



**HEALTHYCLOUD**  
Health Research & Innovation Cloud

## D3.1 – Landscape analysis of FAIRness levels of health-related data using catalogue matrix

### Document Information

Contract Number	965345
Project Website	<a href="http://www.healthycloud.eu/">http://www.healthycloud.eu/</a>
Contractual Deadline	M18, August 2022
Dissemination Level	PU- Public
Nature	R - Report
Author(s)	Shona Cosgrove (Sciensano), Pascal Derycke (Sciensano), Irene Kesisoglou (Sciensano)
Contributor(s)	Alicia Martinez Garcia (SAS) Carlos Luis Parra Calderón (SAS) Celia Alvarez Romero (SAS)
Reviewer(s)	Lorenz Dolanski-Aghamanoukjan (GÖG), Alba Jené (BSC)
Keywords	FAIR, data infrastructures, research



**Notice:** The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement N°965345

© 2021 HealthyCloud Consortium Partners. All rights reserved.

## Change Log

Version	Author	Date	Description of Change
V0.1	Shona Cosgrove, Pascal Derycke, Irene Kesisoglou (Sciensano)	29/07/2022	Initial Draft
V0.2	Lorenz Dolanski- Aghamanoukjan (GÖG)	03/08/2022	Formal review
V0.3	Alba Jené (BSC)	29/07/2022	Formal review
V1.0	Shona Cosgrove, Irene Kesisoglou, Pascal Derycke (Sciensano)	31/08/2022	Final version addressing reviewers' comments
			(Final Change Log entries reserved for releases to the EC)

## Table of contents

Executive summary	4
1. Introduction	6
2. Methods	7
2.1. Survey development	7
2.2. Use cases	8
2.3. Survey dissemination	9
2.4. Analysis	11
2.2.1. Survey analysis: feasibility of linking individual-level data	11
2.2.2. FAIRness evaluation: measuring the compliance of the different data infrastructures with the FAIR principles	11
3. Results	12
3.1. Administrative information on the different data infrastructures	13
3.2. Analysis of the survey results	19
3.3. Data quality	25
3.4. FAIR principles	29
3.4.1. Findability	29
3.4.2. Accessibility	33
3.4.3. Interoperability	39
3.4.1. Re-usability	42
3.5. FAIRness evaluation of the data infrastructures	44
4. Discussion	48
5. Conclusion and next steps	50
Annex 1	51
Annex 2	51
Table 1: Data controller and administrative information	51
Table 2a: Type of source	51
Table 2b: Type of source	51
Table 3: Level of aggregation	51
Table 4: Anonymisation	51
Table 5: Pseudonymisation	51
Table 6a to 6d: Geographical and time coverage	51
Table 6a: Geographical coverage	51

Table 6b: Participating countries	51
Table 6c: Socioeconomic coverage	51
Table 6d: Time coverage	51
Table 7: Ethical approval for storage of data	51
Table 8a and 8b: Data quality controls	51
Table 8a: Are data quality controls applied?	51
Table 8b: Are there minimum levels of quality of the data needed for the data to be included in the data infrastructure?	51
Table 9: Error checking	51
Table 10: Versioning of datasets	51
Table 11: Data source legitimacy	51

## Executive summary

The aim of WP3 of HealthyCloud is to carry out a landscape analysis of available health-related data infrastructures, in order to capture the European health data collections available for research purposes, evaluate their FAIRness level and determine the feasibility to perform individual level data linkages. Within this work, Task 3.1 and 3.2 focus on this landscape analysis and collect information about the data aspects of the available health data infrastructures and their adherence to the FAIR principles (Findable, Accessible, Interoperable, Re-usable)<sup>1</sup>.

To collect such information a survey was designed, in collaboration with the leaders of WP4. As an initial step, the study was focused on health data collections that would be useful to answer the research questions of the two use cases of WP7, the one on cancer and the other on atrial fibrillation.

This document, Deliverable 3.1, presents the final analysis of the survey results. The survey results were analysed to perform a FAIRness evaluation of the data infrastructures that have been selected for the scope of the use cases and also to answer the question of feasibility of linking individual level data. This deliverable builds on Milestone 3.2, which presented the initial analysis of the results relating to the cancer use case, focusing only on the results received from the Finnish data infrastructures. D3.1 extends the analysis to include all the survey results.

In relation to the cancer use case, the list of relevant data collections to survey was identified in collaboration with the cancer use case leaders and HealthyCloud partners from Belgium, Finland, Germany and Spain. The research question on cancer requires individual level data linkage from the health interview survey, health examination survey, cancer registry, genomic data collections and statistical registries for socioeconomic data. The research question on atrial fibrillation requires different types of data from patient registries in Europe, such as ECG, MRI, biomarkers, genomic and clinical data. The list of relevant data collections was identified in collaboration with the atrial fibrillation use case.

We received 20 responses to the survey out of a total of 26 surveys sent.

In this document we present the administrative information about each data infrastructure, information about the data they provide, such as the level of aggregation, whether it is anonymised or pseudonymised, about data quality aspects, coverage and standards used to structure the data and regarding the compliance with the FAIR principles.

The findings supportive of a high level of feasibility of linking individual level data include the fact that all the data infrastructures store individual-level data, have national-level coverage and have pseudonymised data. On the other hand, a finding

---

<sup>1</sup> Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

that may hamper data linkage at individual level between different Finnish data collections, for example, is the lack of interoperability due to the usage of different standards to structure their data or metadata.

Finally, we also published on ZENODO (<https://doi.org/10.5281/zenodo.7038397>) an open notebook (e.g. Rmarkdown notebook<sup>2</sup>) for reproducing the FAIRness evaluation and the general analysis of the data infrastructures performed during this project. This **HealthyCloud FAIRness self-assessment tool** is a 2-in-1 tool allowing the publication of the HealthyCloud FAIRness evaluation survey and the production of a report including pie charts demonstrating the percentage scores for each FAIR principle as well as an overall score.

Following this deliverable we are planning on creating a publicly available online catalogue with the information collected for each of the data infrastructures in order to feed the metadata catalogue being prepared by WP6 and hence make these data infrastructures discoverable to external researchers.

---

<sup>2</sup> <https://jupyter.org/>

## 1. Introduction

The aim of Work Package 3 of HealthyCloud is to carry out a landscape analysis of available health-related data infrastructures, in order to capture the European health data infrastructures available for research purposes, evaluate their compliance with the FAIR principles (Findable, Accessible, Interoperable, Re-usable) and determine the feasibility to perform individual level data linkages.

Within this work, Task 3.1 (led by Sciensano) focuses on performing a landscape analysis of available health-related data infrastructures, collecting information about the infrastructure (such as quality assurance aspects and storage) and adherence to the FAIR principles<sup>3</sup>. To collect such information, a survey was designed in collaboration with the leaders of WP4 in a form of a catalogue matrix.

As an initial step, we decided to focus this study on the health data infrastructures that would be useful to answer the research questions of the two use cases, on cancer and atrial fibrillation.

The research question of the cancer use case (use case 1) assesses how genomic information, gathered at population level, can contribute to developing high-risk profiling for the major risk factors for cancer, e.g. tobacco, alcohol, obesity, sun-exposure, family history, socio-economic status. This question requires linkage of individual level genomic data with cancer registry, health interview survey, health examination survey and socioeconomic data. For further details on use case 1, see D7.1.

The use case on atrial fibrillation (use case 2) aims to identify subgroups of atrial fibrillation patients from the diagnosis stage to develop and apply personalised medicine approaches. To address this issue, this use case would require linkage between different types of data, such as ECG, MRI, biomarkers, genomic and clinical data, from different patient registries in Europe. For further details on use case 2, see D7.2.

This document, Deliverable 3.1, presents the final analysis of the survey results of data infrastructures relevant to the two use cases, as well as the results of the FAIRness evaluation assessment of the different health data infrastructures using an adapted FAIRness evaluation tool to provide a landscape analysis of the FAIRness levels using the survey.

These findings will be further developed and consolidated in Deliverable 3.3 'Landscape analysis using a health related-data catalogue matrix', which is due in April 2023.

---

<sup>3</sup> Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

## 2. Methods

### 2.1. Survey development

The survey used for this study was developed in collaboration with SAS (Servicio Andaluz de Salud), WP3 and WP4 co-leads, and with CRG (Centre for Genomic Regulation), WP4 co-leads, to combine efforts and avoid sending multiple similar surveys to the same data infrastructures.

To develop the survey, we took into consideration the following aspects:

- The organisation and governance of the data infrastructures;
- The nature of the data;
- The type of data sources and level of detail;
- The data storage process;
- The findability, accessibility, interoperability and re-usability of the data and metadata. The compliance with the FAIR principles, as defined by the Research Data Alliance (RDA).

The format used for the survey was a catalogue matrix that responders had to fill in. This catalogue matrix includes over 50 indicators (questions) under the following ten areas:

1. Administrative;
2. Data;
3. Completeness of the data collection;
4. Quality aspects of the data collection;
5. Metadata;
6. Findability;
7. Accessibility;
8. Interoperability;
9. Re-usability;
10. Governance.

The survey underwent several rounds of feedback with the HealthyCloud partners. It was then piloted by four data infrastructures (two data collections and two data hubs)<sup>4</sup> and further refined based on their feedback. The final version of the survey can be found [here](#) (online tool form version)<sup>5</sup> and in [Annex 1](#).

---

<sup>4</sup> See HealthyCloud Glossary for definitions of data collection and data hub:  
<https://zenodo.org/record/6787119#.YvZl1XZByUm>

<sup>5</sup> HealthyCloud WP3 and WP4 survey. Online tool. Available at:  
<https://bsc3.typeform.com/to/zY1FNgSQ>



## 2.2. Use cases

The scope of the first landscape analysis presented in this Deliverable 3.1 was set around the two use cases of HealthyCloud, namely the use case on cancer and the one on atrial fibrillation. Therefore, in collaboration with the task leaders of WP7 responsible for these reference use cases, we identified the data collections that contain the various data essential to conduct the studies of the use cases and answer the research questions.

### **Cancer use case**

The cancer use case aims to evaluate the feasibility of linking individual level data between different data collections within countries in order to study how genetic predisposition and environmental factors interact to increase the susceptibility of a person to develop cancer. It aims to lay the ground for development of polygenic risk scores and understand the cancer risks combining genetic with non-genetic variables. This will have to be conducted within a country as it requires the linkage of individual level data across different data collections, infrastructures or registries, such as the health interview survey with data from the cancer registry and the genomic data collection.

Therefore, we collaborated with HealthyCloud partners from Belgium, Finland, Spain and Germany and the use case leaders and identified the data collections in each country that would be needed to conduct such a study. The most complete lists of identified data infrastructures that would be needed to answer the research question of the first use case were in Belgium and Finland.

After sending the survey to the identified data infrastructures, analysing and extracting the answers received, the cancer use case leaders could contact the data controller and provider of each data collection and request the specificities of the different variables that they collect. Then, according to the accessibility procedure mentioned in their answers to the survey, WP7 partners could request access to the specific datasets they need to perform the research project and answer the research question described above.

### **Atrial fibrillation use case**

Atrial Fibrillation (AF) is the most frequently encountered cardiac arrhythmia in clinical practice. It manifests as an irregular and often rapid heart rate that might increase risk of strokes, heart failure and even death. The main issue with AF is its diagnosis at an early stage as there are a lot of asymptomatic cases. An early diagnosis of AF could prevent strokes by offering anticoagulation treatments.

The use case leaders aim to explore an integrative model considering different modalities of AF incidents. They propose the combination of clinical data, imaging data, biomarkers, electrocardiogram (ECG) signals and genetic variants into an integrative model. This model aims to detect subgroups within the population of AF

patients of the UK Biobank (UKB) cohort in a first stage, and then extended to other cohorts to generalise the model in a federated learning scheme.

The challenge in this AF study is to collect enough data. This is why this use case requires cross-border collection and integration of data. Therefore, the leaders of the use case collaborate with the data controllers from the UK, Spain, France, Germany and European data collections. We hence sent the survey to the different data infrastructures with which they collaborate<sup>6</sup>.

### 2.3. Survey dissemination

After modifications and refinement, the survey was sent to more than 28 data collections in the scope of WP3.

The research question on cancer would require individual level data linkage from the following data collections: the health interview survey, the health examination survey, the cancer registry, the genomic data collection and statistics office for socioeconomic data. The list of relevant data infrastructures was identified in collaboration with the cancer use case leaders and HealthyCloud partners from the use case countries (Belgium, Finland, Germany and Spain).

The research question on atrial fibrillation would require different types of data from patient registries in Europe, such as ECG, MRI, biomarkers, genomic and clinical data. The list of relevant data collections was identified in collaboration with the atrial fibrillation use case leads.

