HEALTHYCLOUD
Health Research & Innovation Cloud

D4.1 Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M18, August 2022
Dissemination Level	PU - Public
Nature	R - Report
Author(s)	Alicia Martínez-García (SAS), Celia Alvarez-Romero (SAS)
Contributor(s)	Carlos Luis Parra Calderón (SAS)
Reviewer(s)	Amy Curwin (CRG), Jordi Rambla (CRG), Lorenz Dolanski-Aghamanoukjan (GÖG)
Keywords	Health data management, health data infrastructure, health data hub, patterns of governance, governance models, actors and business processes, survey.

Notice: *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement Nº965345*

Change Log

Version	Author	Date	Description of Change
v0.1	Alicia Martínez-García	11.04.2022	Initial Table of Content.
v0.2	Alicia Martínez-García Celia Alvarez-Romero Carlos Luis Parra Calderón Silvia Rodríguez Mejías	04.07.2022	First content for review.
v0.3	Lorenz Dolanski-Aghamanoukjan Amy Curwin Jordi Rambla Stefan Klein	04.07.2022 - 20.07.2022	Reviews.
v0.4	Alicia Martínez-García Celia Alvarez-Romero Carlos Luis Parra Calderón Silvia Rodríguez Mejías	22.07.2022	Improvements in terms of english writing, updated percentages, diagrams and tables included, etc. Covering reviewers' comments.
v0.5	Alicia Martínez-García Celia Alvarez-Romero Carlos Luis Parra Calderón	25.07.2022	Second draft sent to HealthyCloud Coordinators and to the data hubs that have answered the survey.
v0.6	HealthyCloud Coordinators	25.07.2022 - 07.08.2022	Reviews.
v1.0	Alicia Martínez-García Celia Alvarez-Romero	08.08.2022	Final version sent to coordinators.
v1.1	Alicia Martínez-García Celia Alvarez-Romero Juan González-García Alba Jene	30.08.2022	Final version submitted.

Table of contents

1. Executive Summary

HealthyCloud WP4 is focused on how health data is managed by dedicated infrastructures: the data hubs.

The first task of WP4 seeks to capture the different **governance and auditing models** behind **data hubs across Europe** managing health data to analyse the existing regional and national initiatives, as well as European projects related to domain-specific data hubs. For this purpose, in collaboration with the leaders of WP3, a survey was designed and carried out.

Deliverable D4.1 includes the **analysis of the survey responses**, through the stratification of the results. As results of this deliverable, patterns of data hub governance are represented. Finally, recommendations for integration in HealthyCloud are shown in the conclusions.

1. Introduction

The deliverable **D4.1 'Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance'** covers recommendations for integration in the HealthyCloud ecosystem, including an analysis of **health data hub patterns of data governance**. To in-depth understand how health data is managed by the dedicated infrastructures called data hubs, a capture of the different governance behind data hubs across Europe managing health data was performed. For that, existing regional and national initiatives, as well as European projects and initiatives related to domain-specific data hubs, were analysed.

In the HealthyCloud project, a **health data hub** is defined as a data infrastructure with the following minimal inclusion criteria [1]: (i) A digital technical infrastructure with the core mission of enabling health data sharing. (ii) It provides health data from different sources. (iii) It allows discovery of health datasets. (iv) It has a metadata discovery service. (v) It has a data accessibility mechanism in accordance with existing regulation. (vi) It has an authorisation functionality, provided by the same Data Hub or by an external institution.

In the HealthyCloud project, **data governance** is defined as the "assembly of policies and processes, coordination aspects, data usage and accessibility principles and data management procedures for a certain health data infrastructure to ensure legal compliance, consistency and good data quality throughout the different stages of the data life cycle" [1].

The milestone **MS4.1 'Community activity: selection of representative data hubs'** [2] was reached in the month M6 (August 2021), collecting a list of representative data hubs in Europe.

Then, WP4 members worked together with WP3 participants, aiming to collaborate on carrying out a survey that meets the purposes of WP3 and WP4, avoiding sending

more surveys than necessary and not overloading the respondents. This survey was sent in M11 (January 2022) to the list of representative data hubs collected in MS4.1. The milestone **MS4.2 'Study: patterns of governance of selected data hubs'** [3] was reached in the month M12 (February 2022), performing a preliminary analysis of the 34 survey responses received until February 2022. In addition, an All Hands Meeting was scheduled in February 2022 with the HealthyCloud consortium with the aim -among others- to gather feedback to improve the survey analysis and complete the deliverables **D3.1 'Landscape analysis of FAIRness levels of health-related data using catalogue matrix'** and **D4.1 'Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance'**.

With this background, the WP4 work continued **delivering this D4.1**: on the one hand, reviewing in-depth the list of data hubs that answered the survey, to identify missed countries to have a broad European representation, and identify missed relevant data hubs, concluding in a lot of reminders and new contacts. And on the other hand, analysing in-depth the 41 survey responses finally received until June 2022, with the final aim to define the health data hub patterns of data governance (Section 4).

2. Methods

The final aim of the related task T4.1 was to **analyse existing data hubs governance models** and **describe patterns of governance for data hubs** generated after identifying commonalities in the governance models of existing data hubs. The fact that every data hub has similar processes like accepting new submissions, getting submissions done, applying quality control, publishing the dataset for discovery, accepting requests, among others, was presupposed.

First of all, a **survey** was designed and developed jointly with WP3 (to join efforts and share outcomes). The main objectives of the survey were: (i) To evaluate the feasibility of linking individual level data between data collections; and (ii) to perform a landscape analysis of the different governance models in those data infrastructures. This survey was developed in an electronic tool (Typeform), after receiving contributions from numerous HealthyCloud partners following meetings and email exchanges. In addition, the survey was sent in November 2021 to 4 pilots (EGA, ELIXIR-LU transmed data-hub, Belgian Cancer Registry, and Healthdata.be infrastructure) in order to receive their feedback and include more improvements. The final version of the survey can be found at this link: https://bsc3.typeform.com/to/zY1FNgSQ, and in Annex 1.

This work was based on the previous identification of **European data infrastructures** (related to milestone MS4.1 reached in August 2021) which are the target audience of the survey. Once the survey was defined, the survey was sent out at the beginning of January 2022 to the list of identified contacts. The survey

was sent to a list of 69 data hubs (and around 30 other relevant European data infrastructures, in relation to WP3 purposes).

On the other hand, this work was also based on the **initial preliminary analysis of the survey responses** (related to milestone MS4.2 [2] reached in February 2022). For this milestone, 34 responses from data hubs (and 17 responses from other kinds of European data infrastructures, in relation to WP3 purposes) received until February 2022 were analysed.

National, European, and Worldwide data hubs were interviewed through the survey. After the initial analysis mentioned previously, an effort was performed to ensure a robust representation of all the data hubs in Europe. Finally, **the survey was sent to a representative list of 99 data hubs** (Figure 1). So, this effort started by surveying 69 existing data hubs and iteratively continued until covering most of the existing patterns by surveying 99 existing data hubs.

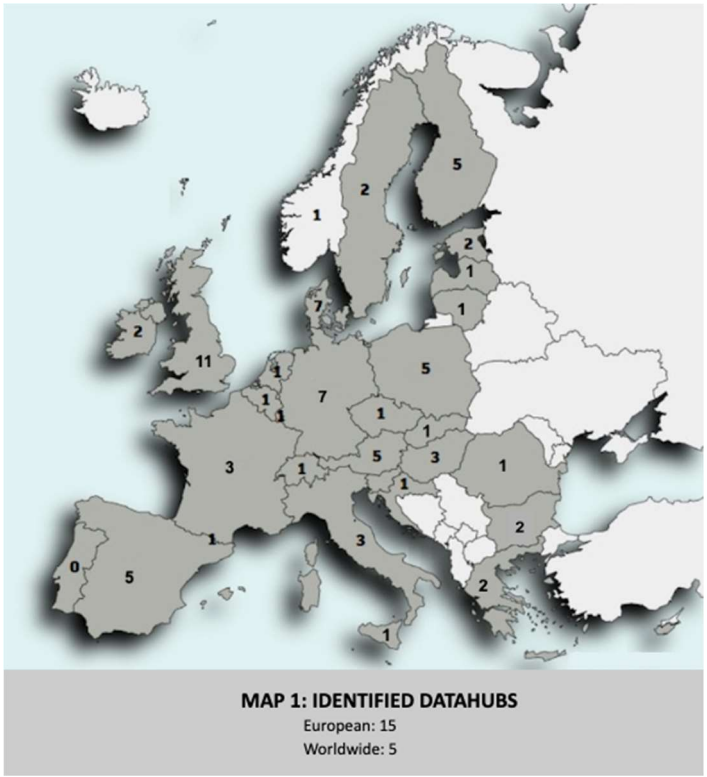

Figure 1: Map with contacted data hubs

41 out of the 99 (41%) contacted data hubs answered the survey until June 2022. The categorisation in terms of geographical coverage was performed depending on the response in the question "What are the participating countries from which you have datasets?" included in section "Completeness of data infrastructure > Participating countries" (Annex 1). Overall, The WP4 team sent more than 400 emails (new contacts, and reminders) to achieve this robust representation.

