Desiderata for the governance of health data hubs for research

Celia Alvarez-Romero (celia.alvrom@gmail.com)
Instituto de Biomedicina de Sevilla
https://orcid.org/0000-0001-8647-9515

Alicia Martínez-García
Instituto de Biomedicina de Sevilla

Máximo Bernabeu-Wittel
Hospital Universitario Virgen del Rocío: Hospital Universitario Virgen del Rocio

Carlos Luis Parra-Calderón
Instituto de Biomedicina de Sevilla

Research Article

Keywords: Health data management, health data infrastructure, health data hub, patterns of governance, governance models, survey

Posted Date: December 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2321504/v1

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

Background

Digital transformation in healthcare and the growth of health data generation and collection include an important challenge for the secondary use of healthcare records in the health research field. Likewise, due to the ethical and legal constraints for using sensitive data, understanding how health data is managed by dedicated infrastructures called data hubs is essential to facilitate data sharing and reuse.

Methods

In order to capture the different data governance behind health data hubs across Europe, a survey focused on analysing the feasibility of linking individual-level data between data collections and the generation of health data governance patterns was carried out. The target audience of this study was National, European, and Worldwide data hubs. In total, the designed survey was sent to a representative list of 99 health data hubs in January 2022.

Results

In total, 41 survey responses received till June 2022 were analysed in-depth. Stratification methods were performed to cover the different levels of granularity identified in some data hubs’ characteristics. Firstly, a general pattern of data governance for data hubs was defined. Afterward, specific profiles were defined, generating specific patterns of data governance through the stratifications in terms of the kind of organisation (centralised vs. decentralised), and role (data controller or data processor) of the health data hubs interviewees.

Conclusions

The in-depth analysis of the responses from health data hubs interviewees across Europe provided a list of the most frequent aspects that concluded a set of specific recommendations on data management and governance, taking into account the constraints of sensitive data. In summary, a data hub should work in a centralised way providing a Data Processing Agreement and a formal procedure to identify data providers, as well as data quality control, data integrity and anonymisation methods.

Background

The study presented in this manuscript was carried out during the Coordination and Support Action HealthyCloud (Health Research & Innovation Cloud), which has received funding from the European Commission, started in March 2021 and will finish in August 2023.
The main aim of HealthyCloud [1] is to align all the knowledge and expertise in health data spread across European and international actors, as well as to lay the foundations for the future European Health Research and Innovation Cloud (HRIC) [2]. HRIC will become a fundamental part of the European Health Data Space (EHDS) [3]. The HRIC will enable the secondary use of data and the capabilities to analyse and share data to drive the limits of health research within an ethically and legally compliant framework that builds and reinforces the trust of patients and citizens.

The digitalization of health systems represents an essential opportunity for health research activities. An enormous amount of health-related data is now generated and collected within healthcare systems [4, 5]. Research networks have assembled and curated health data, at the patient and/or population levels, for multiple diseases in the form of cohorts. Besides, dedicated research infrastructures in Europe have long harmonised the collection and preservation of specific biological specimens and promoted the development of clinical trials, aiming to reuse the results by other researchers [6]. As the reuse of health data is a fast-growing field recognised as essential to realising the potential for high quality healthcare, improved healthcare management, reduced healthcare costs, population health management and effective health research [7]. However, the dispersed nature of data generation and the ethical and legal constraints for using sensitive data represent significant challenges that have strongly restricted the use of health data for research [8, 9]. The significant technical barriers in terms of the need for more structured information and limited interoperability between different health fields also affect the full exploitation of health data for research purposes. So, an appropriate health data management becomes fundamental [10, 11]. In addition, a recent study has concluded that funding agencies do not support data sharing mandates because of data protection regulations. Data sharing mandates are currently absent, which complicates assessing compliance of researchers with funding agencies’ policies and evidence production. Therefore, policy measures that restrict the authority of researchers to make data sharing decisions are often not supported. In this regard, incentive design is paramount if funding agencies do not wish to impose restrictions on the decision-making authority of researchers [12].

In this sense, HealthyCloud execution has included capturing different governance and auditing models behind data hubs across Europe and managing health data to analyse the existing initiatives related to domain-specific data hubs. For this purpose, the definition of important terms related to this study was discussed.

Firstly, in the HealthyCloud project, a health data hub is defined as a data infrastructure with the following minimal inclusion criteria [13]: (i) A digital technical infrastructure with the core mission of enabling health data sharing; (ii) It provides health data from a different source; (iii) It allows discovery of health datasets; (iv) It has a metadata discovery service; (v) It has a data accessibility mechanism following existing regulation; and (vi) It has an authorisation functionality, provided by the same Data Hub or by an external institution.