**Table 1:** Data collections to which the survey was sent.

Cancer use case	
Belgium	Belgian Cancer Registry
Belgium	Belgian Registry on Genomic Data
Belgium	Health Interview Survey and Health Examination Survey
Belgium	Statbel
Finland	Avohilmo, Register of Primary Care Visits
Finland	Care Register for Social Welfare

---

<sup>6</sup> Petersen, S. E et al. "The impact of cardiovascular risk factors on cardiac structure and function: Insights from the UK Biobank imaging enhancement study." PLOS vol 12 (2017):10.

Finland	Findata
Finland	FinHealth 2017 Survey
Finland	Finnish Cancer Registry
Finland	Finnish Social Science Data Archive
Finland	FinSote
Finland	Research Services at Statistics Finland
Finland	THL Biobank
Germany	Survey was sent to German contacts in Charite and TMF for dissemination to relevant data infrastructures
Spain	Cancer Registry of Granada
Spain	Genomics registry SAS
Spain	Red Española de Registros de Cáncer (REDECAN)
Spain	Registro de Cáncer Poblacional de Castilla y León (RECA)
<b>Atrial fibrillation use case</b>	
European	BigData@Heart
Finnish	Biobank of Eastern Finland
French	Atrial Fibrillation registry
French	MICCAI 2012 Right Ventricle Segmentation Challenge
French	MICCAI 2017 ACDC
Germany	Study of Health in Pomerania
Spain	FANTASIIA Registry
Spain	FAPRES Registry
Spain	REVERSE Registry

European Research Infrastructures relevant to both use cases	
European	BBMRI
European	EuroBioImaging

HealthyCloud partners also indicated that the European Research Infrastructures BBMRI and EuroBioimaging could be useful to these two use cases. The survey was sent to these research infrastructures in the scope of WP4, therefore the WP4 leads shared the responses they had received.

## 2.4. Analysis

### 2.2.1. *Survey analysis: feasibility of linking individual-level data*

A qualitative analysis of the survey responses was carried out, with the aim of studying the feasibility of linking individual level data across the included data infrastructures within the specified countries, relevant to the cancer and atrial fibrillation use cases.

### 2.2.2. *FAIRness evaluation: measuring the compliance of the different data infrastructures with the FAIR principles*

The Research Data Alliance's FAIR Data Maturity Model published in June 2020 served as a general framework for the FAIRness evaluation of the collected survey responses<sup>7,8</sup>. We organised a series of workshops where different experts from projects aiming to 'FAIRify' data infrastructures, such as RDA, GO FAIR, FAIR PLUS, FAIR4HEALTH and EJPRD, presented the available FAIRness evaluation tools. After examining in depth the availability of web based tools endorsing the FAIR Data Maturity Model, we decided to use the ARDC FAIR Data self-assessment tool published by the Australian Research Data Commons (ARDC)<sup>9</sup> as a base to assess the FAIRness level of the data infrastructures.

The ARDC FAIR Data self-assessment tool consists of a HTML Web page with functionalities coded in Javascript. We have customised and integrated the existing tool in an Rmarkdown notebook and extended its functionalities. The new **HealthyCloud FAIRness self-assessment tool** is a 2-in-1 tool allowing the publication of the HealthyCloud FAIRness evaluation survey and the production of

<sup>7</sup> FAIR Data Maturity Model: specification and guidelines - RDA FAIR Data Maturity Model Working Group - DOI: 10.15497/rda00050

<sup>8</sup> Bahim, C, et al. 2020. The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. Data Science Journal, 19: 41, pp. 1–7. DOI: <https://doi.org/10.5334/dsj-2020-041>

<sup>9</sup> <https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/>

a report including pie charts demonstrating the percentage scores for each FAIR principle as well as an overall score.

The HealthyCloud FAIRness self-assessment tool has been made freely accessible on a public BinderHub portal hosted by the community at [mybinder.org](https://mybinder.org) allowing any user to produce the FAIRness evaluation and the general analysis of their data collections<sup>10</sup>.

The FAIRness evaluation reports produced by the tool can be updated at any time as a csv file, which can be downloaded and will serve to produce a new updated report from the tool.

The HealthyCloud FAIRness self-assessment tool along with the list of questions we used to gather information about data governance and quality aspects of the different data collections will be published on ZENODO. The FAIRness self-assessment tool is already published on ZENODO (<https://doi.org/10.5281/zenodo.7038397>). By sharing the survey, the methodology and a tool, we offer the means to expand the landscape analysis to more data collections, and expect to facilitate analysis by making it more user friendly.

The HealthyCloud FAIRness self-assessment tool includes quick user instructions on how to proceed with the tool. A Readme file is also accessible on GitHub (<https://github.com/PderyckeSciensano/HEALTHYCLOUD/>) and all information about the tool can be found on ZENODO [<https://doi.org/10.5281/zenodo.7038397>].

### 3. Results

20 responses to the survey were received, as demonstrated in Table 2.

**Table 2:** Data collections from which responses were received and analysed.

Country	Data infrastructures
Belgium	Belgian Cancer Registry
Belgium	Belgian human genomics project
Belgium	Health Examination Survey
Belgium	Health Interview Survey
Belgium	Statistics Belgium
European	BBMRI-ERIC

<sup>10</sup> <https://ovh.mybinder.org/v2/gh/PderyckeSciensano/HEALTHYCLOUD/main?urlpath=rstudio>.

European	EuroBioImaging Italian MMMI Node
Finland	Avohilmo, Register of Primary Health Care Visits
Finland	Care Register for Social Welfare (Sosiaalihuollon hoitoilmoitusrekisteri)
Finland	Findata
Finland	FinHealth 2017 survey, Health Examination Survey
Finland	Finnish Social Science Data Archive
Finland	Finnish Cancer Registry
Finland	FinSote, Health Interview Survey
Finland	Research Services at Statistics Finland
Finland	THL Biobank
Germany	State of Health in Pomerania (SHIP)
Spain	European Genome-Phenome Archive (EGA)
Spain	SAS genomic data, Collaborative Spanish Variant Server
Spain	Plataforma de Información BIGAN, IACS

The analysis of survey results below and the accompanying tables include data infrastructures relevant to both the cancer use case and the atrial fibrillation use case of HealthyCloud.

### 3.1. Administrative information on the different data infrastructures

The survey included questions regarding the data controller and data processor of the data infrastructure.

These terms have previously been defined in the HealthyCloud glossary<sup>11</sup>, following discussion in the glossary working group meetings with consortium members:

---

<sup>11</sup> Irene Kesisoglou, Shona Cosgrove, Pascal Derycke, Petronille Bogaert, Annika Jacobsen, Marco Roos, Anna Niemeyer, Alicia Martinez Garcia, Adrian Thorogood, Petr Holub, Irene Schluender, Salvador Capella, Juan Gonzalez Garcia, & HealthyCloud consortium. (2022). Glossary of commonly

→ **Data controller:**

Under Regulation (EU) 2018/1725, as well as under the GDPR, the data controller is the party that, alone or jointly with others, determines the purposes and means of the processing of personal data. The actual processing may be delegated to another party, called the data processor. The controller is responsible for the lawfulness of the processing, for the protection of the data, and respecting the rights of the data subject. The controller is also the entity that receives requests from data subjects to exercise their rights.<sup>12 13</sup>

→ **Data processor:**

According to Article 3 (12) of Regulation (EU) 2018/1725, a processor shall mean "a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller." The essential element is therefore that the processor only acts "on behalf of the controller" and thus only subject to his instructions.<sup>14</sup>

In some cases, the processor may choose not to process the data himself, but may have recourse to a subcontractor who processes the data on his behalf. In practice, this will depend upon the processor agreement entered into with the controller.

→ **Data provider/data holder:**

Any natural or legal person, which is an entity or a body in the health or care sector, or performing research in relation to these sectors, as well as European Union institutions, bodies, offices and agencies who has the right or obligation, or the ability to make available, including to register, provide, restrict access or exchange certain data.<sup>15</sup>

*Who is the data controller, provider or processor for each data infrastructure?*

---

used terms in the field of health data research - developed by the EU project HealthyCloud (0.1). Zenodo. <https://doi.org/10.5281/zenodo.5998128>

<sup>12</sup> Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018.

Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725&from=EN>

<sup>13</sup> Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN>

<sup>14</sup> Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018.

Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725&from=EN>

<sup>15</sup> Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space (EHDS). 3 May 2022. Available at: [https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC\\_1&format=PDF](https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF)

The table representing our findings for data infrastructures that responded to the survey is available in [Annex 2, Table 1](#): Data controller and administrative information.

### **Defining what criteria describe/respond to a data collection**

When participants were asked how they would define their data infrastructure, and what characteristics describe their data infrastructure, we received the following answers from the data infrastructures surveyed:



**Table 3:** Data infrastructure responses to the question ‘Which of the following characteristics fit your data infrastructure?’

	Finland									Belgium					Spain		Europe		Germany	
	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	Eurobioimaging Italian MIMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
A digital platform that receives and stores data	X			X			X	X	X			X	X		X	X	X	X		11
It receives data from a single source and/or multiple sources	X	X	X	X	X	X	X	X	X			X	X	X	X	X	X	X	X	17
It has control over the data stored	X	X	X	X	X	X	X	X		X	X		X		X	X	X	X	X	16
It allows discovery of health datasets				X	X	X				X	X				X	X	X	X		9
It has a specific thematic, data type that it collects (e.g. a particular disease, a particular data type: genomic data, clinical data, EHRs...)		X		X				X		X	X	X	X			X	X	X		10
It is part of one or more overarching data hubs		X		X									X				X	X		5

It generates data	X							X		X	X	X					X		X	7
A digital technical infrastructure with the core mission of enabling health data sharing				X											X	X	X	X		5
It provides health data from different sources				X	X	X		X	X			X			X			X	X	9
It has a metadata discovery service			X	X	X	X				X	X					X		X	X	9
It has a data accessibility mechanism in accordance with existing regulation	X		X	X	X	X	X	X		X	X		X		X	X		X	X	14
It has an authorization functionality, provided by the same Data Hub or by an external institution			X	X	X			X				X			X				X	7

**Note:** no response from EGA for this question (see limitations section)

Key:

- Red = minimal inclusion criteria for a **data collection**
- Black = other possible characteristics of a **data collection**
- Orange = minimal inclusion criteria for a **data hub**
- Yellow = minimal inclusion criterion for both a **data collection** and a **data hub**

### **Conclusion: minimum criteria to define a data collection**

From the table above we can conclude that the minimal inclusion criteria described in the glossary under the term 'data collection' correspond with the characteristics that describe most of the data infrastructures that participated in the survey. Of the 19 data infrastructures that responded to this question, 17 (89.5%) stated that they receive data from one or multiple sources. 16 (84.2%) of the data infrastructures have control over the data stored. A smaller proportion but still over half of the respondents characterise themselves as a digital platform that receives and stores data (57.9%).

The only characteristic under the inclusion criteria for the term 'data hub' that was selected by over half of the respondents was that of having a data accessibility mechanism in accordance with existing regulation. This was selected by 14 (73.7%) of respondents across all countries. Therefore, we could consider adding this criterion also to the 'data collection' definition and we will discuss it in the next glossary working group.

The three characteristics belonging to the group of 'other possible characteristics of a data collection' in the HealthyCloud Glossary were all selected by a smaller proportion of respondents. 10 (52.6%) of the data infrastructures surveyed have a specific theme. Only 7 (36.8%) generate data, and only 5 (26.3%) are part of one or more overarching data hubs. As previously described in Milestone 3.1 based on the analysis of the Finnish results at the time, this suggests that these last two criteria could be removed from the definition of 'data collection' in the HealthyCloud Glossary, which is now supported by the analysis of responses from other countries.

From this analysis we could conclude that the respondents to the survey were mostly data collections rather than data hubs, which fits with the scope of the use cases requiring individual level linkage of health data.

### 3.2. Analysis of the survey results

#### Type of source

In terms of the data source, responses were received from 19 data infrastructures. The data infrastructures could select multiple options between the following: general population, patient group, experimental setting, or other. The majority (13 data infrastructures, 68.4%), have data from the general population. 6 data infrastructures (31.6%) have data from a patient group. 2 (10.5%) have data from an experimental setting. Finally, 5 data infrastructures responded that they have data from other sources. For instance, the Finnish Cancer Registry has data from cancer screening, and the Belgian Cancer Registry has data from patients diagnosed with cancer or from cancer screening.

The full responses to this question can be found in [Annex 2, Table 2a](#): Type of source.

Regarding the type of data source, the data infrastructures could choose multiple options of types of source (e.g., electronic health records, clinical trials, surveys etc). Only 6 of 19 data infrastructures (31.6%) had a single type of data source. For instance, FinSote in Finland as well as the Health Interview and Health Examination Surveys in Belgium only contain survey data. The EuroBioImaging Italian MMMI Node only contains imaging data. The Belgian Genomic Data Registry and the Collaborative Spanish Variant Server (CSVS) only contain genomic data.