Figure 2 shows the final geographical coverage achieved through the survey responses:

Figure 2: Map with received responses from contacted data hubs

With the final number of 41 responses out of the 99 contacted data hubs, all the material collected through the survey was **analysed** (both structured and free-text questions) focusing on identifying **actors** and **business processes** involved in the hubs' governance, also considering **ELSI** (Ethical, Legal, Societal Impact) **aspects**. In-depth details about this analysis are included in Section 3.

To appropriately cover the different levels of granularity identified in some data hubs' characteristics (such as geographical coverage, kind of data hub organisation, role, etc), **stratifications** (i.e. segmentation of the responses to be analysed) were performed using characteristics such as the kind of data hub organisation (centralised vs. federated, more detail in Section 3.1), the role applied in data management (data controller vs. data processor, Section 3.2), the geographical coverage (i.e. European, Worldwide, Section 3.3), or the kind of data source (i.e. EHRs, administrative data, registry, specific disease data, Section 3.4) delivering specific profiles.

After performing the survey analysis, in Section 4 a general **pattern of governance** is described after identifying commonalities between the analysed governance models. As a result of the stratifications, **profiles** covering specific patterns of governance are described (Sections 4.1 to 4.4). This description is complemented

with key performance indicators (KPIs) (Section 4.5) useful to analyse the performance of a specific governance model.

Finally, to facilitate the process and to promote participation, these resulting patterns of governance (Section 4) were validated with the data hubs that have taken part in the process, involving them in the review phase of this deliverable.

3. Analysis

This section includes the in-depth details obtained during the analysis of the 41 responses of the survey that was carried out. During the analysis, to improve the readability, the decimal places were considered not representative, so all percentages were rounded without using decimal places, taking into account that with 41 responses, the minor step (1 answer more or less) is more than 2%.

Data hub criteria

First of all, it is important to note that the HealthyCloud consortium defined a set of inclusion criteria for defining a data infrastructure as a data hub [1]. The same method was used to define health data collection in order to distinguish between the two terms.

In the case of the health data hub concept, the minimal inclusion criteria are [1]: (i) A digital technical infrastructure with the core mission of enabling health data sharing; (ii) It provides health data from different sources; (iii) It allows discovery of health datasets; (iv) It has a metadata discovery service; (v) It has a data accessibility mechanism in accordance with existing regulation; (vi) It has an authorisation functionality, provided by the same Data Hub or by an external institution.

Apart from these characteristics, from the survey responses in a multiple-choice question we can state that: 27 added to this minimal inclusion criteria the feature "A digital platform that receives and stores data", 30 added the feature "It receives data from a single source and/or multiple sources", and 26 added the feature "It has control over the data stored".

Data hub main features

All data hubs have provided their official titles and websites. On several of the websites, a Data Governance section is included in the website. This finding is an important recommendation included in the patterns of governance included in Section 4. The Governance sections of the different websites were explored to take the content into account in the proposed patterns of governance.

Data governance is the assembly of policies and processes, coordination aspects, data usage and accessibility principles and data management procedures for a certain health data infrastructure to ensure legal compliance, consistency and good

data quality throughout the different stages of the data life cycle [1], defining between other aspects who within an organisation has authority and control over data assets and how those assets may be used. And as we have defined previously, under the European Union (EU) Regulation 2018/1725, as well as under the General Data Protection Regulation (GDPR), the data controller is the party that, alone or jointly with others, determines the purposes and means of the processing of personal data [1]. 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 for respecting the rights of the data subject. The controller is also the entity that receives requests from data subjects to exercise their rights. Therefore, excluding the case of distributed and/or federated data infrastructures, all data hubs identify the data controller, and where applicable the data access provider and the data processor.

Regarding whether the data infrastructure is part of a data hub, 32% of those surveyed answered yes; that is, 13 of 41. All of them provided the name of the associated data hub or data hubs or the link to the web page. From these links, a total of 15 data hubs were obtained, of which 9 were already known to us and 6 were not in our initial list of data hubs. Of these new data hubs, two were from Germany, two from Europe, one from Finland, and one from the UK.

Regarding the data infrastructure organisation, 22% answered "It has a decentralized management", and 70% answered "It is managed centrally", the rest did not apply or did not answer.

Data management

The first question related to data was whether ethical approval is required for data to be stored in the infrastructure. 29% confirmed that it is required. However, 46% answered that it is not required, and 24% that it does not apply to this data infrastructure.

Related to the origin of the data, in a multiple-choice question, the most frequent situation came from a patient group (selected by 24) or the general population (29).

The survey (Annex 1) includes questions about the type of data source that data hubs use, offering between the multiple-choice options electronic health record (EHRs) (selected by 25), administrative data (22), registry data (22), and health care data (22), among others less common. Related to how the data is compiled before storage, in a multiple-choice question the respondents answered "Data retrieval" (selected by 19), "Transforming" (16), "Loading" (19), "ETL methods" (18), among other options less frequent.

Regarding the question about the technologies used for data storage and the format in which the data is stored, they depend on the type of data stored (texts, numbers, files, images, spreadsheets, tissue samples, sounds, among others less common).

Regarding the level of aggregation of the data stored in the data hubs, 46% of the respondents answered "Individual", and 44% selected "Both" (that is, both individual and aggregated). Only 1 respondent (2%) answered "Aggregated", and 3 of the respondents (7%) added that question does not apply to their data infrastructure.

Concerning anonymisation, 65% (26) of the respondents stated anonymisation methods are used in those data hubs: 8 data hubs answered that they use anonymisation methods at the point of collection, 3 before sharing them internally, 11 data hubs before sharing them externally, 3 at the point of publishing, and 1 not specified. On the other hand, 20% (8) do not anonymise data in those data infrastructures. This question does not apply to the 15% (6) of the respondents.

Regarding if the anonymisation is performed by the data infrastructure and/or the data is received already anonymised, this question was not answered from the 35% that in the previous one stated they do not anonymise data or the question does not apply to them. Of the 25 responses with information, this question concluded that 48% of these data hubs perform the anonymisation and 24% receive anonymised data. Both events occur in 28% out of the 25 data hubs.

Regarding pseudonymisation, it has to be noted that 80% of the respondents have pseudonymised data, versus 7% who do not. 10% added that the question does not apply to their data infrastructure, and 2% stated that they did not know.

Completeness of data hubs

Regarding the geographical coverage for the data collections registered/linked in the data hubs, in a multiple-choice question 10 out of 39 data hubs answering to this question reported international level, 13 cover European level, 27 cover national level, and 10 cover regional level, specifying in all cases the participating countries. Related to the socioeconomic coverage following the NUTS classification, 2 interviewees did not answer this question, and of the remaining 39, in a multiple-choice question, 13 respondents answered they did not know this detail, 18 selected NUTS 1, 14 selected NUTS 2, and 18 selected NUTS 3.

Other questions were related to the data collection period, with very varied results. 20 out of 34 respondents have a data collection start date before 2015, although in many cases the date depends on the specific dataset. 14 out of 34 respondents have a data collection start date after 2015. All 35 of the respondents to the question, if data collection was still ongoing, answered with yes. 18 of them further specified that the end date is undetermined, not applicable or depending on the specific dataset, only one data hub specified a future end date. A possible interpretation is

that the 7 respondents without answer in this question meant that the data collection is not ongoing, but none of them specified an end date.

Data quality aspects

The first conclusion of this survey section is that 83% of the respondents stated that data quality controls are applied in their data hubs. 7% answered that they do not use data quality controls and 10% answered that they did not know it. Another finding to note is that only 17 out of 38 respondents stated data is only included if it reaches a certain quality level. 6 out of 38 respondents stated they do quality control for internal use only, and 7 out of 38 answered that minimum levels of quality of the data are not needed for the data to be included in the data infrastructure but the results of the quality control are available when searching for the data. 6 out of 38 do not apply and 2 out of 38 answered “Unknown”.

Regarding the updating periodicity applied to upgrade the datasets, of the 40 data hubs that answered this multiple-choice question, 5 of them stated this characteristic does not apply, and the most selected options were“irregularly” (12), “annually” (11), and “daily” (10).

Another aspect related to data quality is checking for errors and completeness. In this survey, 61% of the respondents stated to use a tool for error checking, compared to 24% who do not. 2 out of 41 respondents (5%) answered that they do not know and 4 out of 41 (10%) stated that the question does not apply to that data infrastructure. Out of the 25 who answered to use a tool for error checking in the previous question, 21 (84%) specified the tool they use. And 7 out of 25 (28%) specified the checksum technique in their answer.

In addition, keeping track of the versions is very common for the data hubs that answered the survey since 24 out of 41 (59%) stated that they have a process to keep track of the different versions of the datasets, versus 8 out of 41 (20%) that stated they do not have this kind of process. 8 out of 41 (20%) answered that the question does not apply to that data infrastructure and 1 out of 41 (2%) answered that they did not know. Out of the 24 who answered to keep track of the version process, 19 (79%) specified the process they use.

To conclude this section, it was asked whether they have a method to check data source legitimacy. Only 26 answers to this question could be analysed, of which 12 (46%) answered yes and 14 no (54%).

Data hub usage

This section relates to data size and/or amount in the 41 data hubs. When asked ‘how much storage capacity is in use up to date?´ there were a large range of answers (500MB, 10GB, 500 GB, more than 60TB, several petabytes, etc), while 7 were unable to answer. In terms of the number of data collections stored in the data hub, the responses again vary significantly, from 1 to "1 billion facts", the

highest number of studies being 7690 and ranging from 1 to hundreds in between, while 9 of 41 were not able to answer.

