with additional support from Spark for scalable data processing.

Denmark: R and RStudio. All libraries were available as Danmarks Statistik and have a local copy of CRAN (official libraries repository).

Genomic data linked to health data, with a focus on cancer, led by ELIXIR

- Hardware

The aim was to analyse VCFs using a specific tool (The Personal Cancer Genome Reporter (PCGR)³¹) provided by the Norwegian node. In order to request the hardware requirements of the SPEs, the reasoning below was followed:

For analysing 1 VCF file from WGS, 1 CPU and ~7Gb of RAM are needed. However, it depends on the size of the VCF (i.e., on the number of somatic mutations detected in each sample). Therefore, a buffer of 1 CPU and 1Gb was added to the previous estimations. Of note, the analysis software can create threads, therefore the addition of 1 CPU would speed up the process.

Decreasing the running time is also possible by analysing several VCFs simultaneously. Thus, the number of CPUs and Gb needed could be increased to analyse several genomes in parallel.

Regarding disk size, the available VCFs are used as a guideline. As the output of the analysis tool are annotated VCFs together with few text files, the disk size required would be two times the size of the original VCFs. Nonetheless, a buffer is also requested.

- Specific software

- Analysis software:

- The Personal Cancer Genome Reporter (PCGR)³²

- Containers:

³¹ <https://doi.org/10.1093/bioinformatics/btx817>

³² <https://doi.org/10.1093/bioinformatics/btx817>

- Similar to many high-performance computing (HPC) environments, the supercomputer designated for genomic analysis in Denmark (GenomeDK) does not allow Docker executions. This restriction arises due to elevated permissions that Docker requires for installation, which regular users do not have. Administrative rights are restricted for security reasons and the large number of users. Conversely, Singularity does not need such privileges and is therefore permissible in these settings. Thus, the use case opted to use Singularity across all nodes. A Singularity image has been provided by the Norwegian team, and it will be equally executed in all nodes enhancing portability and ensuring reproducibility.

11. Conclusions and recommendations

Given all the considerations gathered in the present document, clear guidelines and procedures with regard to data security and protection are of utmost importance in the upcoming HD@EU. Thus, this document could be used as a source of information to further design and develop the infrastructure. Building on the above sections, especially those bringing the lessons learnt from the use cases, these are the recommendations that could be taken into account when writing such document(s):

- Developing the correct guidelines entails to clearly **identify the threats to data protection and security** for those data that are part of the HD@EU.
- Apparently, taking the practical example of the use cases and considering the PETs applied by them, the **re-identification of individuals is a prominent threat within the HD@EU**. Thus, the implementing act should give the correct guidelines to apply PETs covering both prevention and mitigation. Specifically, **pseudonymisation and data minimisation** are the ones mostly used in the use cases, together with the **aggregation of the results**.
- In practice, the use cases are more concerned about **unauthorised access** than data loss as their main data security measures are focused on **managing restricted access to data**.
- The **EHDS implementation should avoid fragmentation in data protection approaches** as, despite the GDPR, differences in how Member States interpret and enforce data protection laws can create barriers to interoperability and data reuse. These inconsistencies might continue to impede the use of sensitive data, especially particularly sensitive data such as genomics.
- **SPEs** are out of the scope of this WP. However, they are used by the use cases and therefore repeatedly mentioned in this report. In general, clear and common guidelines are needed for such environments. The experiences gathered here from the use cases are very in line with the outcomes of the workshop on “Elements of Secure Processing Environments” celebrated in 2023³³.
- As the EHDS envisions a federated ecosystem, we recommend that different **scenarios of federated querying and analysis are considered** when building the data security and protection measures for the infrastructure, reflecting the differences seen in the use cases. However, agnostic of the approach, how to **analyse the software from the users** for working with health data is something that needs to be tackled. In addition, even though this pilot successfully demonstrated a technical method for retrieving

³³ Schlünder, I., Mayrhofer, M. T., & Ene, E. (2023). Elements of Secure Processing Environments (Workshop Report) V1.0. Zenodo. <https://doi.org/10.5281/zenodo.8341642>

aggregated data from multiple sources using predefined queries, it was limited to a specific use case. Therefore, further federated querying pilot projects would help to shape scalable implementation of this process within the HealthData@EU. Similarly, since piloting the entire user journey through all steps of the Health data request process in the EHDS Regulation was beyond the scope of this project, further efforts are needed to develop and refine this feature of the upcoming infrastructure.

- Cross-border data analysis is done following a **federated approach** in all use cases. Nonetheless, this approach is not enough for some specific tasks and the **data needed to be centralised** to perform the analyses. Therefore, this scenario should be envisioned and supported.
- The generation and use of **synthetic data** is becoming more and more common, especially thanks to the development of AI. Thus, it is important to have clear guidelines on how to assess the anonymity of such data.
- The heterogeneity of computational environments, database environments and the volume of data makes it difficult to design efficient and working algorithms for all situations. The following recommendations should be taken into account when going forward:
 - **Use containers and harmonise their use.** At the moment, the use of such applications is different among institutions and countries, even forbidden in some cases. However, if the proper security measures are applied, they are useful tools for ensuring the reproducibility of the analyses and avoiding incompatibilities.
 - Be mindful of the **data size and prepare the SPEs** for analysing big amounts of data.

Relevant references

During the generation of this report, we have identified some sources of information that, even not directly mentioned in the text, provide further insights on the topics of this deliverable:

- OECD (2023), "Emerging privacy-enhancing technologies: Current regulatory and policy approaches", OECD Digital Economy Papers, No. 351, OECD Publishing, Paris, <https://doi.org/10.1787/bf121be4-en>.
- SPUHiN, Norwegian EU funded project (direct grant) on SPEs. As part of this project, a report on a gap analysis was published: <https://www.fhi.no/globalassets/dokumenterfiler/helsedata/gap-analysis-report.pdf>
- TEHDAS Deliverable 7.2: Options for the services and services architecture and infrastructure for secondary use of data in the EHDS: <https://tehdas.eu/app/uploads/2023/07/tehdas-options-for-the-services-and-services-architecture-and-infrastructure.pdf>