All other data infrastructures contain data from multiple types of data sources. Only one data infrastructure (BBMRI-ERIC) contains data from clinical trials. The full responses to this question can be found in [Annex 2, Table 2b](#): Type of source.

#### Level of aggregation

In terms of the level of aggregation for the data stored in the data infrastructure (i.e., aggregated, individual or both), 19 out of 20 respondents (95%) have individual level data (15 have only individual level data, and 4 have both individual and aggregated data). Only one data infrastructure relevant to the cancer use case, the Collaborative Spanish Variant Server (CSVS) in Spain, only has aggregated data.

This is a key finding as the scope of the research question is to determine the feasibility of linking *individual*-level data across data infrastructures.

The responses can be found in [Annex 2, Table 3](#): Level of aggregation.

#### Anonymisation/pseudonymisation techniques used

**Anonymisation** techniques differ between the infrastructures surveyed.

Overall, over two thirds of data infrastructures (12 of 19 responses received, 63.2%) perform anonymisation at some point of the data life cycle. The most common response, with 6 of 19 responses (31.6%) was that anonymisation is performed

before sharing data externally. 4 data infrastructures (21%) anonymise data at the point of publishing (i.e., they perform analyses with identifiable data and only anonymise when publishing the results/paper). Only 2 data infrastructures (10.5%) relevant to the cancer use case - the Avohilmo Register of Primary Care Visits in Finland and the Collaborative Spanish Variant Server (CSVS) in Spain - anonymise data at the point of collection. This is an important finding, as anonymisation at the point of collection reduces the feasibility of linking individual level data.

5 of the data infrastructures (26.3%) do not anonymise data at all.

The responses to this question can be seen in the table below:

**Table 4:** Anonymisation methods used by the data infrastructures surveyed

	Finland									Belgium					Spain			Europe		Germany	
Are anonymisation methods used with the data?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
Yes: at the point of collection									X							X					2
Yes: before sharing them externally				X		X		X		X	X				X						6
Yes: before sharing them internally																					0
Yes: at the point of publishing		X	X											X						X	4
No: we do not anonymise data	X				X		X						X					X	X		5
I don't know																					0
This question doesn't apply to this data infrastructure												X									1

**Note:** No response from EGA for this question (see limitations section)

Of the 12 data infrastructures who work with anonymised data, 11 perform the anonymisation themselves (91.7%) while only one Spanish data infrastructure relevant to the cancer use case - the Collaborative Spanish Variant Server (CSVS) - receives already anonymised data. Only one data infrastructure for the cancer use case - the Finnish Social Science Data Archive - stated that it can both receive already anonymised data or perform the anonymisation in-house.

The responses to this question can be found in [Annex 2, Table 4](#): Anonymisation.

In terms of **pseudonymisation**, 17 of the 19 respondents (89.5%) have pseudonymised data. Only 2 data infrastructures - the Collaborative Spanish Variant Server (CSVS) in Spain and the EuroBioImaging Italian MMMI Node - do not have pseudonymised data. The responses to this question can be found in [Annex 2, Table 5](#): Pseudonymisation.

For the data infrastructures that have pseudonymised data, the organisation that holds the method to reverse the pseudonymisation process differs. Some data infrastructures have a trusted third party (TTP) that holds the method to reverse the pseudonymisation, while others hold the method to reverse it themselves. The responses to this question are shown in the following table:

**Table 5:** Organisation that holds the method to reverse the pseudonymisation process.

	Finland									Belgium					Spain		Europe	Germany
If yes, who (name of the organisation or stakeholder) holds the method to reverse the pseudonymisation process? (e.g. key, dictionary, map, table)	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	EGA	BBMRI-ERIC	State of Health in Pomerania (SHIP)
Free text	Finnish Institute for Health and Welfare (THL)	THL	Us at Research Services	The data depositors	The biobank keeps the pseudonymization keys	Findata	THL	Finnish Cancer Registry		Statbel	Statbel	TTP eHealth	Sciensano DS epidemiology and public health	Statbel	Servicio Aragonés de Salud, IACS	The data owners	Data source	Us

**Note:** One data infrastructure which has pseudonymised data (Avohilmo Register of Primary Care Visits) did not provide a response for the organisation who holds the method to reverse the pseudonymisation process.



## Geographical coverage

In terms of geographical coverage, 16 data infrastructures (80%) contain only national-level data.

3 data infrastructures (15%) contain international and European-level data. The Finnish Social Science Data Archive contains international, European, national and regional level data, with a broad representation of participating countries from every continent. BBMRI-ERIC contains only international and European-level data. The participating countries depend on the particular data collection. It can be any BBMRI-ERIC member/observer country, or, for COVID-19 and rare diseases, it can be completely global. EGA contains data from participating countries worldwide.

Only 4 data infrastructures (20%) across both the cancer and atrial fibrillation use cases have regional-level data: the Finnish Social Sciences Data Archive, Statbel, Plataforma de Información BIGAN and the Study of Health in Pomerania (SHIP). Of these, only the latter two have data solely at the regional level, without national-level data.

This is an important finding, as having solely regional coverage reduces the feasibility of linking individual-level data within a country or across countries.

The responses to this question can be found in [Annex 2, Table 6a](#): Geographical coverage and Table 6b: Participating countries.

The data infrastructures were asked the socioeconomic coverage of the data in their data infrastructure (based on the NUTS classification). The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU and the UK for the purpose of collection, development and harmonisation of European regional statistics:

- NUTS 1: major socioeconomic regions
- NUTS 2: basic regions for the application of regional policies
- NUTS 3: small regions for specific diagnoses

19 data infrastructures responded to this question. 8 (42.1%) have coverage across all three NUTS levels. 4 data infrastructures (21.1%) have NUT3 coverage only: FinHealth 2017 survey, FinSote, the Belgian Cancer Registry, and Plataforma de Información BIGAN in Spain. Another 4 data infrastructures (21.1%) have NUTS1 coverage only: the Belgian Genomic Data Registry, the Collaborative Spanish Variant Server (CSVS), the EuroBioImaging Italian MMMI Node and BBMRI-ERIC. One data infrastructure - THL Biobank in Finland - has coverage across NUTS 1 and 2. One data infrastructure stated that they did not know (Findata).

The responses to this question can be found in [Annex 2, Table 6c](#): Socioeconomic coverage.

In terms of **time coverage**, 15 of the data infrastructures have ongoing data collection. 5 of these stated that there is no specific end data for the data collection. Conversely, other data collections (such as FinHealth 2017 Survey in Finland and the Health Interview Survey and Health Examination Survey in Belgium) have specified periods of data collection, 2017 and 2020 respectively.

The full responses to these questions can be found in [Annex 2, Table 6d](#): Time coverage.

### **Ethical approval for storage of data**

Regarding whether ethical approval is required for data to be stored in the data collections, the picture is varied.

Just below half of data infrastructures surveyed (8 respondents, 42.1%) do not require ethical approval for data to be stored in their data infrastructure. 6 infrastructures (31.6%) do require ethical approval for data to be stored. 1 infrastructure responded that they do not know, while 5 stated that this question is not relevant to their data infrastructure.

The responses to this question can be found in [Annex 2, Table 7](#): Ethical approval for storage of data.

## **3.3. Data quality**

### **Data quality controls**

In terms of data quality, 95% (19 of 20) of data infrastructures surveyed apply data quality controls on the data. EuroBioImaging Italian MMMI Node was the only data infrastructure that reported that it does not apply data quality controls. However, it is important to note that this data infrastructure responded to the next question that quality controls are applied for internal use only, indicating that some quality controls are used. This may indicate a misinterpretation of the first question.

12 of 20 respondents (60%) only include data in their data infrastructure if it reaches a minimum quality level. 6 data infrastructures (30%) apply quality controls for internal use only, but do not apply minimum levels of quality for inclusion. 1 data infrastructure, EGA, does not apply minimum quality levels for inclusion, but the results of the quality control are available when searching for the data. Finally, one data infrastructure, Findata, responded that this question does not apply to them.

The responses to this question can be found in [Annex 2, Tables 8a and 8b](#): Data quality controls.

### **Updating periodicity**

The updating periodicity between the data infrastructures varied widely, with the majority of data infrastructures updating their data irregularly or at long time intervals.

The most common response (7 data infrastructures, 35%), was that they update their data only irregularly. The next most common response (6 data infrastructures, 30%) was that data is updated on an annual basis. 4 data infrastructures (20%) report that they perform one-time collection. Only 3 data infrastructures (15%) update data on a daily basis, 2 (10%) do so on a monthly basis, and 1 (5%) on a weekly basis.

Overall, the majority of data infrastructures report updating data only irregularly, annually, or have once only collection. The responses from this question are shown in the table below:

**Table 6:** Updating periodicity of the data infrastructures surveyed.

	Finland									Belgium					Spain			Europe		Germany	
How often do you update the datasets?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
Daily									X						X					X	3
Weekly															X						1
Monthly			X												X						2
Annually		X	X				X	X						X	X						6
Biannually																					0
Every 2+ years																					0
Every 5+ years										X	X										2
Irregularly	X		X		X							X			X			X	X		7
One time collection						X							X			X			X		4
I don't know																					0
This doesn't apply to this data infrastructure				X																	1

## Error checking

14 data infrastructures (70%) stated that they use a tool to check for errors and completeness of data, whilst 6 (30%) do not. The responses to this question can be found in [Annex 2, Table 9](#): Error checking.

Those who use a tool were asked which tool they used, with varied responses. 3 of the data infrastructures who use a tool for error checking did not report which tool. Of those who did respond, responses varied, with some using proprietary tools and others using tools such as Checksum.

**Table 7:** Error checking tool

	Finland			Belgium				Spain		Europe	Germany
If yes, what tool do you use (e.g., Checksum)?	Finnish Social Science Data Archive	THL Biobank	Finnish Cancer Registry	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Collaborative Spanish Variant Server (CSVS)	EGA	BBMRI-ERIC	State of Health in Pomerania (SHIP)
Free text	Checksum	Different tools are used for different datasets	IARC checking tool, own checking algorithms, change tracking, time comparison of statistics	Internal coherence is checked	Internal coherence is checked	Custom built tool	Checksums	QC control checks data integrity as well	We check data when enters and exit. we check file matches type to what is uploaded	TLS-backed TCP transmission (hence checksums)	Proprietary tool

## Versioning of datasets

The data infrastructures were asked if they have a process to keep track of the different versions of datasets.

12 data infrastructures (60%) have such a process, whilst 5 (25%) do not, and 3 (15%) stated that this does not apply to their data infrastructure. The responses to this question can be found in [Annex 2, Table 10](#): Versioning of datasets.

Of the data infrastructures who do have a process to keep track of the versions, the processes varied:

**Table 8:** Processes for keeping track of versions of datasets.

	Do you have a process to keep track of different versions of datasets? If yes, please specify the process.	
Finland	Research Services at Statistics Finland	Date – name

	<b>Finnish Social Science Data Archive</b>	First version of the dataset gets the version number 1.0. The version number is updated, if changes or updates are made into the dataset (Major/minor change -> first or second number). We describe the changes and keep track of them in our internal database.
	<b>THL Biobank</b>	We have many different datatypes, each has different process for versioning, depending also on the database in which the data is stored.
	<b>FinSote</b>	Through relational database
	<b>Finnish Cancer Registry</b>	Every 6 monthly release of cancer data is kept and can be traced back. The system has been ongoing since the year 2014
<b>Belgium</b>	<b>Health Interview Survey</b>	Indication of the version
	<b>Health Examination Survey</b>	Indication of the version
<b>Spain</b>	<b>Collaborative Spanish Variant Server (CSVS)</b>	Every update
<b>European</b>	<b>BBMRI-ERIC</b>	Internal in the database on any database update.
<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	Versioning options within PostgreSQL

2 data infrastructures (Avohilmo Register of Primary Care Visits in Finland and Plataforma de Información BIGAN in Spain) responded that they have a process for versioning of datasets but did not specify further.

### Data source legitimacy

8 data infrastructures (40%) responded that they do not have a method to check data source legitimacy, 1 data infrastructure responded that they did not understand the question, and 4 data infrastructures did not respond to the question.

Of the 7 data infrastructures (35%) who do have such a method, the method used varied. For instance, Findata stated that its data is only coming from official health and social care sectors. The Belgian Cancer Registry stated that registered data needs to meet ENCR, IARC International Guidelines. The Belgian Genomic Data Registry does not have such a method yet, however a working group is expected to deliver standards on data quality in the coming years that will be followed.

The full responses to this question can be found in [Annex 2, Table 11](#): Data source legitimacy.

## 3.4. FAIR principles

### 3.4.1. Findability

#### Metadata

Having a publicly available metadata catalogue, presenting information regarding the datasets stored or controlled by a data infrastructure and the way to access them is essential for discovery and re-use by researchers or other users that need access to these datasets.