When asked about estimated annual growth, 15 were unable to answer. Of those that answered it ranged from 5 collections/year, 8TB, 500 up to 200000GB. The questions were given as free text, therefore the responses also vary in their output.

The number of users again varied extensively; submitters e.g. 1, 10, 20, 40, 500, 700 (13 did not answer); users with access e.g. 0, 20, 100, 300, 26000+ (11 did not answer).

GDPR compliance

As part of legal characteristics, one of the relevant aspects included in the survey is related to the national rules existing and used in addition to the GDPR in each specific country, if any. 24 of 41 interviewees answered this question. 8 of these 24 stated this data hub does not use any additional rule to the GDPR or the respondent does not know this detail, so 16 of 24 data hubs (67%) declared they use national rules. 2 of these 16 stated that this data infrastructure has performed national interpretations of the GDPR but did not specify information related to these interpretations. The regulations mentioned by the 15 data hubs with tangible information were explored to take it into account to deliver this analysis.

In the scope of the EU GDPR, and in relation to personal data, 28% and 28% of the respondents declared their organisations have the role of Data controller and Data processor, respectively. 31% stated their organisations have different roles depending on the specific situation. And 13% declared they are in a situation different from the above, e.g. a decentralised data management strategy. In this case, 39 of 41 answers cover this question.

In terms of logging and auditing processes of user actions, in a multiple-choice question 10 of 35 (excluding 6 empty answers) declared this does not apply to their organisation. About the remaining 25: (i) 20 declared their organisation uses a record of user deposition including date and time, (ii) 17 of 25 declared their organisation uses a record of user contact to client service, (iii) 18 declared their organisation uses a record of user application for data use (download and/or see).

Data management

This subsection covers data management aspects, strongly linked with legal aspects in general and GDPR compliance.

The survey asked if there was a formal procedure to know who provides the data. 4 of the 41 answers did not complete this question. Of the remaining 37 answers, 16% stated they do not use a formal procedure to know who provides the data, but 84% do so. For these, the survey asked about specific procedures (i.e. contracts, agreements, open information in the organisation), obtaining in the responses several specific procedures: legal contracts, different kinds of agreements

(collaboration, accreditation data access, confidentiality, data transfer, data sharing, data processing, use, deposition, etc.), regulations, open information in the organisation, queryable resource information on data access and data re-use conditions, terms of use, licences, the user needs to register, mandatory institute email address, information about the principal investigators and the project, alliance membership, assigned Data Access Committee, data permissions based on the Act on a secondary use.

Related to a Data Access Agreement (DAA) to be signed between data providers and data requesters, 38 of 41 answers cover this question. 55% (21) of the 38 interviewed data hubs' provide a DAA, 24% do not, and 21% selected "Others" stating, among others, that it depends on the specific resource queried or that only employees access the data directly. 52% of the 21 with DAA use a non-negotiable DAA form, and 48% provide a DAA template that may be modified under the agreement. In terms of a Data Processing Agreement (DPA) to be signed with the data providers, 38 of 41 answers cover this question. 47% (18) of these provide a DPA, 32% do not, and 21% detailed other options such as they have pending to cover the DPA management. 39% of the 18 with DPA use a non-negotiable DPA form, and 61% provide a DPA template which may be modified under the agreement. Regarding if the data hub has a Data Protection Impact Assessment (DPIA) model, 36 of 41 answers cover this question. 56% of the 36 data hubs use a DPIA model, and 44% do not.

About the kind of access control mechanisms (authentication and authorisation) implemented by the data hubs, in a multiple-choice question, 14 of 41 answers are empty, so there are 27 answers with information. 4 stated this data hub has no access control mechanisms. The following access control mechanisms are used: OAuth2 (9), OpenID Connect over HTTPs (8), authorisation over SSH (7), authorisation with web services backed by a database (10), and authorisation via (web) Rest API (2). No data hub uses authorisation (read) over AMQPs. And 2 stated they use other kinds of mechanisms without specifying if the mechanism is programatically implemented or not.

Funding

As part of the sustainability plan, the survey goes in-depth about the type of funding and the sustainability plan of this current funding. In this case, 38 of 41 answers cover this question regarding type of funding: national funding for the Hub core function (66%), participation in projects (16%), European or international funding (11%), and private funding (8%).

Regarding the sustainability plan**:** 42% receives stable funding (of which 39% stated this stable funding is of national origin), 13% present funding from private profits (i.e. data licence fees, pay for customer use, etc.), 32% are applying to infrastructure funding (national, European, and/or international), and 6% stated their plan to apply for competitive plans or projects related to research funding. Related to the

geographical scope of these fundings, including stable, non-stable, and expected profits, 77% of the data hubs stated they received funding from regional or national organisations, and 32% from European or international organisations (42% of the data hubs did not specify the geographical scope, so these numbers could be biassed).

Other data governance aspects

Concerning a catalogue of the different data sources, 34 of 41 data hubs covered this question. 21% of these 34 do not offer this kind of catalogue, because this specific data hub is connected only to a unique data source. And 79% of 34 provide a catalogue of different data sources.

In terms of the process to connect with the external data, a specific data hub could receive and store the data (centralised), or could link to the data remaining in the original place (federated). 39 of 41 data hubs covered this question**.** 77% and 23% of these 39 stated they are a centralised or federated data hub, respectively.

About the services used for data sharing (a specific data hub could use a different kind of service depending on the need), analysing a multiple choice question, it is concluded that 55%, 23%, 23% uses website/webportal, APIs (REST, RDF, or other kinds of APIs), or FTP/SFTP, respectively.

The relevant documentation related to Data Policy, Licence Model and Terms of Use provided by the 15 interviewees that answered this question with tangible material were explored to take it into account to deliver this analysis.

Related to the Standard Operating Procedures (SOPs) that the data hub's organisations follow and update regularly, 34 of 41 interviewees answered this question, stating that 79% use and 21% do not use this kind of procedures.

Reflection: The most frequent aspects

After performing this in-depth analysis of the 41 responses of the survey that was carried out, below the most frequent aspects and the corresponding percentages are included below.

Simple-choice questions (with percentages):

- [Procedure] Formal procedure to find out who provides the data (84%).
- [Quality] Quality control is applied to the data (83%).
- [Catalogue] A catalogue of the different data sources is provided (79%).
- [SOPs] There are Standard Operating Procedures (SOPs) that are followed and regularly updated (79%).
- [Receive] They receive health data from different sources (76%).
- [Centrally] The data infrastructure is centrally managed (75%).

- [Anonymisation - Pseudonymisation] Data anonymisation methods are used (65%), using pseudonymised data (80%).
- [Errors] A tool is used to check for errors and data integrity (61%).

Multiple-choice questions (with absolute values):

- [Text-Numbers] One type of collected data is "text" (34) or "numbers" (34).
- [Population - Patients] Data come from the general population (29) or from a group of patients (24).
- [Accessibility] A data accessibility mechanism is available in accordance with current regulations (28).
- [National - National funding] The coverage of the data infrastructure is national (27), receiving national funding (19).
- [Provide] They provide health data from different sources (28).
- [Receives and stores] They are a digital platform that receives and stores data (27).
- [Findability] They allow the discovery (findability) of health data sets (26).
- [Control] They have control over stored data (26).
- [Discoverability] Enable discoverability of health data sets (26).
- [Authorisation] They have authorisation functionality, provided by the organisation itself or by an external institution (25).
- [EHR] The type of data source used is the electronic health record (EHR) (25).

Using the conclusions obtained in the in-depth analysis of the 41 survey responses, and taking into account this previous list of more frequent aspects, the Section 4 (Results: patterns of governance) is defined.

3.1. Stratification depending on the kind of data hub organisation

To cover this stratification, the question “How is the data infrastructure organised?” was analysed. From the 41 surveyed data hubs, 40 have answered and 1 has not answered. Of these 40 responses, 30 (75%) answered “It is managed centrally” , 9 (22%) answered “It has a decentralised management” and 1 (2%) answered “This does not apply to this data infrastructure”. Figure 3 shows this distribution related to how the data infrastructure is organised.

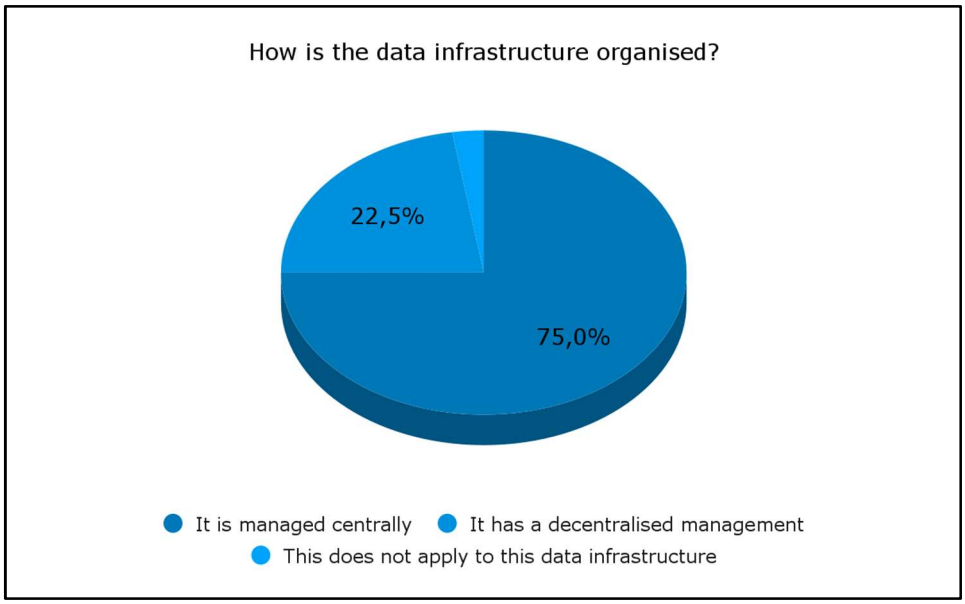

Figure 3: Distribution related to how the data infrastructure is organised

Analysing by subgroups, two profiles are proposed (Section 4.1): “data hubs managed centralised” and “data hubs managed decentralised”.