Secondly, HealthyCloud defines data governance 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” [13].

**Methods**

The study described in this manuscript covers an analysis of health data governance patterns generated after identifying commonalities in the governance models of existing data hubs. To in-depth understand how health data is managed by the dedicated infrastructures called data hubs [14, 15], a capture of the different data governance behind health data hubs across Europe was carried out. Existing initiatives and projects related to domain-specific data hubs at regional, national, European and Worldwide levels were analysed. After, a list of 99 representative data hubs in Europe was collected.

To gather the feedback from the representative data hubs collected, a collaborative survey was designed. The survey’s main objectives 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. The survey was developed in an electronic tool (Typeform.com), including the contributions and improvements detected by the HealthyCloud researchers. Finally, the survey was sent at the beginning of January 2022 to the target audience, which was the National, European, and Worldwide data hubs identified previously. Concretely, the survey was sent to a representative list of 99 data hubs (see Fig. 1), after the effort to ensure a robust representation of all the data hubs in Europe.

Finally, 41 out of the 99 (41%) contacted data hubs answered the survey until June 2022. Figure 2 shows the final geographical coverage achieved through the survey responses.

All the material collected through the survey was analysed (both structured and free-text questions), focusing on identifying actors, data aspects and business processes involved in the hubs’ governance, also considering ELSI (Ethical, Legal, Societal Impact) aspects.

To appropriately cover the different levels of granularity identified in some data hubs’ characteristics (such as 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), and the role applied in data management (data controller vs. data processor), delivering specific profiles.

Finally, a set of recommendations related to data governance patterns were identified by analysing the list of the most frequent aspects of data hubs interviewees.

**Survey analysis**

The in-depth details obtained during the analysis of the 41 survey representative responses are detailed below. 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

Apart from the characteristics defined for the health data hub concept [13], from the survey responses in a multiple-choice question: 27 respondents 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 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.

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

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 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 by 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% of the 25 data hubs. Concerning 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.

Data quality aspects

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”.

Another aspect related to data quality is checking for errors and completeness. 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.

Furthermore, 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.

Data management

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.

**Funding**
As part of the sustainability plan, the survey went in-depth about the type of funding and the sustainability plan of this current funding. In this case, 38 of 41 answers covered this question on the type of funding: national funding for the Hub core function (66%), participation in projects (16%), European or international funding (11%), and private funding (8%). Concerning the sustainability plan, 42% received stable funding (of which 39% stated this stable funding is of national origin), 13% presented funding from private profits (i.e. data licence fees, pay for customer use, etc.), 32% were 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 did not offer this kind of catalogue, because this specific data hub was connected only to a unique data source. And 79% of 34 provided 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.

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 these kinds of procedures.

**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 answered and 1 did not answer. Of these 40 responses, 30 (75%) answered “It is managed centrally”, 9 (22%) answered “It has a decentralised management”, and 1 (2%) “This does not apply to this data infrastructure”.

On “data hubs managed centralised”, 23 had control of the data stored. In addition, 25 received and stored data from a single source and/or from multiple sources. 90% pseudonymised data, 90% applied data quality control, 81% established standard operating procedures (SOPs) that the organisation followed and updated regularly, 89% there was a formal procedure to know who provided the data, and 83% required legal approval for the data.
On “data hubs managed decentralised”, they may not have a single data controller and may not have a data management strategy. 9 (all of them) allowed the discovery (findability) of health datasets and 8 were a digital technical infrastructure with the core mission of enabling health data sharing. 8 host data that came from "patient groups", 7 from "general population" and 7 from "experimental settings". In 7, the data was stored in "xml" format.

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 answered and 2 did not answer. 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”.

Regarding “data controller”, 82% were managed centrally, 100% pseudonymised data, 10 had control over the data stored, and 9 received data from a single source and/or multiple sources. 9 of them had data from "general population". 82% had a process to keep track of the different versions of the datasets and 90% had a formal procedure to know who provided the data. 81% established SOPs that the organisation followed and updated regularly and 81% provided a catalogue of the different data sources.

Regarding “data processor”, 80% were managed centrally, 80% had pseudonymised data, and 9 were a digital platform that received and stored the data. Furthermore, 90% had an authorisation functionality provided by the organisation itself or by an external institution, and 90% had a data accessibility mechanism in accordance with existing regulations. 91% had a formal procedure to know who provided the data, and 80% had established SOPs that the organisation followed and updated regularly.

Results

A general pattern of data governance for data hubs is defined below, using the conclusions obtained in the in-depth analysis of the 41 survey responses. 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, using the conclusions obtained in the stratifications in terms of the kind of organisation (centralised vs. decentralised), and role (controller or processor). For the specific patterns of data governance, it has been counted from 75% (prevalence). 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 fewer 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.