The survey results revealed that 13 out of the 20 (65%) data infrastructures that responded produce or collect metadata for the datasets they are storing or are data controllers of. 14 out of 20 (70%) responded that they also have a public metadata catalogue service available where a researcher can find information about their data collection. However, only 12 out of 14 data infrastructures provided the URL of the publicly available metadata catalogue.

Interestingly, 4 out of 20 data infrastructures also have an online catalogue with the datasets they store and control but this is accessible only using a proprietary search engine.

Finally, only 7/20 data infrastructures had a metadata record API endpoint in place (see table interoperability). This is important because it determines the readability and compatibility of this metadata record with other existing metadata catalogues.

**Table 9:** Data infrastructures responses in relation to metadata and metadata catalogues

	Data infrastructure	Do you produce or collect metadata for your data?	Do you have a public metadata catalogue service? If yes, what is the URL?
Belgium	Belgian Cancer Registry	Handbook, description variables, Guideline data infrastructure tool, registration handbooks,...)	No
	Health Examination Survey	Codebook + Manual for external users	Yes
	Health Interview Survey	Codebook + Manual for external users	Yes
	Genomic data registry	No	No
	Statbel		No
European	BBMRI-ERIC	Yes (typically according to MIABIS model)	Yes, <a href="https://directory.bbmri-eric.eu/">https://directory.bbmri-eric.eu/</a>
	EuroBioImaging Italian MMMI Node	Yes	No
Finland	Avohilmo, Register of Primary Care Visits		Yes, <a href="https://www.julkari.fi/bitstream/handle/10024/138288/URN_ISBN_978-952-343-346-5.pdf?sequence=1&amp;isAllowed=y">https://www.julkari.fi/bitstream/handle/10024/138288/URN_ISBN_978-952-343-346-5.pdf?sequence=1&amp;isAllowed=y</a>

	<b>The Care Register for Social Welfare</b>		Yes, Sosiaalihuollon hoitoilmoitusrekisteri 1995- (Sosiaalihilmo)
	<b>Findata</b>	Data controllers expected to provide the data descriptions in Aineistoeditori (a tailor-made tool)	Yes, <a href="https://aineistokatalogi.fi/catalog">https://aineistokatalogi.fi/catalog</a>
	<b>FinHealth 2017 Survey</b>	No	No
	<b>Finnish Cancer Registry</b>	yes, <a href="https://aineistokatalogi.fi/catalog/studies/21085403-7be8-4f93-bf05-231518c642a0">https://aineistokatalogi.fi/catalog/studies/21085403-7be8-4f93-bf05-231518c642a0</a> . <a href="https://cancerregistry.fi/services/information-requests/">https://cancerregistry.fi/services/information-requests/</a>	Yes, <a href="https://aineistokatalogi.fi/catalog/studies/21085403-7be8-4f93-bf05-231518c642a0">https://aineistokatalogi.fi/catalog/studies/21085403-7be8-4f93-bf05-231518c642a0</a>
	<b>Finnish Social Science Data Archive</b>	Yes, a description of the used format and metadata we provide: <a href="https://www.fsd.tuni.fi/en/services/depositing-data/ddi/">https://www.fsd.tuni.fi/en/services/depositing-data/ddi/</a>	Yes, <a href="https://services.fsd.tuni.fi/index?lang=en">https://services.fsd.tuni.fi/index?lang=en</a>
	<b>FinSote</b>	No	Yes, <a href="https://aineistokatalogi.fi/catalog">https://aineistokatalogi.fi/catalog</a>
	<b>Research Services at Statistics Finland</b>	Yes we do, but some are only readily available within Statistics Finland and can be obtained only by asking separately	Yes, <a href="https://taika.stat.fi/en/">https://taika.stat.fi/en/</a>
	<b>THL Biobank</b>	We produce metadata for different datasets, as well as collect documentations from research data returned to the biobank.	Yes, <a href="https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections">https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections</a>
<b>Spain</b>	<b>Collaborative Spanish Variant Server (CSVS)</b>	Yes, and it is displayed in the CSVS documentation	Yes, <a href="https://github.com/babelomics/CSVS/wiki">https://github.com/babelomics/CSVS/wiki</a>
	<b>European Genome-phenome Archive (EGA)</b>	we collect metadata that the users (data controllers) submit to us together with the genomic files	Yes, <a href="https://ega-archive.org/studies">https://ega-archive.org/studies</a>
	<b>Plataforma de Información BIGAN</b>		No
<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	Yes	Yes, <a href="http://www2.mezizin.uni-greifswald.de/cm/fv/ship.html">http://www2.mezizin.uni-greifswald.de/cm/fv/ship.html</a>

### Unique identifier for the data and metadata

7 out of 20 data infrastructures have a unique identifier for the datasets they store and control. They have either a PubMed ID, a Uniform Resource Name (URN) or an internal ID as a unique identifier. 5 out of 20 data infrastructures have a unique identifier for their metadata. This unique identifier is either in a UUID format or a



URN. In the BBMRI directory datasets and metadata are saved using the biobank ID or the collection ID.

**Table 10:** Data infrastructure responses relating to unique identifiers

	Data infrastructure	Do you have a unique identifier for your data?	If yes, what type of unique identifier (example: DOI, PubMed ID)?	Do you have a unique identifier for your metadata (ex: uuid)?	If yes, what type of unique identifier (example: uuid)?
<b>Belgium</b>	<b>Belgian Cancer Registry</b>	Yes	ID	Yes	UUID, Increment interger
	<b>Health Examination Survey</b>	Yes		yes	UUID: cd8ec871-81a9-45a4-931d-ee41cd2e6988
	<b>Health Interview Survey</b>	Yes		yes	UUID: 79643855-6a56-4604-91f4-e92728afd54d
	<b>Genomic data registry</b>	No		This question doesn't apply to this data infrastructure	
	<b>Statbel</b>	Yes	ID	No	
<b>European</b>	<b>BBMRI-ERIC</b>	Yes	biobankID or collectionID in the Directory, plus ongoing work on EPIC PIDs	Yes	biobankID or collectionID in the Directory, plus ongoing work on EPIC PIDs
	<b>EuroBioImaging Italian MMMI Node</b>	No		No	
<b>Finland</b>	<b>Avohilmo, Register of Primary Care Visits</b>	I don't know		I don't know	
	<b>The Care Register for Social Welfare</b>	I don't know		I don't know	
	<b>Findata</b>	This doesn't apply to this data infrastructure		This doesn't apply to this data infrastructure	
	<b>FinHealth 2017 Survey</b>	No		No	
	<b>Finnish Cancer Registry</b>	This doesn't apply to this data infrastructure		I don't know	

	<b>Finnish Social Science Data Archive</b>	Yes	URN	Yes	URN
	<b>FinSote</b>	No		No	
	<b>Research Services at Statistics Finland</b>	No		No	
	<b>THL Biobank</b>	This doesn't apply to this data infrastructure	Not applicable	This doesn't apply to this data infrastructure	not applicable
<b>Spain</b>	<b>Collaborative Spanish Variant Server (CSVs)</b>	No	Data is aggregated	No	
	<b>European Genome-phenome Archive (EGA)</b>	Yes	EGA study ID or EGA dataset ID	No	
	<b>Plataforma de Información BIGAN</b>	No		No	
<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	No		No	unique identifier only provided in MDM repository

### 3.4.2. Accessibility

The table below presents in brief the accessibility mechanism in place and whether the accessibility conditions are publicly available. The table also explains whether it is possible to extract the data from the data collection, and if yes how, or whether there is a secure processing environment to analyse the data remotely and extract only the aggregated results. Moreover, we provide information on the requirement of a registration and/or legal approval prior to the data access (75% require legal approval).

Finally, interestingly this table also reveals that in 7 out of 20 data infrastructures (35%) it takes more than 3 months to access the data from the moment the researcher has applied. This timeframe is reported to highly depend on the level of aggregation that is needed, the requirement for linkage of individual level data and the need for an approval by a committee.

**Table 11:** Access conditions across the data infrastructures

	Data infrastructure	How is the data accessed (e.g. template of how to request data, access request form (link), flow chart)? Please specify or provide a URL.	Are the conditions of access published?	Is it possible to extract the data from the data infrastructure (e.g. download) or do they have to stay in the data infrastructure?	If we cannot extract the data, is there a safe space to analyse the data?	Do third party users have to register to the data infrastructure and have an account in order to access the data?	Does the requestor need a privacy and/or legal approval to access the data?	How long does it take to provide access to the requested data to the researcher after the query has been launched or the application for access has been submitted?
Belgium	Belgian Cancer Registry	Not applicable	No	Certain BCR employees can extract data from the data infrastructure. No external users can access the infrastructure.	Yes through a Secure, remote environment	<p>Depends on the type of user (internal/external).</p> <p>External users cannot access our data infrastructure. Access to data is provided via a different way, for which the external user needs to register and needs to have an account.</p> <p>Internal users need to register and have an account to access the data infrastructure.</p>	Yes	Very variable. Depends on the request, the need to link additional data sources,...

	<b>Health Examination Survey</b>	<a href="https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx">https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx</a>	yes, <a href="https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx">https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx</a>	Once given access, the requested data file is secured and transferred		Yes	Yes	Around 6 weeks, if all goes well. Longer if the request has to go through the Information Security Council, then it is variable
	<b>Health Interview Survey</b>	<a href="https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx">https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx</a>	yes, <a href="https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx">https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx</a>	Once given access, the requested data, file is secured and transferred		Yes	Yes	Around 6 weeks, if all goes well. Longer if the request has to go through the Information Security Council, then it is variable
	<b>Genomic data registry</b>	No mechanisms are in place	No	Data can currently not be extracted from the data infrastructure	No	This question doesn't apply to this data infrastructure	I don't know	
	<b>Statbel</b>	<a href="https://statbel.fgov.be/nl/over-statbel/wat-doen-we/microdata-voor-onderzoek">https://statbel.fgov.be/nl/over-statbel/wat-doen-we/microdata-voor-onderzoek</a>	Yes	No, the microdata or aggregated data is transferred in a secure manner	No	No	Yes	3 weeks
<b>European</b>	<b>BBMRI-ERIC</b>	Via BBMRI-ERIC Negotiator	Yes, Basic conditions in BBMRI-ERIC Directory - plus details are negotiated via BBMRI-ERIC Negotiator	Data retrieval possible.	No	Yes	Yes	Depends largely - typical minimum is 1 month.
	<b>EuroBioImaging Italian MMMI Node</b>		No	Yes	Yes, <a href="http://cim-xnat.unito.it/app/template/Login.vm">http://cim-xnat.unito.it/app/template/Login.vm</a>	Yes	No	days

<b>Finland</b>	<b>Avohilmo, Register of Primary Care Visits</b>	<a href="https://sampo.thl.fi/pivot/prod/fi/avopika/pikarap01/summary_kaynnitkkvko">https://sampo.thl.fi/pivot/prod/fi/avopika/pikarap01/summary_kaynnitkkvko</a>	Yes, <a href="https://thl.fi/fi/tilas/tot-ja-data/aineistot-ja-palvelut/avoin-data#Perusterveidenhuolto">https://thl.fi/fi/tilas/tot-ja-data/aineistot-ja-palvelut/avoin-data#Perusterveidenhuolto</a>	Yes	This doesn't apply to this data infrastructure, <a href="https://thl.fi/fi/tilas/tot-ja-data/aineistot-ja-palvelut/avoin-data#Perusterveidenhuolto">https://thl.fi/fi/tilas/tot-ja-data/aineistot-ja-palvelut/avoin-data#Perusterveidenhuolto</a>	No	Yes	
	<b>The Care Register for Social Welfare</b>	<a href="https://thl.fi/en/web/thlfi-en/statistics-and-data/data-and-services/data-requests-and-analytical-services">https://thl.fi/en/web/thlfi-en/statistics-and-data/data-and-services/data-requests-and-analytical-services</a>	Yes	Yes	Yes	I don't know	I don't know	
	<b>Findata</b>	Via remote access environment Kapseli	Yes, <a href="https://findata.fi/en/kapseli/">https://findata.fi/en/kapseli/</a>	Not possible	Yes, <a href="https://findata.fi/en/kapseli/">https://findata.fi/en/kapseli/</a>	Yes	Yes	Depends on the case, current median time is 68 days
	<b>FinHealth 2017 Survey</b>	<a href="https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study">https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study</a>	Yes, <a href="https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study">https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study</a>	No	No	This doesn't apply to this data infrastructure	Yes	6-12 months
	<b>Finnish Cancer Registry</b>	<a href="https://syoparekisteri.fi/palvelut/tietopyynnot/">https://syoparekisteri.fi/palvelut/tietopyynnot/</a> <a href="https://findata.fi/en/">https://findata.fi/en/</a>	Yes, Aggregated: <a href="https://syoparekisteri.fi/palvelut/tietopyynnot/">https://syoparekisteri.fi/palvelut/tietopyynnot/</a> Individual: <a href="https://findata.fi/en/">https://findata.fi/en/</a>	It is possible to extract from Findata	Yes, <a href="https://findata.fi/en/">https://findata.fi/en/</a>	No	Yes	The permission process takes multiple months. When the requester has the legal approval, 2-4 weeks to get access to the data.