Regarding “data hubs managed centralised'', 23 have control of the data stored, and the type of data used is "Text" in 26, and "Numbers" in 26. In addition, 25 receive and store data from a single source and/or from multiple sources. 90% pseudonymise data, 90% apply data quality control, 81% establish standard operating procedures (SOPs) that the organisation follows and updates regularly, in 89% there is a formal procedure to know who provides the data, and 83% require legal approval for the data.

Regarding “data hubs managed decentralised'', they may not have a single data controller and may not have a data management strategy. 9 (all of them) allow the discovery (findability) of health datasets and 8 are a digital technical infrastructure with the core mission of enabling health data sharing. 8 host data that comes from "patient groups", 7 from "general population" and 7 from "experimental settings". In 7, the data is stored in "xml" format. The type of data collected is "text" in 7, "images" in 7, and "numbers" in 7.

3.2. Stratification depending on the role

For this stratification, the question “What is your organisation's role in relation to personal data?” was analysed. From the 41 surveyed data hubs, 39 have answered and 2 have not answered. Of these 39 answers, 11 (28%) answered “Data controller”, 11 (28%) answered “Data processor”, 12 (33%) answered “We have different roles in different situations” and 4 (10%) answered “None of the above”. Figure 4 shows the distribution related to the infrastructure role.

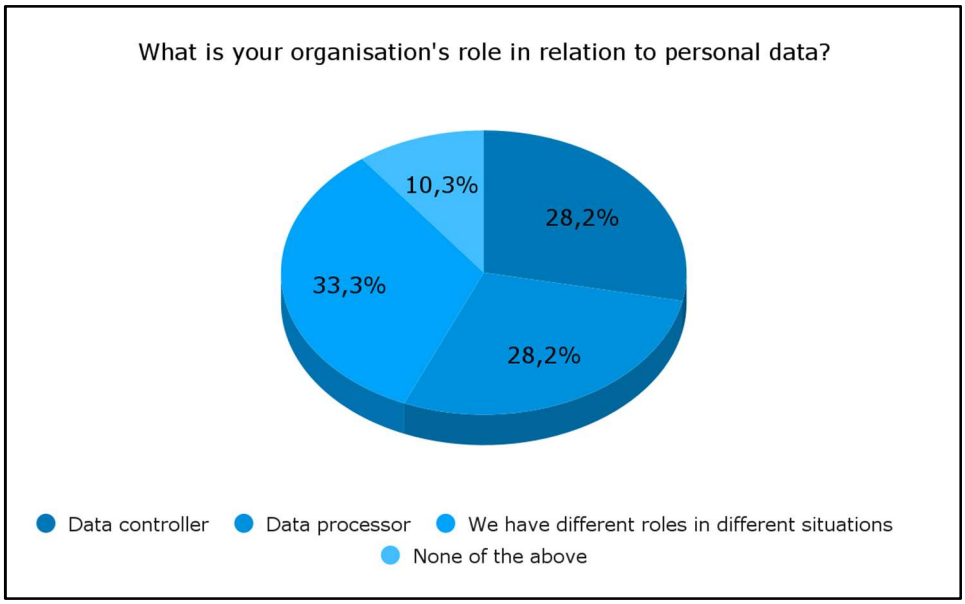

Figure 4: Distribution related to the infrastructure role

Analysing by subgroups, two profiles are proposed (Section 4.2): “data controller” and “data processor”.

In the case of “data controller”, 82% manage centrally, 100% pseudonymise data, 10 have control over the data stored, and 9 receive data from a single source and/or multiple sources. 9 of them have data from "general population". 10 of them uses "text" as data type. 82% have a process to keep track of the different versions of the datasets and 90% havea formal procedure to know who provides the data. 81% established SOPs that the organisation follows and updates regularly and 81% provide a catalogue of the different data sources.

Regarding “data processor”, 80% manage centrally, 80% have pseudonymised data, and 9 are a digital platform that receives and stores the data. In addition, 90% have an authorisation functionality provided by the organisation itself or by an external institution, and 90% have a data accessibility mechanism in accordance with existing regulations. 91% have a formal procedure to know who provides the data, and 80% have established SOPs that the organisation follows and updates regularly.

3.3. Stratification depending on the geographical coverage

To carry out this stratification, the question “What are the participating countries from which you have datasets?” was analysed. From the 41 surveyed data hubs, 40 have answered and 1 have not answered. Of these 40 responses, 7 (17%) are European, and 6 (15%) are worldwide. The rest, 27 (67%) have national or regional coverage in each country. Figure 5 shows the distribution related to the geographical coverage.

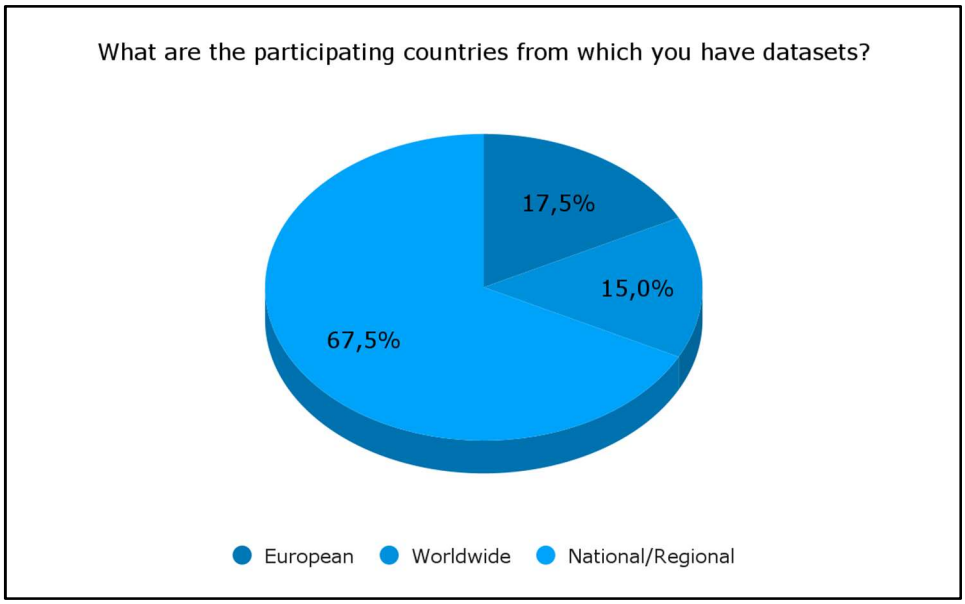

Figure 5: Distribution related to the geographical coverage

Analysing by subgroups and to achieve more global conclusions similar to HealthyCloud's geographic approach, two profiles are proposed in this stratification (Section 4.3): "European data hubs" and "worldwide data hubs" depending the geographical coverage, since the rest only have national coverage and each country should be grouped individually.

In the case of "European data hubs", 6 have data originating from "General population". 6 use the type of data "Numbers". 86% have a formal procedure to know who provides the data, and 86% have established SOPs that the organisation follows and updates regularly.

Regarding "worldwide data hubs", 5 have a data accessibility mechanism in accordance with existing regulation, 5 use the data source type "Socioeconomic data". 100% have a formal procedure to know who provides the data, 100% provide a DAA, 83% have a DPA, and 83% provide a catalogue of the different data sources.

3.4. Stratification depending on the source of the data

For this stratification, the question "What is the type of data source that you are using?" was analysed. From the 41 surveyed data hubs, 40 have answered and 1 has not answered. Of these 40 multiple-choice answers, the most frequent options are: 25 data hubs selected the option "Electronic health records", 22 data hubs selected the option "Administrative data", 22 data hubs selected the option "Registry" and 20 data hubs selected the option "Specific disease data". Figure 6 shows the distribution related to the source of the data.

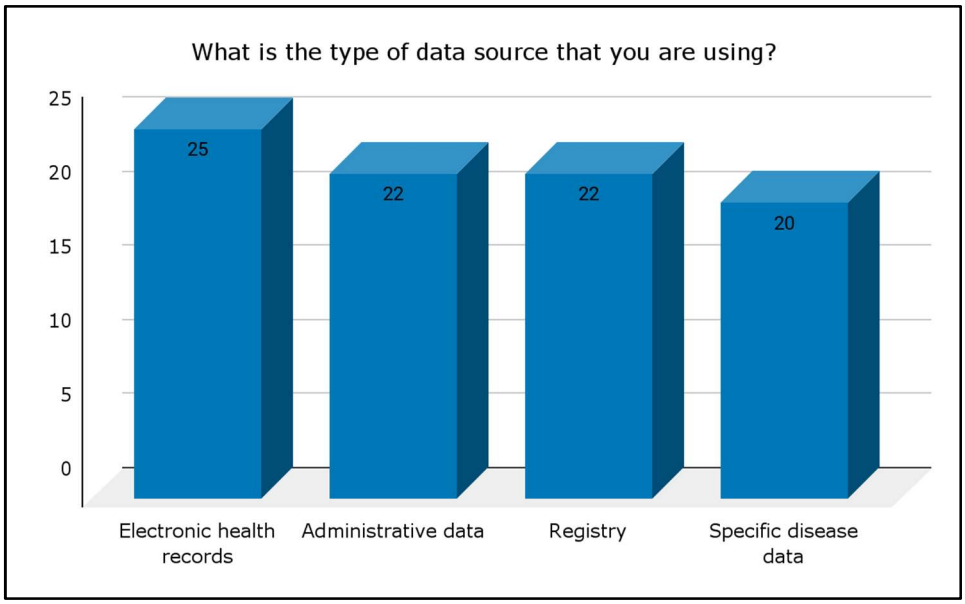

Figure 6: Distribution related to the source of the data

Analysing by subgroups, four profiles are proposed (Section 4.4): "data hubs using EHRs as one type of data source", "data hubs using administrative data as one type of data source", "data hubs using registry as one type of data source" and "data hubs using specific disease data as one type of data source".

Regarding "data hubs using EHRs as one type of data source", 76% manage centrally, 19 receive data from a single source and/or multiple sources, 21 provide health data from different sources, and 19 have a data accessibility mechanism in accordance with existing regulation. Data comes from "General population" in 20. The type of data source used is "Healthcare data" in 20 of them. The types of data collected are "Text" in 23, and "Numbers" in 21. 88% anonymise and 88% have pseudonymised data. 92% apply quality control, and 86% have a formal procedure to know who provides the data. 82% receive national funding, and 90% provide a catalogue of the different data sources.

Regarding "data hubs using administrative data as one type of data source", 77% manage centrally, 18 of them receive data from a single source and/or multiple sources, and 18 of them have control over the data stored. In addition, 19 provide health data from different sources, 17 have a data accessibility mechanism in accordance with existing regulation, and 17 have data from "General population". 18 of them use the type of data "Electronic Health Records", and 17 "Healthcare data". The types of data collected are "Text" in 20, and "Numbers" in 20. Also, 77% anonymise data, 86% have pseudonymised data, and 91% apply data quality control. 79% have a formal procedure to know who provides the data, and 77% receive national funding.

Regarding "data hubs using registry as one type of data source", 82% manage centrally, 19 receive data from a single source and/or multiple sources, and 17 have

control over the data stored. 17 of them have data originating from "General population", and the type of data collected are "Text" in 18 and "Numbers" in 19. In addition, 82% have pseudonymised data, 86% apply data quality control, and 89% have a formal procedure to know who provides the data. Also, 78% provide a catalogue of the different data sources, and 82% have established SOPs that the organisation follows and updates regularly.

Regarding "data hubs using specific disease data as one type of data source", 75% manage centrally, 17 receive data from a single source and/or multiple sources, and 15 have control over the data stored. 16 provides health data from different sources, 15 allow discovery (findability) of health datasets, and 15 have a metadata discovery service. In addition, 16 have a data accessibility mechanism in accordance with existing regulation. The types of source used are "Electronic Health Records" in 16, and "Administrative data" in 15. The types of data collected are "Text" in 17, and "Numbers" in 18. 75% anonymise data, and 85% have pseudonymised data. Also, 90% apply quality control, 15 use "time stamp of data deposition" for the logging and auditing of user actions, and 94% have a formal procedure to know who provides the data. Besides, 87% provide a catalogue of the different data sources , 83% have "National funding", and 81% receive data from different sources. 80% have established SOPs that the organisation follows and updates regularly.

4. Results: patterns of data hub governance

This section defines a general pattern of data governance for data hubs, using the conclusions obtained in the in-depth analysis of the 41 survey responses (Section 3), and taking into account the list of more frequent aspects (Section 3 > Reflection: The most frequent aspects). To define the general pattern of governance, a common characteristic was considered if the respondents coincided by at least 60%.

Hereafter, specific profiles are defined generating specific patterns of data governance for data hubs (Sections 4.1 to 4.4), using the conclusions obtained in the stratifications (Section 3) in terms of the kind (centralised vs. decentralised), role (controller or processor), geographical coverage (European or worldwide), and data source. For the specific patterns of governance, it has been counted from 75%. It was needed to reduce this percentage in the case of the general one (from 75% to 60%), because when all the responses together were analysed less commonalities were found.

For each pattern of data governance (both the general one, and the specifics), data aspects, business models, and ELSI aspects are defined, preceded by the list of actors involved in these processes.

After that, Section 4.5 defines a list of KPIs to analyse the data hubs performance, and Section 4.6 includes a list of terms used in the definition of the patterns of governance for data hubs.

Actors

In a data hub, the data controller refers to the "party that, alone or jointly with others, determines the purposes and means of the processing of personal data" [1]. Depending on the data hub, there may be one or more data controllers and sometimes there is a data controller for each data set. The data controller can be any institution, such as a research institute, university, hospital, health service, etc.

The data access provider is defined as "an entity which makes data available for secondary use" [1]. There may be one or several, it may be a person or a set of mechanisms.

The data processor determines who is in charge of data processing, "which processes personal data on behalf of the controller" [1]. This actor can vary depending on the particular case, it can be the same data hub, another institution, or there can be no data processor.

We also find other relevant actors such as researchers, ethical and scientific committees, advisory committees, management boards (government bodies that evaluate applications) or data protection agencies among others.

Regarding the organisation's role in personal data, data hubs are data controllers or data processors. They also can have both roles depending on the specific situation.

Data aspects

Concerning characteristics that are frequently present in the data hubs, these kind of data infrastructures usually:

- are digital platforms that receive and store data,
- have control over the stored data, receive data from a single source and/or multiple sources,
- are a digital technical infrastructure with the core mission of enabling health data sharing and providing health services data from different sources enabling the discovery of health data sets by having a published metadata discovery service and data accessibility mechanism in accordance with existing regulation that has an authorisation functionality,
- provided by the data hub itself or by an external institution.

In addition, although less common, a data hub can have characteristics such as generating data, being part of one or more overarching data hubs, or having a specific thematic or collected data type (e.g., a particular disease, a particular data type, etc), among others.

Related to the geographical coverage of the data infrastructure, it can be national, which is the most common, or with less frequency European, regional or international.

As far as the organisation of the data infrastructure, the most common is in a centralised way, and less frequently in decentralised (federated) way. A data hub can also be part of another data hub, although this characteristic is not very frequent.

With regards to the origin of the data, health data usually comes from the general population or from a patient group. With less frequency, health data comes from an experimental setting, among others.

Common types of data source are EHRs, administrative data, registry data, and health care data, such as, prescriptions, diagnoses, laboratory data, treatment, surgery, etc. Nevertheless, other types of data sources can also be clinical trials, surveys, cohorts, biobanks (biological samples), Picture Archiving and Communication System (PACS), imaging data, medical devices, clinical Research data, genomic data, biometric data, molecular data, socioeconomic data, specific disease data, survival data, population health data, interview data, customer record data or observational study data, among others less common.

Regarding the type of data collected frequently the data hubs work with texts, numbers and files, but can also gather images, spreadsheets, tissue samples, sounds, or others less common.

Related to the level of aggregation of the data stored (individual vs. aggregated), the data hubs frequently present an individual level or both, but it also (although less common) can be aggregated only.

Data hubs use storage capacity in the range from MB up to PB. They can store up to billions of data sets/studies/collections. Like their storage capacity, they have a very varied annual growth, and can even reach several terabytes per year.

In relation to the number of sustained users submitting or storing data, this can range from one to millions. Similarly, the number of sustained users accessing data can range from one to millions.

Business processes

Related to how the data is compiled: stored in the data hub, data retrieval, loading, ETL methods, transforming or passing, among others, can be used. The storage can be supported by technologies such as SQL, relational database, Sorl, MongoDB, Oracle, Cloud data lakes, DataOntap, DICOM, XML, RDF, CSV, JSON, DBs, or a self-developed database/geographic information system. Data can be stored in several formats such as plain text, XML, or files (which are the most common), but also in others like JSON, DICOM, tsv, RDF, FASTA, Dublin core, Parquet, Nifti, FHIR, Oracle tables, OMOP Common Data Model, SAS Data Set, etc.

Data hubs usually apply quality controls to their data and require a minimum level of data quality to be included in the data infrastructure. Sometimes, a data hub applies quality controls only for internal use. Frequently, passing quality control is not mandatory for the data but the results of quality control are available when searching the data.

It is relevant for data hubs to use tools for checking errors and completeness of data. The most used is Checksum, but there are also many others such as HEX/SHACL, XSD Schemas, SQL-Scripts, R-dlookr, or even an automatic web-based check, a data submission portal and manual checks of certain variables or a specific software developed for the purpose of the network, or other options.

Data hubs with a low frequency use methods to check data source legitimacy, such as ISO standard, a Data Utility Framework, an accreditation of the data provider institute, an authentication of the data providing individual, quality/FAIRness/sustainability assessments, etc.