	<b>Finnish Social Science Data Archive</b>	<a href="https://services.fsd.tuni.fi/index?lang=en">https://services.fsd.tuni.fi/index?lang=en</a>	Yes, <a href="https://services.fsd.tuni.fi/help?lang=en">https://services.fsd.tuni.fi/help?lang=en</a>	Customers download the data for themselves	This doesn't apply to this data infrastructure	Yes	No	For most of the cases the customer can download the requested data right away (automatic authentication and approval). If the dataset requires permission from the data depositor, it may take from a few days to a couple of weeks.
	<b>FinSote</b>	Through Findata	This doesn't apply to this data infrastructure	No	No	This doesn't apply to this data infrastructure	Yes	6-12 months
	<b>Research Services at Statistics Finland</b>	<a href="https://www2.tilastokeskus.fi/tup/mikroaineistot/ohjeita_uttukijalle_en.html">https://www2.tilastokeskus.fi/tup/mikroaineistot/ohjeita_uttukijalle_en.html</a> <a href="https://www2.tilastokeskus.fi/sivusto/loomakkeet/index_en.html">https://www2.tilastokeskus.fi/sivusto/loomakkeet/index_en.html</a>	Yes, <a href="https://www2.tilastokeskus.fi/tup/mikroaineistot/ohjeita_uttukijalle_en.html">https://www2.tilastokeskus.fi/tup/mikroaineistot/ohjeita_uttukijalle_en.html</a>	The data has to be handled over a remote access system. Researchers can download aggregated data and results from the remote access system	Yes, <a href="https://www2.tilastokeskus.fi/tup/mikroaineistot/etakeytt_o_en.html">https://www2.tilastokeskus.fi/tup/mikroaineistot/etakeytt_o_en.html</a>	Yes	Yes	Depending on the type of data , 1 - 6 months
	<b>THL Biobank</b>	<a href="https://thl.fi/en/web/thl-biobank/for-researchers/application-process">https://thl.fi/en/web/thl-biobank/for-researchers/application-process</a>	Yes, <a href="https://thl.fi/en/web/thl-biobank/for-researchers/application-process/principles-of-access">https://thl.fi/en/web/thl-biobank/for-researchers/application-process/principles-of-access</a>	A copy of the specific data is provided to researchers with approved research application and signed MTA	This doesn't apply to this data infrastructure	This doesn't apply to this data infrastructure	Yes	Depending on many factors, because access requires approval of application and signed MTA.
<b>Spain</b>	<b>Collaborative Spanish Variant Server (CSVs)</b>	<a href="http://csvs.babelomics.org/">http://csvs.babelomics.org/</a>	Yes, <a href="https://github.com/babelomics/CSVs/wiki">https://github.com/babelomics/CSVs/wiki</a>	Stay in the infrastructure. There is a matchmaking service.	No	No	No	

	<b>European Genome-phenome Archive (EGA)</b>	<a href="https://ega-archive.org/access/data-access">https://ega-archive.org/access/data-access</a>	Yes, <a href="https://ega-archive.org/access/data-access">https://ega-archive.org/access/data-access</a>	they can be downloaded		Yes	Yes	roughly 2 months but it greatly depend on the the specific data controllers
	<b>Plataforma de Información BIGAN</b>	Access Request Form ( <a href="https://www.iacs.es/instituto-aragones-ciencias-la-salud/oficina-virtual/solicitud-de-acceso-a-datos-para-realizacion-de-un-proyecto-de-investigacion-rpi01-3a/">https://www.iacs.es/instituto-aragones-ciencias-la-salud/oficina-virtual/solicitud-de-acceso-a-datos-para-realizacion-de-un-proyecto-de-investigacion-rpi01-3a/</a> )	Yes	Download available		No	Yes	
<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	<a href="http://www2.medizin.uni-greifswald.de/cm/fv/ship.html">http://www2.medizin.uni-greifswald.de/cm/fv/ship.html</a>	Yes, <a href="http://www2.medizin.uni-greifswald.de/cm/fv/ship.html">http://www2.medizin.uni-greifswald.de/cm/fv/ship.html</a>	yes		Yes	Yes	1-4 months, depends on contract issues

### 3.4.3. *Interoperability*

One of the most important factors to link individual level data or datasets across different member states is interoperability. This can be affected by the format in which datasets have been stored in, the semantic interoperability standards used, such as ICD11 or SNOMED CT, the common data model used to describe them, such as OMOP, or the standard used to transfer data, such as HL7 FHIR.

#### **Format of the data**

From the survey results we can conclude that there is a wide variety of data formats used across data infrastructures depending on the kind of data. For example, the Health Interview Survey results are stored in files using plain text, medical images are stored using the DICOM standards and other health data are stored in either JSON, XML or FASTA.

#### **Semantic interoperability and data exchange standards**

Table 12 below presents the standards used by the different data collections to structure their data and metadata. 9/20 data collections use the same ICD-10 semantic interoperability standard to structure their data and only 3/20 use SNOMED-CT. Some of these data infrastructures use nationally developed standards.

The data exchange standard HL7-FHIR is used only by 2 data infrastructures.

#### **Format for distributing the data**

Most of the data infrastructures (14/20) distribute health data in csv files. Data is also distributed in R, SAS, SPSS, DICOM, PDF, JSON and Stata file formats.

This lack of interoperability observed between these data infrastructures might cause a challenge to a research project that aims at linking individual level data between these data collections, e.g. the cancer use case research question.



**Table 12:** Data infrastructure responses to interoperability questions

	Data infrastructure	Which community-recognised vocabularies, standards or methodologies are used for metadata and data to facilitate interoperability?	What is the format(s) for distributing data?	Do you have a metadata record API endpoint (m2m) in place?	What is the format in which the data is stored?
Belgium	Belgian Cancer Registry	/ ICD-10 / ICD-O-3 / TNM	/ csv / R / SAS	No	/ Data is encrypted when stored. / Plain text / XML / JSON / Files / Other
	Health Examination Survey	Other: look at the codebook (NACE, ISCO)	Any format that is requested	yes	Plain text
	Health Interview Survey	Other: look at the codebook (NACE, ISCO)	Any format that is requested	yes	Plain text
	Genomic data registry	/ This doesn't apply to this data infrastructure	/ FASTQ, BAM, VCF - but no data exchange policies are in place	/ No	/ Other: FASTQ, BAM, VCF
	Statbel	/ This doesn't apply to this data infrastructure	/ csv / json / SAS	No	/ Files
European	BBMRI-ERIC	/ HL7 FHIR / SNOMED CT / LOINC / ICD-10 / OMOP, CDISC	/ csv / xml / json / Id-json / DB dumps, and other formats possible.	Yes	/ Plain text / FASTA / XML / RDF / tsv / JSON / DICOM / Files / Other: This is really heterogeneous for different cases.
	EuroBioImaging Italian MMMI Node	/ I don't know	DICOM	No	/ XML / DICOM / Other
Finland	Avohilmo, Register of Primary Care Visits	/ HL7 / LOINC / ICD-10 / ICPC-2, THL-Toimenpide	/ csv / xml / json / pdf / R / SAS	I don't know	/ JSON
	The Care Register for Social Welfare				I don't know

	<b>Findata</b>	/ SNOMED-CT / ICD-10	/ csv / xml / pdf / R / SAS	I don't know	Files
	<b>FinHealth 2017 Survey</b>	This doesn't apply to this data infrastructure	/ csv / R / SAS / SPSS/ Stata	Yes	/ Files
	<b>Finnish Cancer Registry</b>	/ ICD-10 / ICD-O-3	/ csv / pdf / R / SAS / xls, dat, txt	I don't know	I don't know
	<b>Finnish Social Science Data Archive</b>	DDI, CESSDA Vocabularies, YSO/Finto (General Finnish Ontology). <a href="https://www.fsd.tuni.fi/en/services/data-management-guidelines/examples-and-vocabularies/">https://www.fsd.tuni.fi/en/services/data-management-guidelines/examples-and-vocabularies/</a>	/ csv / PDF / por, odt, txt, html ( <a href="https://www.fsd.tuni.fi/en/data-archive/documents/records-management-and-archives-formation-plan/file-formats/">https://www.fsd.tuni.fi/en/data-archive/documents/records-management-and-archives-formation-plan/file-formats/</a> )	Yes	Files
	<b>FinSote</b>	This doesn't apply to this data infrastructure	/ csv / R	Yes	/ Files
	<b>Research Services at Statistics Finland</b>	/ ICD-10 / Standard classifications for education (ISCED), occupation (ISCO), etc.	/ csv / SAS	I don't know	Files
	<b>THL Biobank</b>	This doesn't apply to this data infrastructure	csv; different datasets are in different data format. Format is provided as the researchers request.	I don't know	/ Plain text / FASTA / Files / Other  Many different types of data formats
<b>Spain</b>	<b>Collaborative Spanish Variant Server (CSVS)</b>	ICD-10	csv	No	/ FASTA / Other: Indexed in OpenCGA
	<b>European Genome-phenome Archive (EGA)</b>			Yes	
	<b>Plataforma de Información BIGAN</b>	/ SNOMED CT / LOINC / ICD-10 / ICD-9; ICPC; DICOM; ATC	csv	No	/ Plain text / FASTA / tsv / JSON / DICOM / Parquet / Files

<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	ICD-10 maelstrom Taxonomy, UMLS	/ csv / xml / SAS /R, Stata, SPSS	No	/ Plain text / JSON / DICOM / Other: you did not mention data types within PostGreQL
----------------	--	---------------------------------------	--	----	---

### 3.4.1. Re-usability

Secondary use of health data would only be possible if the data infrastructures allow the re-use of data they control.

The table below reports which data infrastructures allow external users to access the data and re-use it for more than one purpose and whether there is a clear procedure to request the re-use of the data. According to the results of the survey, in 12/20 (60%) of the data infrastructures, data can be re-used by external users either for a single or multiple projects. Respondents also reported the procedures that external, third party users, need to follow to request the reuse of the datasets controlled by these data infrastructures.

Finally, we asked whether the information they provided in this survey has already been placed in an open access source. Around half of the respondents replied that this information has already been placed online and they provided the URL to this online location.

**Table 13:** Data infrastructure responses to questions on re-use of data

	Data infrastructure	Is it possible for third party users to access the data and re-use it for more than one purpose/project?	Is there a clear procedure for third party users to request (the licence) for data re-use?	Have you placed the metadata related to your data infrastructure (that is, the above information provided in this survey) in another available source already?
<b>Belgium</b>	<b>Belgian Cancer Registry</b>	Yes, Third party users don't access the data infrastructure, but a copy of the data in our infrastructure can be made available via a standard operating procedure.	Yes	No
	<b>Health Examination Survey</b>	No	/ This doesn't apply to this data infrastructure	No
	<b>Health Interview Survey</b>	No	/ This doesn't apply to this data infrastructure	No
	<b>Genomic data registry</b>	/No	/No	No

	<b>Statbel</b>	No	Yes, <a href="https://statbel.fgov.be/nl/over-statbel/wat-doen-we/microdata-voor-onderzoekhe-procedure">https://statbel.fgov.be/nl/over-statbel/wat-doen-we/microdata-voor-onderzoekhe-procedure</a>	Yes, <a href="https://statbel.fgov.be/nl/themas/bevolking/sterfte-en-levensverwachting/sterfte#documents">https://statbel.fgov.be/nl/themas/bevolking/sterfte-en-levensverwachting/sterfte#documents</a>
<b>European</b>	<b>BBMRI-ERIC</b>	Yes	Data is typically accessed on DTA/MTA basis. So not a license, but another type of contract.	Yes, BBMRI-ERIC Directory
	<b>EuroBioImaging Italian MMMI Node</b>	Yes	I don't know	No
<b>Finland</b>	<b>Avohilmo, Register of Primary Care Visits</b>	Yes	I don't know	Yes, ELIXIR-ES
	<b>The Care Register for Social Welfare</b>	Yes	Yes	No
	<b>Findata</b>	No	Yes, Data permit is study-specific; to use the data for other purposes you need another application	Yes, Aineistoeditori
	<b>FinHealth 2017 Survey</b>	Yes	Yes	Yes, <a href="https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study">https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study</a>
	<b>Finnish Cancer Registry</b>	No	No	No
	<b>Finnish Social Science Data Archive</b>	Yes	Yes, see before from "data access"	Yes, <a href="https://www.coretrustseal.org/wp-content/uploads/2020/11/Finnish-Social-Science-Data-Archive.pdf">https://www.coretrustseal.org/wp-content/uploads/2020/11/Finnish-Social-Science-Data-Archive.pdf</a> ; <a href="https://www.fsd.tuni.fi/en/">https://www.fsd.tuni.fi/en/</a> ; <a href="https://www.fsd.tuni.fi/en/data-archive/documents/records-management-and-archives-formation-plan/">https://www.fsd.tuni.fi/en/data-archive/documents/records-management-and-archives-formation-plan/</a>
	<b>FinSote</b>	Yes	Yes	Yes, Aineistokatalogi <a href="https://aineistokatalogi.fi/catalog/studies/76c9e6e8-e3ce-469d-bae4-dc6e8abe2ca6">https://aineistokatalogi.fi/catalog/studies/76c9e6e8-e3ce-469d-bae4-dc6e8abe2ca6</a>
	<b>Research Services at Statistics Finland</b>	Yes	Yes, Apply for a new licence or changes to existing licence	Yes, Metadata Catalogue Taika <a href="https://taika.stat.fi/en/">https://taika.stat.fi/en/</a> and Data Resource Catalogue <a href="https://aineistokatalogi.fi/catalog">https://aineistokatalogi.fi/catalog</a>