Related to how often the data sets are updated, this characteristic depends on each specific data set, and the most usual is to update annually, daily or irregularly, although they can also be updated monthly, weekly or even every 12 hours, among others. Another option is to perform a one time collection without updates.

Data hubs have processes to keep track of the different versions of datasets, such as manually creating versions by saving the date and name of each update, applying a different PID each time a version is stored, tracking model or software changes,

documented in the metadata management, or storing it in the log history. Also, each data type may have a different process for versioning.

On the subject of describing the logging and auditing of user actions, data hubs can time stamp the data deposition, time stamp the user contact to client service, and/or time stamp the user application to download or see the health data.

Data hubs commonly have a formal procedure to know who provides the data, practically materialised in contracts, agreements, regulations, terms of use, licence, accreditation - authentication, alliance membership, a law framework making formal requests for data collection mandatory approvals, records on data processing and provision, among others.

It is also important to highlight that data hubs frequently establish standard operating procedures (SOPs) that the organisation follows and updates regularly.

It is highly recommended for data hubs to include in their websites a Data Governance section describing the used data governance model, it can be in the form of a detailed document or in a paragraph.

ELSI aspects

Regarding ethical aspects, before accepting new submissions data hubs may require ethical approval for data to be stored on the infrastructure. After receiving the ethical approval the submission can be done.

Regarding anonymisation and pseudonymisation of data, data hubs usually use anonymisation methods with the data. The data can arrive already anonymised, which is not the most common. Additionally, the data hub itself can be in charge of anonymisation. The process can be done at the point of collection, before sharing it externally (these two are the most common), before sharing it internally, or at the point of publication. Almost all data hubs pseudonymise their data, this can be done by the data hub itself or by another external organisation.

Related to the legal aspects, when a data requester asks to access data and a data provider accepts the specific request, data hubs may offer a DAA (Data Access Agreement) to be signed between data providers and data requesters. It also can be done by a data permission or by accepting use policy. Data hubs may have a DPA (Data Processing Agreement) to be signed with the Data providers but it also can be by accepting use policy or to depend on contracting situations. Besides, data hubs may have a DPIA (Data Protection Impact Assessment) model.

Data hubs usually implement mechanisms to control the access of the data (authentication and authorisation) such as authorisation with web services backed by a database, OAuth2, OpenID Connect (over HTTPs), or other options.

Most of the data hubs have a funding sustainability plan. The data hub can receive national funding (it is the most common), or international, regional, from a hospital, European, related to participation in projects, international, or private fundings.

Data hubs receive data from different sources, providing a catalogue of these different data sources. Data is shared through a website, a secure data exchange portal, APIs, FTP, SFTP, DICOM transfer, among other options. This characteristic can depend on the specific usage request.

Related to societal impact aspects, the socioeconomic coverage of the data infrastructure can be NUTS1, NUST2 or/and NUTS3. Usually, the data hubs included in NUTS 2 are also included in NUTS1 and NUTS3.

Data hubs have national rules under the GDPR, such as data protection articles, rules for secondary use of data or laws for research organisations or health data documentation, among others. Data hubs also have documentation on data policy, licensing model and terms of use.

4.1. Profiling kinds of data hub organisation

Specific profile: data hubs managed centralised

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: these kinds of data hubs control the data stored, and provide health data from "General population" using "Text" and "Numbers" as types of data. In addition, the centralised data hubs receive and store data from a single source and/or from multiple sources.

Business processes: the centralised data hubs apply data quality control and establish standard operating procedures (SOPs) that the organisation follows and updates regularly, and use a formal procedure to know who provides the data.

ELSI aspects: this kind of data hubs pseudonymise data, and require legal approval for the data.

Specific profile: data hubs managed decentralised

Actors: may not have a single data controller, and may not have a data management strategy.

Data aspects: the decentralised data hubs are digital technical infrastructures with the core mission of enabling health data sharing that allows the discovery (findability) of health datasets. The data come from "Patient groups", "General population", and "Experimental settings". These data are stored in "XML" format, and are "Text", "Images", "Numbers".

Business processes: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Business processes).

ELSI aspects: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > ELSI aspects).

At a glance

	Data hubs managed centralised	**Data hubs managed decentralised**
Actors	No peculiarities.	No single data controller. No data management strategy.
Data aspects	Control the data stored. Data from "General population". Use "Text", "Numbers". Receive and store data from: single source, multiple sources.	Data from "Patient groups", "General population", "Experimental settings". Use "Text", "Images", "Numbers". Data stored in "XML".
Business processes	Data quality control. SOPs. Procedure to know who provides data.	No peculiarities.
ELSI aspects	Pseudonymised data. Require legal approval.	No peculiarities.

Table 1: Profiles depending on the kind of data hub organisation

4.2. Profiling roles

Specific profile: data hubs acting as data controller

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: the data hubs acting as data controllers manage centrally, pseudonymise data, have control over the data stored and receive data from a single source and/or multiple sources. Data are from "General population" and are "Text" type.

Business processes: the data hubs with this role have a process to keep track of the different versions of the datasets and use a formal procedure to know who provides the data.

ELSI aspects: establishes SOPs that the organisation follows and updates regularly, and provides a catalogue of the different data sources.

Specific profile: data hubs acting as data processor

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: a data hub acting as data processor is managed centrally, and is a digital platform that receives and stores the data. In addition, it has an authorisation functionality provided by the organisation itself or by an external institution, and a data accessibility mechanism in accordance with existing regulations.

Business processes: this kind of data hub uses a formal procedure to know who provides the data.

ELSI aspects: the data hubs with this role have pseudonymised data, and have established SOPs that the organisation follows and updates regularly.

At a glance

	Data hubs acting as data controller	Data hubs acting as data processor
Actors	No peculiarities.	No peculiarities.
Data aspects	Managed centrally. Pseudonymised data. Receive and store data from: single source, multiple sources. Data from "General population". Use "Text".	Managed centrally. Receives and stores the data. Functional authorisation. Data accessibility mechanism.
Business processes	Procedure to keep track of datasets versions. Procedure to know who provides data.	Procedure to know who provides data.
ELSI aspects	SOPs. Catalogue of data sources.	SOPs. Pseudonymised data.

Table 2: Profiles depending on the role

4.3. Profiling geographical coverage

Specific profile: European data hubs

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: data are from “General population”, and the type of data used is “Numbers”.

Business processes: the European data hubs have a formal procedure to know who provides the data.

ELSI aspects: this kind of data hub has established SOPs that the organisation follows and updates regularly.

Specific profile: worldwide data hubs

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: the worldwide data hubs have a data accessibility mechanism in accordance with existing regulation, and data source type is “Socioeconomic data”.

Business processes: this kind of data hub has a formal procedure to know who provides the data.

ELSI aspects: the data hub provides a DAA, has a DPA, and provides a catalogue of the different data sources.

At a glance

	European data hubs	Worldwide data hubs
Actors	No peculiarities.	No peculiarities.
Data aspects	Data from "General population". Use "Numbers".	Data accessibility mechanism. Data source: "Socioeconomic data".
Business processes	Procedure to know who provides data.	Procedure to know who provides data .
ELSI aspects	SOPs.	DAA. DPA. Catalogue of data sources.

Table 3: Profiles depending on geographical coverage

4.4. Profiling source of the data

Specific profile: data hubs using EHRs as one type of data source

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: the data hubs that use EHRs as one type of data source are managed centrally, receive data from a single source and/or multiple sources, provide health data from different sources, and have a data accessibility mechanism in accordance with existing regulation. Data comes from "General population". The type of data source used is "Healthcare data". The type of data collected is "Text" and/or "Numbers".

Business processes: these kinds of data hubs apply quality control, and have a formal procedure to know who provides the data.

ELSI aspects: these kinds of data hubs anonymise data, have pseudonymised data, receive national funding, and provide a catalogue of the different data sources.

Specific profile: data hubs using administrative data as one type of data source

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: the data hubs that use administrative data as one type of data source are managed centrally, receive data from a single source and/or multiple sources, have control over the data stored, provide health data from different sources, and have a data accessibility mechanism in accordance with existing regulation. Data comes from "General population", and the type of data used is "Electronic Health Records" and/or "Healthcare data". In addition, the types of data collected are "Text" and/or "Numbers".

Business processes: these data hubs apply data quality control, and have a formal procedure to know who provides the data.

ELSI aspects: these data hubs anonymise data, have pseudonymised data, and receive national funding.

Specific profile: data hubs using registry as one type of data source

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: the data hubs that use registry as one type of data source are managed centrally, receive data from a single source and/or multiple sources, and have control over the data stored. Data comes from "General population", and the type of data collected are "Text" and/or "Numbers".

Business processes: these data hubs apply data quality control, have a formal procedure to know who provides the data, and provide a catalogue of the different data sources.

ELSI aspects: these data hubs have pseudonymised data, and have SOPs that the organisation follows and updates regularly.

Specific profile: data hubs using specific disease data as one type of data source

Actors: no peculiarities found compared to those described in the general pattern of data governance (Section 4 > Actors).

Data aspects: the data hubs that use specific disease data as one type of data source are managed centrally, receive data from a single source and/or multiple sources, and have control over the data stored. In addition, they provide health data from different sources, allow discovery (findability) of health datasets, and have a metadata discovery service. Also, they have a data accessibility mechanism in accordance with existing regulation. The types of source used are "Electronic health records" and/or "Administrative data", and the types of data collected are "Text" and/or "Numbers".