	<b>THL Biobank</b>	Yes	Yes, New data generated in biobank projects must be returned to the biobank and can be provided to other researchers.	Yes, The metadata is available through different catalogs, in addition to the biobank's own webpages
<b>Spain</b>	<b>Collaborative Spanish Variant Server (CSVs)</b>	No	No	Yes
	<b>European Genome-phenome Archive (EGA)</b>	Yes	Yes, the requester ask to the EGA. we move the request to the DAC and manage the needed documentation, like DAA. when done, we open the data for the requester.	No
	<b>Plataforma de Información BIGAN</b>	No	This doesn't apply for this data infrastructure	No
<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	Yes	Yes	Yes, e.g. Maelstrom, euCanShare

### 3.5. FAIRness evaluation of the data infrastructures

To evaluate the compliance of these data infrastructures we adapted an already existing FAIRness evaluation tool (the ARDC tool; see the *HealthyCloud FAIRness self-assessment tool* in section 3.4 in this document) to fit with the exact questions we asked in the survey. We then asked the partners of WP3 to use this online tool to evaluate the different data infrastructures in order to provide a score for their compliance with the FAIR principles. The graphs below present the results of this evaluation and, more specifically, the overall FAIRness score. As we can observe, most data infrastructures have a FAIRness score above 50%, which is promising for the execution of the use cases.



In order to better overall FAIRness observe the areas improvement in infrastructure in

understand the score and that would need each data order to become

more FAIR, we have added below the tables with the detailed score for each letter of the FAIR principles.

**Table 14:** FAIR assessment of Finnish data infrastructures surveyed

	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits
<b>F</b>	0%	41%	41%	100%	41%	47%	23%	41%	23%
<b>A</b>	50%	80%	90%	80%	60%	70%	40%	70%	60%
<b>I</b>	50%	0%	62%	62%	50%	62%	75%	62%	62%

<b>R</b>	100%	71%	100%	100%	100%	100%	100%	0%	57%
<b>FAIRness evaluation (total%)</b>	66%	49%	73%	85%	62%	69%	59%	43%	50%

**Table 15:** FAIRness assessment of Belgian data infrastructures surveyed

	<b>Health Interview Survey</b>	<b>Health Examination Survey</b>	<b>Belgian Cancer Registry</b>	<b>Genomic data registry</b>	<b>Statbel</b>
<b>F</b>	100%	100%	70%	0%	29%
<b>A</b>	70%	70%	10%	0%	60%
<b>I</b>	37%	37%	62%	50%	50%
<b>R</b>	0%	0%	71%	0%	71%
<b>FAIRness evaluation (total%)</b>	51%	51%	53%	12%	52%

**Table 16:** FAIRness assessment of Spanish data infrastructures surveyed

	<b>Plataforma de Información BIGAN</b>	<b>Collaborative Spanish Variant Server (CSVS)</b>	<b>European Genome-phenome Archive (EGA)</b>
<b>F</b>	0%	41%	70%
<b>A</b>	70%	40%	40%
<b>I</b>	62%	62%	25%
<b>R</b>	0%	28%	71%
<b>FAIRness evaluation (total%)</b>	33%	43%	51%

**Table 17:** FAIRness assessment of European Research Infrastructures surveyed

	<b>Eurobioimaging Italian MMMI Node</b>	<b>BBMRI-ERIC</b>
<b>F</b>	17%	100%
<b>A</b>	70%	70%
<b>I</b>	50%	87%
<b>R</b>	28%	100%
<b>FAIRness evaluation (total%)</b>	41%	89%

**Table 18:** FAIRness assessment of the German data infrastructure surveyed (atrial fibrillation use case)

	SHIP
<b>F</b>	41%
<b>A</b>	70%
<b>I</b>	62%
<b>R</b>	100%
<b>FAIRness evaluation (total%)</b>	68%

This adapted FAIRness evaluation tool is now available and ready to use in a Binder environment and any data controller can use it to evaluate how FAIR their data collection is. The link to access the tool is: <https://ovh.mybinder.org/v2/gh/PderyckeSciensano/HEALTHYCLOUD/main?urlpat h=rstudio>.

To use the online FAIRness evaluation tool please follow the stepwise approach presented below:

Step by step guide for users in myBinder/Rstudio:

- Open and knit the FAIR\_TOOL.rmd Rmd notebook
- Input your answers for each question in the tool
- Click 'Download' and save the csv file
- Upload the csv file in the Rstudio environment
- Check and edit if needed the path and name of the csv file in the rmd code (line 39)
- Re-knit the Rmd notebook
- This creates a FAIRness report, including pie charts demonstrating the percentage scores for each principle as well as an overall score.
- Upload and share the FAIRness "FAIR\_TOOL.html" report.

There is an option to select 'I don't know' under each question. However, we would encourage users to consider if there is someone else in their organisation who does know the answer to that question, to increase the accuracy of the assessment. You would need to share the FAIRness report within your organisation. At each updating step, a new csv file can be produced and used to generate an updated FAIRness report.

[\*] The tool can be downloaded on [ZENODO.org](https://zenodo.org/doi/10.5281/zenodo.7038397) (<https://doi.org/10.5281/zenodo.7038397>) (R studio is required) or accessed on [MyBinder](https://ovh.mybinder.org/v2/gh/PderyckeSciensano/HEALTHYCLOUD/main?urlpath=rstudio) (no installation required).



## 4. Discussion

### Main conclusions of the analysis

This document presents the analysis of the WP3 survey in relation to the data infrastructures relevant to the cancer use case from Finland, Belgium, Spain and the two European research infrastructures, namely BBMRI-ERIC and the Eurobioimaging (Italian MMMI Node). Moreover, we also analysed the results from one of the data sources used to answer the atrial fibrillation use case, namely the State of Health in Pomerania (SHIP).

The areas analysed were: type of data source, level of aggregation, anonymisation and pseudonymisation methods, geographical and time coverage, ethical requirements for data storage, data quality controls of the various data infrastructures and, finally, the compliance with the FAIR principles.

The findability of these datasets by a potential user is relatively high as 65% of the data infrastructures produce or collect metadata for the datasets they are storing or are data controllers of and 70% responded that they also have a public metadata catalogue service available where a researcher can find information about their data collection.

The facts that most of the data infrastructures store individual-level data, have pseudonymised data, and have national-level coverage provide good chances to successfully link individual level data. However, two of the data infrastructures - the Plataforma de Información BIGAN for the cancer use case, and the State of Health in Pomerania (SHIP) for the atrial fibrillation use case - only have regional-level data. This could reduce the feasibility of linking individual level data.

Another finding that hampers individual-level data linkage is the fact that the Spanish Combined Variant Server (CSVS) (relevant to the cancer use case) collects already aggregated data and does not have individual-level data, reducing the ability to link with data from this data infrastructure. Promisingly, it is the only data infrastructure that does not have individual-level data.

In addition, whilst most of the data infrastructures store pseudonymised data, there are two infrastructures relevant to the cancer use case that anonymise the data at the point of collection: the Avohilmo Register of Primary Care Visits in Finland and the Collaborative Spanish Variant Server (CSVS) in Spain, reducing the feasibility of linkage. In addition, almost a third of the data infrastructures anonymise data before sharing it externally which highlights the importance of and compliance with privacy preservation of sensitive personal data.

Another finding that may hamper data linkage at individual level between different data collections is the lack of interoperability due to the usage of different standards to structure their data or metadata. This could be mitigated by using crosswalks,

building tables that translate the code for each term and reference in the different standards used.

After WP7 completes the cancer use case research project in at least one of these member states, the output and feedback from the researchers would determine the minimum level of FAIRness needed in order to use these data infrastructures.

### **Limitations**

As described above on the cancer use case, at the time of writing we had received very few responses from the Spanish data infrastructures and none from the German data infrastructures. This is possibly due to both of these countries having a decentralised federated organisation and thus complicating the identification and contact of the right data providers to conduct the research study. Moreover, there seems to be a lack of a common metadata catalogue that would compile all available data collections in the country. With the cancer use case team we have liaised with German partners from Charité and TMF to discuss the relevant data infrastructures in Germany and we hope to have their information for the Deliverable 3.3 (due by April 2023).

Similarly, unfortunately, only one response was received from the data infrastructures relevant for the atrial fibrillation use case. The findings from this data infrastructure are incorporated into this deliverable. If further responses are received, they will be incorporated into Deliverable 3.3.

In addition, EGA was one of the data infrastructures which piloted the WP3 survey prior to its finalisation. This means that there are several questions where we do not have answers from EGA, as these questions were adapted or added after the feedback from the piloting phase.

## 5. Conclusion and next steps

In conclusion, this deliverable presents the analysis of the results of the WP3 survey. The survey results are being analysed to perform a FAIRness evaluation of the data infrastructures that have been selected for the scope of the use cases and also to answer the question of feasibility of linking individual level data. This landscape analysis will expand, and more data collections will be added and analysed in the Deliverable 3.3.

### *Next steps*

The survey and the results will be transformed into a digital notebook and a catalogue matrix that will be publicly available online, more user friendly and queryable. This will allow the expansion of this study and add more data collections, share the FAIRness evaluation of the European health related data collections we are exploring and create a source where researchers can access and find more information on the data collections they would be interested to use.

The aim is to further use these findings to start building an online, publicly available metadata catalogue of health data infrastructures with their key description and information. This will also populate the portal that is being designed by WP6. The descriptive metadata template used to inventorise these data collections will be based on the DCAT-AP standard template that will be presented in Deliverable 3.2 and optionally on the health DCAT-AP extension if this is made available publicly by then.

## Annex 1

HealthyCloud			
ID	INDICATORS	Description of the indicator (example)	Format of the input
<b>Part 1: Data</b>			
Administrative	Title	Title or name of the data infrastructure (data collection or data hub)	Free text
	Abbreviation or alternative title	Abbreviation or alternative title	Free text
	Website	Website of the data infrastructure (collection or hub)	URL
	Data controller	Who is the data controller organisation?	Free text
	Data controller	Contact details (full name and email address of the data controller)	Free text
	Contact details of the data access provider (Provides the availability of data, through a metadata catalogue)	Full name of the contact person	Free text
		Email address	Free text
		URL	URL
	Data processor	Who is the data processor organisation, if any?	Free text
	Data hub	Which of the following characteristics fit your data infrastructure?	Multiple choice: / A digital platform that receives and stores data / It receives data from a single source and/or multiple sources / It has control over the data stored / It has a specific thematic, data type that it collects (e.g. a particular disease, a particular data type: genomic data, clinical data, EHRs...) / It is part of one or more overarching data hubs / It generates data / A digital technical infrastructure with the core mission of enabling health data sharing / It provides health data from different sources / It allows discovery of health datasets / It has a metadata discovery service / It has a data accessibility mechanism in accordance with

			existing regulation / It has an authorization functionality, provided by the same Data Hub or by an external institution
		If your data infrastructure is part of a data hub, what is the name and URL of the data hub?	Name and URL of data hub
		How is the data infrastructure organised?	Drop down menu: / It is managed centrally / It is a decentralised management / I don't know / This doesn't apply to this data infrastructure / Other
Data	Data storage	Do you require ethical approval for the data to be stored in your infrastructure?	Yes No I don't know This doesn't apply to this infrastructure
	Type of source	Does the data originate from a patient group, the general population or an experimental setting, or other?	Drop down menu: / Patient group / General population / Experimental setting / Other / I don't know / This doesn't apply to this data infrastructure  If 'Other', please specify.
		What is the type of data source that you are using? You can choose multiple options.	Multiple choice: / Electronic health records (EHR) / Clinical trials / Survey / Cohorts / Biobanks (biological samples) / Picture Archiving and Communication System (PACS) / Imaging data / Medical devices / Clinical Research data / Genomic data (Whole Genome sequencing / Whole exome sequencing / targeted sequencing / epigenetic-sensitive sequencing / other genomic data) / Biometric data / Molecular data / Socioeconomic data / Specific disease data / Survival data / Population health data / Interview data / Administrative data