Business processes: these data hubs apply quality control, use time stamp of data deposition for the logging and auditing of user actions, have a formal procedure to know who provides the data, and provide a catalogue of the different data sources.

ELSI aspects: these data hubs anonymise data, have pseudonymised data, receive national funding, and from the perspective of where the data is stored receive data from different sources. In addition, they established SOPs that the organisation follows and updates regularly.

At a glance

	Data hubs using EHRs	**Data hubs using administrative data**
Actors	No peculiarities.	No peculiarities.
Data aspects	Managed centrally. Receive data from: single source,	Managed centrally. Receive data from: single source, multiple

	multiple sources. Provide health data. Data accessibility mechanism. Data from "General population". Data source: "Healthcare data". Use "Text", "Numbers".	sources. Control the data stored. Provide health data. Data accessibility mechanism. Data from "General population". Data source: "Electronic Health Records", "Healthcare data". Use "Text", "Numbers".
Business processes	Data quality control. Procedure to know who provides the data.	Data quality control. Procedure to know who provides the data.
ELSI aspects	Anonymised and pseudonymised data. National funding. Catalogue of data sources.	Anonymised and pseudonymised data. National funding.

Table 4: Profiles depending on the source of the data (I)

	Data hubs using registry	**Data hubs using specific disease data**
Actors	No peculiarities.	No peculiarities.
Data aspects	Managed centrally. Receive data from: single source, multiple sources. Control the data stored. Data from "General population". Use "Text", "Numbers".	Managed centrally. Receive data from: single source, multiple sources. Control the data stored. Provide health data. Allow discovery of health datasets. Metadata discovery service. Data accessibility mechanism. Data source: "Electronic health records", "Administrative data". Use "Text", "Numbers".
Business processes	Data quality control. Procedure to know who provides the data. Catalogue of data sources.	Data quality control. Time stamp of data deposition. Procedure to know who provides the data. Catalogue of data sources.
ELSI aspects	Pseudonymised data. SOPs.	Anonymised and pseudonymised data. National funding. Data from different sources. SOPs.

Table 5: Profiles depending on the source of the data (II)

4.5. Key performance indicators

The proposed patterns of governance (the general one described in Section 4; and the specific profiles described in Sections 4.1 to 4.4) may be complemented with key performance indicators (KPIs).

This is a first draft list of KPIs.

- **Average time for approval** (in days). Average time between the day the request is posted including the mandatory documentation, and the final

answer is given obtaining the approval (intermediate answers or requests for clarifications are not considered accountable) (e.g. 7 days / 14 days / more than 14 days).

- **Average time for access** (in days). Average time between the day the request is posted, and the final answer is given accessing the health data (intermediate communications for clarifications are not considered accountable) (e.g. 7 days / 30 days / more than 14 days).
- **Cost**. Average cost related to accessing health data (several ranges can be defined to cover the different situations in terms of rate per hours / a month / a year; or in terms of normal / extensive data permit).
- **Data quality controls**. Minimum levels of quality of the data (results from quality controls) needed for the data to be included in the data infrastructure.

This suggested list of KPIs have to be reviewed with a specific data hub to check if these ideas are relevant and feasible for the specific data hub, removing, updating, and/or adding KPIs if it is needed, and defining the excellent/medium/minimum performance levels. The final list of KPIs created in collaboration with the specific data hub taking into account the peculiarities of the specific governance model, could be applied to analyse the related performance.

4.6. List of used terms

➢ Actor. Person participating to carry out a specific data governance model with a specific role.

➢ Anonymisation. The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject [1].

➢ Business process. Mechanism part of a specific data governance model. Every data hub has similar processes like accepting new submissions, getting submissions done, applying quality control, publishing the dataset for discovery, accepting requests, etc.

➢ Data Access Agreement or DAA. Negotiable under agreement or non-negotiable document to specify the agreed terms between data access provider and data processor in terms of accessibility.

➢ Data aspects. Data characteristics part of a specific data governance model.

➢ 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 [1].

➢ Data governance. Assembly of policies and processes, coordination aspects, data usage and accessibility principles and data management procedures for a certain health data infrastructure to ensure legal compliance, consistency

and good data quality throughout the different stages of the data life cycle [1].

- Data governance model. List of data governance characteristics applied by a specific existing data hub.
- Data hub. Data infrastructure that fulfils the following minimal inclusion criteria: (i) A digital technical infrastructure with the core mission of enabling health data sharing. (ii) It provides health data from different sources. (iii) It allows discovery of health datasets. (iv) It has a metadata discovery service. (v) It has a data accessibility mechanism in accordance with existing regulation. (vi) It has an authorisation functionality, provided by the same data hub or by an external institution [1].
- Data Processing Agreement or DPA. Negotiable under agreement or non-negotiable document to specify the agreed terms between data access provider and data processor in terms of processing.
- 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" [1].
- Data Protection Impact Assessment model or DPIA model. Document to carry out the process to identify and minimise the data protection risks through data sharing.
- Data provider. An entity which makes data available for secondary use [1].
- ELSI aspects. Characteristics related to Ethical, Legal, and Societal Impact issues.
- General Data Protection Regulation or GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.
- Level of aggregation. Characteristic of the data contained in a specific data infrastructure describing if the data was stored in an individual or aggregated way.
- NUTS: Used to analyse the socioeconomic coverage, 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. NUTS 2: basic regions for the application of regional policies. NUTS 3: small regions for specific diagnoses.
- Pattern of data governance. List of commonalities identified after analysis of a list of specific governance models of existing data hubs.

- Pseudonymisation. The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person [1].
- Standard Operating Procedures or SOPs. Consistent description of the mandatory operational steps to be followed related to processes or policies.
- Key performance indicators. Measurements that can be applied in the case of the governance model of a specific data hub to analyse the related performance.

5. Conclusions

This deliverable **D4.1 'Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance'** covers recommendations for integration in the HealthyCloud ecosystem, including a definition of health data hub patterns of data governance: one general (Section 4), and some specifics (Sections 4.1 to 4.4); derived from an analysis of the governance models of the interviewed data hubs. Accommodating data hubs with different governance models was essential to enable the decentralised ecosystem for health research across Europe.

The work covered in this deliverable will be complemented with deliverable **D4.2 'Report on current discoverability solutions and FAIR adoption level'** that will be delivered in December 2022 through the execution of the related task **T4.2 'Analysis of the data hubs operation related to the reference guidelines defined in WP3'** describing the tools, methods, process, procedures and/or mechanisms for data access and discovery (using -completely or partially- the FAIR principles) used by the interviewed data hubs. D4.2 will be strongly connected with **MS4.3 'Study: data hubs usage current metrics'** [4] that also covered a preliminary analysis of the survey answers, but focusing on the data hubs usage and the FAIR metrics, as well as with **MS3.3 'Guidelines: standardised guidelines for FAIRness maturity levels completed'** [5] that provides a set of clear guidelines to apply the FAIR principles in health data infrastructures at European level. The questions included in the sections Metadata, Findable, Accessible, Interoperable, Reusable in the survey (Annex 1) have not been analysed in this deliverable D4.1 (will be explored in T4.2/D4.2).

In terms of **limitations in the execution of the study**, it is relevant to mention the difficulties to identify the list of representative data hubs, due to the inexistence of a repository of contacts for the representative data hubs in Europe. Finally a robust list of 99 data hubs was used. Additionally, the participation of the data hubs through a survey was not easy due to availability matters, finally 41 of 99 contacted data hubs (41%) answered the survey. In terms of analysing the responses: in the case of non-mandatory questions some data hubs did not fulfil some questions (detailed information about the number of non-empty responses by question has been included in Annex 2), 65% (35 of 54) questions offered the possibility to include free text (directly answering the question, or through the 'Others' option in a structured question) adding a subjective interpretation in the analysis, 4 of these 35 free-text questions asked for URLs linking to a lot of material to explore, and some free text responses could not be used due to problems in the interpretation (e.g., an estimation of size specifying the number without specifying the unit).

Special attention was put in understanding potential **limitations and constraints in existing governance models**, and here the reflections are described. Most of the data hubs include related costs to access the data as part of their data governance model. This limitation slows down the progress in Open Science. The time spent for

ethical approval and for accessing the data itself is a constraint in the final use of the data. In some cases, the absence of a sustainability plan was identified, this fact endangers continuity of the data infrastructures. To ensure working in a secure environment, anonymisation and/or pseudonymisation methods, and logging and auditing mechanisms including access control mechanisms (authentication and authorisation) must be used. And finally, it is relevant to mention that, in order to have high quality data, tools, processes or methods must be applied in terms of errors checking, completeness, versions tracking, legitimacy. Not all data hubs cover these kinds of mechanisms.

5.1. Conclusions from survey analysis and results

The patterns of governance described in Section 4, together with constraints and limitations described in the last paragraph of Section 5, will be an **essential contribution** to the ready-to-implement **roadmap** for the HealthyCloud ecosystem that will be defined in the task **T4.5 'Development of a roadmap for the coordination of health data hubs in the context of the future HRIC ecosystem'**. Therefore, the recommendations included in this subsection will be completed during task T4.5.

First of all, the issue to be solved is: **"I manage a data hub and we want to integrate it in HealthyCloud. What to do?"**. From the analysis (Section 3) and results (Section 4) sections, we suggest the following steps:

1. **Identify the certain kind of data hub** (according to the given classification). Section 4 includes a general pattern of data governance for data hubs, using the findings obtained in the in-depth analysis of the 41 survey responses. Specific profiles have been defined generating specific patterns of data governance for data hubs. Concretely, for each pattern of data governance, data aspects, business models, and ELSI aspects have been defined, preceded by the list of actors involved in these processes.
    a. Kind of data hub organisation
        i. If you are a centralised data hub, please, consider Section 4.1 > Specific profile: data hubs managed centralised.
        ii. If you are a decentralised data hub, please, consider Section 4.1 > Specific profile: data hubs managed decentralised.
    b. Role
        i. If you act as data controller, please, consider Section 4.2 > Specific profile: data hubs acting as data controller.
        ii. If you act as data processor, please, consider Section 4.2 > Specific profile: data hubs acting as data processor.
    c. Geographical coverage
        i. If you manage European data, please, consider Section 4.3 > Specific profile: European data hubs.

ii. If you manage worldwide data, please, consider Section 4.3 > Specific profile: worldwide data hubs.

d. Source of the data
   i. If you use EHRs as one type of data source, please, consider Section 4.4 > Specific profile: data hubs using EHRs as one type of data source.
   ii. If you use administrative data as one type of data source, please, consider Section 4.4 > Specific profile: data hubs using administrative data as one type of data source.
   iii. If you use registry as one type of data source, please, consider Section 4.4 > Specific profile: data hubs using registry as one type of data source.
   iv. If you use specific disease data as one type of data source, please, consider Section 4.4 > Specific profile: data hubs using specific disease data as one type of data source.

2. **Comply to these given recommendations and KPIs**. The suggested recommendations are included in Section 4 as features of the data hubs for its integration in HealthyCloud. Likewise, a proposed list of KPIs has been added to be reviewed with data hubs and check if these ideas are relevant and feasible, removing, updating, and/or adding KPIs if it is needed.

5.2. Recommendations for integration in HealthyCloud

This subsection gathers a first draft of the recommendations proposed for the integration of a specific data hub in HealthyCloud.

A first list of **the most relevant recommendations** for integration in HealthyCloud is drafted in the following table:

Recommendation	**Description/Example**
Configure your data hub in a centralised way	That is, it requires a connection process whom data hub receives and stores the data directly. For example, a specific data hub have the control of the data stored and can receive and store data from a single source and/or from multiple sources.
Complete and sign a Data Processing Agreement (DPA)	The DPA includes the data use policy and contracting situations, as well as the agreed terms between data access provider and data processor in terms of processing.
Apply mechanisms of quality control to the data	For instance, a data hub can include data only if it reaches a certain quality level or perform data quality controls for internal use.

Define a formal procedure to find out who provides the data	In this sense, for data management it is relevant to know who provides the data through a formal procedure (i.e. legal contracts, agreements, or open information in the organisation).
Provide a catalogue of the different data sources	For example, that catalogue is really useful in the case of a data hub that connects to several data sources.
Apply anonymisation and/or pseudonymised methods	For instance, in the case of health data hubs that do not receive anonymised data, anonymisation and/or pseudonymised methods are recommended as applicable in order to comply with GDPR rules.
Use some tool to check for errors and data integrity	This recommendation is included because checking for errors and completeness is another important aspect of data quality in data hubs. For example, tools like Checksum, HEX/SHACL, XSD Schemas, SQL-Scripts, R-dlookr, or even an automatic web-based check, a data submission portal and manual checks of certain variables or a specific software developed for the purpose of the network.
Include in the data hub website a Data Governance section describing the used data governance model	Important information related to data governance model or data management can be provided by data hubs through their websites.

Table 6: The most relevant recommendations for integration in HealthyCloud

As was detailed in the second paragraph of Section 5 > Conclusions, these recommendations will be complemented with tools, methods, process, procedures and/or mechanisms for data findability, access, interoperability and discovery in deliverable **D4.2 'Report on current discoverability solutions and FAIR adoption level'**.

6. References

[1] Glossary of commonly used terms in the field of health data research - developed by the EU project HealthyCloud: https://zenodo.org/record/5998128#.YhOMWDHMKUm

[2] HealthyCloud MS4.1 – Community activity: selection of representative data hubs.

[3] HealthyCloud MS4.2 – Study: patterns of governance of selected data hubs.

[4] HealthyCloud MS4.3 – Study: data hubs usage current metrics.

[5] HealthyCloud MS3.3 – Guidelines: standardised guidelines for FAIRness maturity levels completed.

7. Annex 1: Survey

HealthyCloud			
ID	INDICATORS	Description of the indicator (example)	Format of the input
Part 1: Data			
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 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
		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.

ID	INDICATORS	Description of the indicator (example)	Format of the input
	Type of source	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

ID	INDICATORS	Description of the indicator (example)	Format of the input
	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 - NUTS 2: basic regions for the application of regional policies - NUTS 3: small regions for specific diagnoses	Multiple choice: / NUTS1 / NUTS2 / NUTS3 / I don't know
	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

ID	INDICATORS	Description of the indicator (example)	Format of the input
	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	Do you have a public metadata catalogue service?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
			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	Do you have a public data catalogue?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
			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

ID	INDICATORS	Description of the indicator (example)	Format of the input
	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	What is the format(s) for distributing data?	Multiple choice: / csv / xml / json / ld-json / pdf / R / SAS / Other / I don't know / This doesn't apply to this data infrastructure
		If other, please specify	Free text
	Metadata record	Do you have a metadata record API endpoint (m2m) in place?	Drop down menu: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
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

ID	INDICATORS	Description of the indicator (example)	Format of the input
Part 2: Governance / Management / data hub specific questions		**Please, could you answer the following questions if it is applicable to your case?**	
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? i.e. Data controller/Joint controller/Data processor/None of the above	add an option we have differnt roles in different situations, select multiple options
			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

8. Annex 2: Number of non-empty responses

This Annex details, by question, the number of non-empty responses used and analysed, out of the total of 41 survey responses received. Responses stating "Not applicable" have been counted as non-empty responses, because they help to understand that this characteristic is not applicable in the case of the specific data hub.

Questions	**Nº of non-empty responses (out of 41)**
Part 1: Data	
Administrative	
Title or name of the data infrastructure	41
Abbreviation or alternative title	36
Website of the data infrastructure	40
Who is the data controller organisation?	40
Contact details	35
Data access provider	40
Data processor	39
Which of the following characteristics fit your data infrastructure?	40
If your data infrastructure is part of a data hub, what is the name and URL of the data hub?	21
How is the data infrastructure organised?	40
Data	
Do you require ethical approval for the data to be stored in your infrastructure?	41
Does the data originate from a patient group, the general population or an experimental setting, or other?	40
What is the type of data source that you are using?	40
How is the data that is stored in the data infrastructure compiled?	40
Describe the technologies used for data storage.	38
What is the format in which the data is stored?	40
Specify the type of data collected	40
What is the level of aggregation of the data stored in this data infrastructure?	41
Are anonymisation methods used with the data?	40
Is the anonymisation performed by your data infrastructure and/or do	25

you receive already anonymised data?	
Do you have pseudonymised data?	40
If yes, who holds the method to reverse the pseudonymisation process?	33
Completeness of data infrastructure	
What is the geographical coverage of the data infrastructure?	40
What is the socioeconomic coverage of the data infrastructure?	39
What are the participating countries from which you have datasets?	41
When did your data infrastructure start collecting data?	38
Is the data collection period still ongoing?	35
What is the end date of the data collection period?	19
Data quality aspects	
Are data quality controls applied?	41
Are there minimum levels of quality of the data needed for the data to be included in the data infrastructure?	38
How often do you update the datasets?	40
Do you use a tool to check for errors and completeness?	41
If yes, what tool do you use	22
Do you have a process to keep track of the different versions of the datasets?	41
If yes, please specify the process.	20
Do you have a method to check data source legitimacy? Please specify.	27
<u>Part 2: Governance / Management / data hub specific questions</u>	
Technical	
How much storage capacity is in use up to date?	38
Until today, how many datasets are stored in your data collection, or studies/data collections stored in your data hub?	33
What is the estimated annual growth of the data infrastructure in size or number of datasets?	32
Number of sustained users who submit or store data up to date	30
Number of sustained users who access data up to date	30
Legal aspects	
Are there any national rules additional to the GDPR in your country? If yes, which ones?	21
In the scope of the EU GDPR, what is your organisation's role in	39

relation to personal data? i.e. Data controller/Joint controller/Data processor/None of the above	
Please, describe the logging and auditing of user actions	35
Is there a formal procedure to know who provides the data?	37
If 'Yes', please specify the procedure (i.e. contracts, agreements, open information in the organization, etc)	26
Does the data hub provide a DAA to be signed between data providers and data requesters?	39
Does the data hub have a DPA to be signed with the Data providers?	38
Does the data hub have a DPIA model?	36
Has access control mechanism been implemented?	27
Has access control mechanism been implemented (authentication and authorization)?	26
What type of funding does the data hub receive?	37
What is the sustainability plan of the data hub funding?	31
Does the data hub provide a catalogue of different data sources?	34
From the perspective of where is the data stored. Does the data hub receive data from different sources?	34
Please, describe the services through which data is shared	31
Feel free to provide names and/or links to relevant documentation regarding Data Policy, License Model and Terms of Use.	16
Operational	
Do you have established SOPs that your organization follows and updates regularly?	34
Other comments	5