			/ Registry data / Customer record data / Observational study data / Healthcare data (Prescriptions / Diagnoses / Laboratory data/ Treatment / Surgery/ Other) / Other (can choose multiple options) If 'Other', please specify
	Data compilation methods	How is the data that is stored in the data infrastructure compiled?	Multiple choice: / Data retrieval / Parsing / Transforming / Loading / ETL methods / Other / I don't know / This doesn't apply to this data infrastructure If 'Other', please specify.
	Technologies used for data storage	Describe the technologies used for data storage. E.g. relational database (SQL, ...), NoSQL (..), Graph db	Free text
	Data format	What is the format in which the data is stored?	Multiple choice: / Plain text / FASTA / XML / RDF / Dublin Core / tsv / JSON / DICOM / Parquet / Files / Other / I don't know / This doesn't apply to this data infrastructure If 'Other', please specify
	Type of data	Specify the type of data collected	Multiple choice: / Images / Text / Numbers / Files / Tissue samples / Sounds / Multidimensional array / Spreadsheet / Other (please specify)
	Level of aggregation	What is the level of aggregation of the data stored in this data infrastructure? e.g. aggregated, individual, both	Drop down menu: / Individual / Aggregated / Both / I don't know

			/ This question doesn't apply to this data infrastructure
	Anonymisation	Are anonymisation methods used with the data?	Drop down menu: / Yes: at the point of collection / Yes: before sharing them externally / Yes: before sharing them internally / Yes: at the point of publishing / No: we do not anonymise data / I don't know / This question doesn't apply to this data infrastructure
		Is the anonymisation performed by your data infrastructure and/or do you receive already anonymised data?	Drop down menu: / We perform the anonymisation / We receive anonymised data / Both
	Pseudonymisation	Do you have pseudonymised data?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		If yes, who (name of the organisation or stakeholder) holds the method to reverse the pseudonymisation process? (e.g. key, dictionary, map, table)	Free text
Completeness of data infrastructure	Geographical coverage	What is the geographical coverage of the data infrastructure (datasets registered in your data collection, or data collections registered/linked in your data hub)?	Multiple choice: / International / European / National / Regional / I don't know / This question doesn't apply to my data infrastructure
		What is the socioeconomic coverage of the data infrastructure (datasets registered in your data collection, or data collections registered/linked in your data hub)?  NB: The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU and the UK for the purpose of collection, development and harmonisation of European regional statistics. - NUTS 1: major socio-economic regions	Multiple choice: / NUTS1 / NUTS2 / NUTS3 / I don't know

		- NUTS 2: basic regions for the application of regional policies - NUTS 3: small regions for specific diagnoses	
	Participating countries	What are the participating countries from which you have datasets?	Free text
	Data collection start date	When did your data infrastructure start collecting data? If this applies to your data infrastructure.	Free text
	Data collection period	Is the data collection period still ongoing? If this applies to your data infrastructure.	Yes/No
	Data collection end date	What is the end date of the data collection period? If this applies to your data infrastructure.	Free text
Data quality aspects	Data quality control	Are data quality controls applied?	Drop down menu: / Yes / No / I don't know
		Are there minimum levels of quality of the data (results from quality controls) needed for the data to be included in the data infrastructure?	Drop down menu: / Yes, data is only included if it reaches a certain quality level / No, we do quality control for internal use only / No, but the results of the quality control are available when searching for the data / Does not apply / Unknown
	Updating periodicity	How often do you update the datasets ?	Multiple choice: / Weekly / Monthly / Annually / Biannually / Every 2+ years / Every 5+ years / Irregularly / One time collection / I don't know / This doesn't apply to this data infrastructure
	Error checking	Do you use a tool to check for errors and completeness (e.g., Checksum tool)?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		If yes, what tool do you use (e.g., Checksum)	Free text
	Versioning of datasets	Do you have a process to keep track of the different versions of the datasets?	Drop down menu: / Yes / No



			/ I don't know / This question doesn't apply to this data infrastructure
		If yes, please specify the process.	Free text
	Data source legitimacy	Do you have a method to check data source legitimacy (e.g. ISO standard on data quality)? Please specify.	Free text
Metadata	Metadata related to data infrastructure	Have you placed the metadata related to your data infrastructure (that is, the above information provided in this survey) in another available source already?	Drop down menu: / Yes / No / I don't know
		If yes, where is it?	Free text or URL
	Metadata related to data	Do you produce or collect metadata for all your data (e.g. handbook, guide for users, description, keywords, timestamp, spatial coverage etc.)? Please specify.	Free text
	Metadata catalogue		Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		Do you have a public metadata catalogue service?	If yes, what is the URL?
Findable	Unique identifier for data	Do you have a unique identifier for your data ?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		If yes, what type of unique identifier (example: DOI, PubMed ID)?	Free text
	Unique identifier for metadata	Do you have a unique identifier for your metadata (ex: uuid)?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		If yes, what type of unique identifier (example: uuid)?	Free text
	Data catalogue		Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		Do you have a public data catalogue?	If yes what is the URL?

	Technical solution	What type of search engine do you use (e.g. proprietary or open source solution)?	Drop down menu: / Proprietary / Open source / I don't know / This doesn't apply to this data infrastructure
Accessible	Data access	Do you provide access to individual and/or aggregated data (for third party users)?	Multiple choice: / Individual / Aggregated / I don't know / This doesn't apply to this data infrastructure
		How is the data accessed (e.g. template of how to request data, access request form (link), flow chart)? Please specify or provide a URL.	Free text or URL
		Are the conditions of access published?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure If yes, please provide the URL.
		Is it possible to extract the data from the data infrastructure (e.g. download) or do they have to stay in the data infrastructure?	Free text
		If we cannot extract the data, is there a safe space to analyse the data?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure If yes, please provide the URL of the safe space to analyse data
	Registration	Do third party users have to register to the data infrastructure and have an account in order to access the data?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
	Encryption	Does the data infrastructure encrypt the data?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		Is the data encrypted when stored or only when transferred?	Multiple choice: / Encrypted when stored / Encrypted when transferred
		How is the data encrypted? Please specify the encryption protocol.	Free text

	Legal approval	Does the requestor need a privacy and/or legal approval to access the data?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		How long does it take to provide access to the requested data to the researcher after the query has been launched or the application for access has been submitted?	Free text
Interoperability	Standards used for metadata and data	Which community-recognised vocabularies, standards or methodologies are used for metadata and data to facilitate interoperability?	Multiple choice: / HL7 / FHIR / SNOMED CT / LOINC / ICD-10 / Other / I don't know / This doesn't apply to this data infrastructure
		If other, please specify	Free text
	Data format for exchange		Multiple choice: / csv / xml / json / Id-json / pdf / R / SAS / Other / I don't know / This doesn't apply to this data infrastructure
		What is the format(s) for distributing data?	Free text
	Metadata record		Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		Do you have a metadata record API endpoint (m2m) in place?	
Re-usable	Data re-use	Is it possible for third party users to access the data and re-use it for more than one purpose/project?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
		Is there a clear procedure for third party users to request (the license) for data re-use?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure

			If Yes, please specify the procedure
	Legal officer	Do you have a legal officer/data owner contact?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure If yes, please provide the full name and email address of the person
	Ethical and legal approval for re-use of data	Does the requestor need ethical approval for the secondary use of health data?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure If Yes, please specify the procedure
		Does the requestor need privacy and/or legal approval for secondary use of health data? e.g. ensuring that the patient cannot be identified	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure If Yes, please specify the procedure
<b>Part 2: Governance / Management / data hub specific questions</b>		<b>Please, could you answer the following questions if it is applicable to your case?</b>	
Technical	Size	How much storage capacity is in use up to date?	number
		Until today, how many datasets are stored in your data collection, or studies/data collections stored in your data hub?	number
	Estimated annual growth	What is the estimated annual growth of the data infrastructure (repository or hub) in size or number of datasets?	number
	Data infrastructure Users	Number of sustained users who submit or store data up to date	number
		Number of sustained users who access data up to date	number
Legal aspects	GDPR compliance	Are there any national rules additional to the GDPR in your country? If yes, which ones?	Names and/or links to the laws and regulations that include aspects that are not developed in the GDPR at the regional and national level
		In the scope of the EU GDPR, what is your organisation's role in relation to personal data?	add an option we have different roles in different situations, select multiple options

		i.e. Data controller/Joint controller/Data processor/None of the above	If 'None of the above', please specify.
		Please, describe the logging and auditing of user actions	record of user deposition date and time / record of user contact to client service / record of user application for data use / none of the above/ others / this does not apply to my organization
		Does the data hub provide a DAA (Data Access Agreement) to be signed between data providers and data requesters?	No / Yes, data hub has a non-negotiable DAA form / Yes, data hub provides a DAA template which may be modified under agreement / Other
			If 'Other', please specify
		Does the data hub have a DPA (Data Processor Agreement) to be signed with the Data providers?	No / Yes, data hub has a non-negotiable DAA form / Yes, data hub provides a DAA template which may be modified under agreement / Other
			If 'Other', please specify
		Does the data hub have a DPIA (Data Protection Impact Assessment) model?	Yes / No
	Has access control mechanism been implemented (authentication and authorization)?	no/ OAuth2 / OpenID Connect (over HTTPs) / Authorization over SSH / Authorization with Web services backed by a database / Authorization via (web) Rest API / Authorization (read) over AMQPs / others	
	Sustainability	What is the sustainability plan of the data hub funding?	free text (i.e.stable national or international funding/applying to european infrastructure funding/applying to competitive plans)
	Governance	Does the data hub provide a catalogue of different data sources?	Yes / No, the data hub is connected only to an unique data source
From the perspective of where is the data stored. Does the data hub receive data from different sources?		Yes, data is sent to the data hub and stored there (centralised) /No, data stay only at original place and it is linked at the data hub (federated)	
Please, describe the services through which data is shared e.g. website, APIs, FTP			
Operational	Others	Do you have established standard operating procedures (SOPs) that your organization follows and updates regularly?	yes/No
		Other comments	free text



## Annex 2

**Table 1:** Data controller and administrative information

	Data infrastructure	Data controller	Contact details of data controller	Data access provider (Provides the availability of data, through a metadata catalogue):	Contact details of data access provider: Email	Contact details of data access provider: URL	Data processor
<b>Belgium</b>	<b>Belgian Cancer Registry</b>	Belgian Cancer Registry	Belgian Cancer Registry info@kankerregister.org	Not applicable	Not applicable	Not applicable	Axians (Fit IT nv) admin.be@axians.com www.axians.be
	<b>Health Examination Survey</b>	Data Protection Officer	Melissa van Bossuyt	Stefaan Demarest	<a href="mailto:stefaan.demarest@sciensano.be">stefaan.demarest@sciensano.be</a>	<a href="http://www.sciensano.be">www.sciensano.be</a>	Sciensano
	<b>Health Interview Survey</b>	Data Protection Officer	Melissa van Bossuyt	Stefaan Demarest	<a href="mailto:stefaan.demarest@sciensano.be">stefaan.demarest@sciensano.be</a>	<a href="http://www.sciensano.be">www.sciensano.be</a>	Sciensano
	<b>Genomic data registry</b>	Sciensano	Not yet defined, Marc Van den Bulcke or Karin de Ridder	Not yet defined	Not yet defined	Not yet defined	Not yet defined
	<b>Statbel</b>	Statbel, represented by the Director General	Statbel, Koning Albert II-laan 16 - 1000 Brussel. Directeur-generaal a.i. Philippe.Mauroy@economie.fgov.be	Gisele Vandervelpen	Gisele.Vandervelpen@economie.fgov.be	<a href="http://www.Statbel.fgov.Be">www.Statbel.fgov.Be</a>	Not applicable
<b>European</b>	<b>BBMRI-ERIC</b>	BBMRI-ERIC or a data source (depends on the situation, the question is not unambiguous for us)	BBMRI-ERIC, Neue Stiftingtalstraße 2/B/6, 8010 Graz, AT contact@bbmri-eric.eu (please note that data controller is *the institution* and not any specific person .... the institution has a DPO, but you are not asking for that)	We have 600+ of those, plus BBMRI-ERIC itself (either as data controller or as facilitator). In case it's BBMRI-ERIC, we have institutional mechanisms for negotiating access (via BBMRI-ERIC Negotiator) and not a single person. Hence this question is not clear to us.			
	<b>EuroBioImaging Italian MMMI Node</b>	University of Torino - Molecular Imaging Center	Alessandra Viale (alessandra.viale@unito.it)				Molecular Imaging Center

Finland	<b>Avohilmo, Register of Primary Care Visits</b>	Finnish institute for health and welfare	avohilmo@thl.fi	Kaisa Mölläri	avohilmo@thl.fi		
	<b>The Care Register for Social Welfare</b>	The Finnish Institute for Health and Welfare (THL)	Riikka Väyrynen riikka.vayrynen(at)thl.fi	Data requests and analytical services	tietopyynnot(at)thl.fi	<a href="https://thl.fi/en/web/thlfi-en/statistics-and-data/data-and-services/data-requests-and-analytical-services">https://thl.fi/en/web/thlfi-en/statistics-and-data/data-and-services/data-requests-and-analytical-services</a>	THL
	<b>Findata</b>	Social care and health care providers, national registries, Findata	info@findata.fi		info@findata.fi		Findata
	<b>FinHealth 2017 Survey</b>	Finnish Institute for Health and Welfare (THL)				Seppo Koskinen, seppo.koskinen@thl.fi	
	<b>Finnish Cancer Registry</b>	Finnish Institute for Health and Welfare	kirjaamo@thl.fi	Elli Hirvonen	kirjaamo@cancer.fi	<a href="https://cancerregistry.fi/services/information-requests/">https://cancerregistry.fi/services/information-requests/</a>	Cancer Society of Finland
	<b>Finnish Social Science Data Archive</b>	Finnish Social Science Data Archive (FSD)	user-services.fsd@tuni.fi	FSD user services	services.fsd@tuni.fi	<a href="https://www.fsd.tuni.fi/en/">https://www.fsd.tuni.fi/en/</a>	FSD
	<b>FinSote</b>	Finnish Institute for Health and Welfare (THL)	Finnish Institute for Health and Welfare (THL)	Seppo Koskinen (seppo.koskinen@thl.fi) and Anne Lounamaa (anne.lounamaa@thl.fi)			Finnish Institute for Health and Welfare (THL)
	<b>Research Services at Statistics Finland</b>	Statistics Finland	Statistics Finland, FI-00022 Statistics Finland	Registrar's Office of Statistics Finland	kirjaamo@stat.fi	<a href="https://www2.tilastokeskus.fi/meta/tietosuoja/kaytola_en.html">https://www2.tilastokeskus.fi/meta/tietosuoja/kaytola_en.html</a>	CSC – IT CENTER FOR SCIENCE LTD
	<b>THL Biobank</b>	Finnish Institute for Health and Welfare	Sirpa Soini, sirpa.soini (at) thl.fi		admin.biobank (at) thl.fi	<a href="https://thl.fi/en/web/thl-biobank/for-researchers/application-process">https://thl.fi/en/web/thl-biobank/for-researchers/application-process</a>	No data processor organisation
Spain	<b>Collaborative Spanish Variant Server (CSVS)</b>	Fundacion Progreso y Salud	Javier Perez Florido (javier.perez.florido.sspa@juntadeandalucia.es)	Javier Perez Florido	<a href="mailto:javier.perez.florido.sspa@juntadeandalucia.es">javier.perez.florido.sspa@juntadeandalucia.es</a>		Fundación Progreso y Salud
	<b>European Genome-phenome Archive (EGA)</b>	multiple data controllers, one for each dataset					EGA



	<b>Plataforma de Información BIGAN</b>	Departamento de Sanidad de Aragón / Servicio Aragonés de Salud		IACS	bigan.iacs@aragon.es	<a href="https://www.iacs.es/instituto-aragones-ciencias-la-salud/oficina-virtual/solicitud-de-acceso-a-datos-para-realizacion-de-un-proyecto-de-investigacion-rpi01-3a/">https://www.iacs.es/instituto-aragones-ciencias-la-salud/oficina-virtual/solicitud-de-acceso-a-datos-para-realizacion-de-un-proyecto-de-investigacion-rpi01-3a/</a>	IACS
<b>Germany</b>	<b>State of Health in Pomerania (SHIP)</b>	University Medicine Greifswald		<a href="https://www.fvcm.med.uni-greifswald.de/">https://www.fvcm.med.uni-greifswald.de/</a>			University Medicine Greifswald

**Table 2a: Type of source**

	Finland									Belgium					Spain		Europe		Germany	
Does the data originate from a patient group, the general population or an experimental setting, or other?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	Eurobioimaging Italian MIMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Patient group		X				X		X	X								X	X		6
/ General population	X		X			X	X		X	X	X		X	X	X	X		X	X	13
/ Experimental setting																	X	X		2
/ Other				X	X			X				X						X		5
/ I don't know																				0
/ This doesn't apply to this data infrastructure																				0
/ If 'Other', please specify				Many different datasets, in which the universe and sampling procedures vary (all options above possible)	The data originates from research studies that are transferred to the biobank			Cancer screening				Patients diagnosed with cancer and/or patients that underwent cancer screening						Also non-human data and exposure data		

**Table 2b:** Type of source

	Finland									Belgium					Spain			Europe		Germany	
What is the type of data source that you are using? You can choose multiple options.	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
Electronic health records (EHR)						X		X	X			X			X				X	X	7
Clinical trials																			X		1
Survey	X		X	X	X	X	X			X	X								X	X	10
Cohorts				X	X	X									X				X	X	6
Biobanks (biological samples)	X				X										X				X	X	5
Picture Archiving and Communication System (PACS)															X				X	X	3
Imaging data															X			X	X	X	4
Medical devices																			X	X	2
Clinical Research data																			X	X	2
Genomic data	X				X							X	X		X	X			X	X	8
Biometric data	X				X														X		3
Molecular data					X							X							X	X	4

Socioeconomic data			X	X	X	X									X				X	X	7
Specific disease data					X	X						X			X				X	X	6
Survival data												X							X	X	3
Population health data					X	X						X		X	X				X	X	7
Interview data			X	X	X	X													X		5
Administrative data			X		X	X						X		X	X				X	X	8
Registry data		X	X	X		X		X							X				X		7
Customer record data		X			X																2
Observational study data				X															X	X	3
Healthcare data (Prescriptions / Diagnoses / Laboratory data/ Treatment / Surgery/ Other)						X		X	X			X			X				X	X	7
Other					X			X													2
If 'Other', please specify					Research collections			Pathology reports													

**Table 3:** Level of aggregation

	Finland									Belgium					Spain			Europe		Germany	
What is the level of aggregation of the data stored in this data infrastructure? e.g. aggregated, individual, both	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
Individual	X	X	X		X	X	X	X	X	X	X	X		X			X	X		X	15
Aggregated																X					1
Both				X									X		X				X		4

**Table 4: Anonymisation**

	Finland									Belgium					Spain			Europe		Germany	
Is the anonymisation performed by your data infrastructure and/or do you receive already anonymised data?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ We perform the anonymisation		X	X			X		X	X	X	X	X		X	X					X	11
/ We receive anonymised data																X					1
/ We do not anonymise data	X				X		X						X					X	X		6
/ Both				X																	1

**Table 5:** Pseudonymisation

	Finland									Belgium					Spain			Europe		Germany	
Do you have pseudonymised data?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Yes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X				X	X	17
/ No																X		X			2
/ I don't know																					0
/ This question doesn't apply to this data infrastructure																					0

**Table 6a to 6d:** Geographical and time coverage**Table 6a:** Geographical coverage

	Finland									Belgium					Spain			Europe		Germany	
What is the geographical coverage of the data infrastructure (datasets registered in your data collection, or data collections registered/linked in your data hub)?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MIMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
International				X															X		2
European				X															X		2
National	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		X			16
Regional				X										X	X					X	4
I don't know																					0
This question doesn't apply to this data infrastructure																					0



**Table 6b:** Participating countries

	Finland									Belgium					Spain			Europe		Germany
What are the participating countries from which you have datasets?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)
Free text	Finland	Finland	Finland	Broad representation of countries from every continent	Finland	Finland	Finland	Finland	Finland	Belgium	Belgium	Only data on National level	Belgium	Belgium	Spain	Spain	Worldwide	Italy	Depends on particular collection. It can be any BBMRI-ERIC member/observer country, and for COVID-19 and rare diseases, where it can be completely global.	Germany

**Table 6c:** Socioeconomic coverage

	Finland									Belgium					Spain			Europe		Germany	
What is the socioeconomic coverage of the data infrastructure (datasets registered in your data collection, or data collections registered/linked in your data hub)?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
NUTS 1		X	X	X	X			X	X	X	X		X	X		X	X	X		13	
NUTS 2		X	X	X	X			X	X	X	X			X					X	10	
NUTS 3	X	X	X	X			X	X	X	X	X	X		X	X					12	
I don't know						X														1	

**Explanatory text:** The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU and the UK for the purpose of collection, development and harmonisation of European regional statistics. NUTS 1: major socioeconomic regions. NUTS 2: basic regions for the application of regional policies. NUTS 3: small regions for specific diagnoses.

**Note:** No response was received from EGA for this question (see limitations section)

**Table 6d:** Time coverage

	Finland									Belgium					Spain			Europe		Germany
Time coverage	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)
When did your data infrastructure start collecting data? If this applies to your data infrastructure	Jan 2017	June 2005	May 2005	Doesn't apply	THL Biobank stores research collections collected since the 1960s.	Early 1950s	January 2020	Cancer information from 1953 and Screening information: cervical cancer 1991, breast cancer 1992	2011	January 2018	January 2018	Started data collection in 2004	July 2005	Continuous since 1841	Depends on the data set. First data sets (Hospital Discharge Database) from 1996	2010	2008	June 2020	2017	1998
Is the data collection period still ongoing? If this applies to your data infrastructure.	No	Yes	Yes		Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

<b>What is the end date of the data collection period? If this applies to your data infrastructure.</b>	October 2017				Some research collections stored at THL Biobank still actively collect new data, there is no specific end date.		November 2020	Current 2020		December 2018	December 2018	Not applicable	No end date	No end data		No end date	No end date		Depends on specific collection	
---	--------------	--	--	--	---	--	---------------	--------------	--	---------------	---------------	----------------	-------------	-------------	--	-------------	-------------	--	--------------------------------	--

**Table 7:** Ethical approval for storage of data

	Finland									Belgium					Spain			Europe		Germany	
Do you require ethical approval for the data to be stored in your infrastructure?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Yes	X						X		X	X	X									X	6
/ No			X	X		X		X				X		X	X				X		8
/ I don't know													X								1
/ This question doesn't apply to this data infrastructure		X			X											X	X	X			5

**Table 8a and 8b:** Data quality controls**Table 8a:** Are data quality controls applied?

	Finland									Belgium					Spain			Europe		Germany	
Are data quality controls applied?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Yes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	19
/ No																		X			1
/ I don't know																					0

**Table 8b:** Are there minimum levels of quality of the data needed for the data to be included in the data infrastructure?

	Finland									Belgium					Spain			Europe		Germany	
Are there minimum levels of quality of the data (results from quality controls) needed for the data to be included in the data infrastructure?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Yes, data is only included if it reaches a certain quality level		X	X	X	X			X	X	X	X	X				X			X	X	12
/ No, we do quality control for internal use only	X						X						X	X	X			X			6
/ No, but the results of the quality control are available when searching for the data																	X				1
/ Does not apply						X															1
/ Unknown																					0

**Table 9: Error checking**

	Finland									Belgium					Spain			Europe		Germany	
Do you use a tool to check for errors and completeness (e.g., Checksum tool)?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Yes		X		X	X	X		X	X	X	X	X	X			X	X		X	X	14
/ No	X		X				X							X	X			X			6
/ I don't know																					0
/ This question doesn't apply to this data infrastructure																					0



**Table 10:** Versioning of datasets

	Finland									Belgium					Spain			Europe		Germany	
Do you have a process to keep track of the different versions of the datasets?	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Plataforma de Información BIGAN	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)	Total
/ Yes			X	X	X		X	X	X	X	X				X	X			X	X	12
/ No	X	X											X	X				X			5
/ I don't know																					0
/ This question doesn't apply to this data infrastructure						X						X					X				3

**Table 11:** Data source legitimacy

	Finland						Belgium					Spain		Europe		Germany
Do you have a method to check data source legitimacy (e.g., ISO standard on data quality)? Please specify.	FinHealth 2017 Survey	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel	Collaborative Spanish Variant Server (CSVS)	EGA	Eurobioimaging Italian MMMI Node	BBMRI-ERIC	State of Health in Pomerania (SHIP)
Free text	No	No	No	Only official health and social care providers	No	No	Data are collected by us	Data are collected by us	Our registered data needs to meet ENCR, IARC International Guidelines.	Not yet. A EIM working group is expected to deliver standards on data quality in the coming years that will be followed	No	We only trace the provenance laboratory	We do not	No	I don't understand the question.	Yes

**Note:** No response was received from the following data infrastructures for this question: the Care Register for Social Welfare (Finland), Research Services at Statistics Finland (Finland), Avohilmo Register of Primary Care Visits (Finland), Plataforma de Información BIGAN (Spain)