The most frequent aspects
After performing the in-depth analysis of the 41 responses of the survey, the most frequent aspects are listed below.

Concerning the simple-choice questions (with percentages): (i) Formal procedure to find out who provides the data (84%); (ii) Quality control is applied to the data (83%); (iii) A catalogue of the different data sources is provided (79%); (iv) There are Standard Operating Procedures (SOPs) that are followed and regularly updated (79%); (v) They receive health data from different sources (76%); (vi) The data infrastructure is centrally managed (75%); (vii) Data anonymisation methods are used (65%), using pseudonymised data (80%); and (viii) A tool is used to check for errors and data integrity (61%). Figure 3 shows a graphical representation of these percentages.

Regarding multiple-choice questions (with absolute values): (i) Data come from the general population (29) or from a group of patients (24); (ii) A data accessibility mechanism is available in accordance with current regulations (28); (iii) The coverage of the data infrastructure is national (27), receiving national funding (19); (iv) They provide health data from different sources (28); (v) They are a digital platform that receives and stores data (27); (vi) They allow the discovery (findability) of health data sets (26); (vii) They have control over stored data (26); (viii) Enable discoverability of health data sets (26); (ix) They have authorisation functionality, provided by the organisation itself or by an external institution (25); and (x) The type of data source used is the electronic health record (EHR) (25). Figure 4 shows a graphical representation of these percentages.

Identifying common aspects involved in the data hubs’ governance

Below, a general pattern of data governance for data hubs is presented defining common aspects involved in the data hubs’ governance models, using the conclusions obtained in the in-depth analysis of the 41 survey responses, and taking into account the list of more frequent aspects. Data aspects, business models, and ELSI aspects are defined, preceded by the list of actors involved in these processes.

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” [13]. 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 processor determines who is in charge of data processing, “which processes personal data on behalf of the controller” [13]. 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.

Regarding the organisation's role in personal data, data hubs can be data controllers or data processors. They also can have both roles depending on the specific situation.
The data access provider is defined as “an entity which makes data available for secondary use” [13]. There may be one or several, it may be a person or a set of mechanisms. 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 can be found.

**Data aspects**

Concerning data characteristics that are frequently present in the data hubs, these kinds of data infrastructures usually: (i) are digital platforms that receive and store data, (ii) have control over the stored data, receive data from a single source and/or multiple sources, (iii) 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, and (iv) 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.

Regarding 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 sources are EHRs, administrative data, registry data, and healthcare 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.

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.

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 can 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.

**Business processes**

Related to how the data is compiled and 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 low-frequency use methods to check data source legitimacy, such as a Data Utility Framework, 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**

Concerning 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.

Related to anonymisation and pseudonymisation of data, data hubs usually use anonymisation methods. 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 data permission or by accepting a 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 depending 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.

**Profiling kinds of data hub organisation**

Analysing by subgroups, two profiles were proposed, depending on the kind of data hub organisation: “data hubs managed centralised” and “data hubs managed decentralised”. The specifications or peculiarities of these profiles compared to those described in the general pattern of data governance are presented in Table 1.

Table 1. Profiles depending on the kind of data hub organisation.
### Profiling roles

Analysing by subgroups, two profiles were proposed depending on the role of the data hub: “data hubs acting as data controller” and “data hubs acting as data processor”. The specifications or peculiarities found of these profiles compared to those described in the general pattern of data governance are presented in Table 2.

Table 2. Profiles depending on the role performed by the data hub.

<table>
<thead>
<tr>
<th></th>
<th>Data hubs managed centralised</th>
<th>Data hubs managed decentralised</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actors</strong></td>
<td>No peculiarities.</td>
<td>May not a single data controller.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No data management strategy.</td>
</tr>
<tr>
<td><strong>Data aspects</strong></td>
<td>Control the data stored.</td>
<td>Data from &quot;Patient groups&quot;, &quot;General population&quot;, &quot;Experimental settings&quot;.</td>
</tr>
<tr>
<td></td>
<td>Data from &quot;General population&quot;.</td>
<td>Use &quot;Text&quot;, &quot;Images&quot;, &quot;Numbers&quot;.</td>
</tr>
<tr>
<td></td>
<td>Use &quot;Text&quot;, &quot;Numbers&quot;.</td>
<td>Data stored in &quot;XML&quot;.</td>
</tr>
<tr>
<td></td>
<td>Receive and store data from: single source, multiple sources.</td>
<td></td>
</tr>
<tr>
<td><strong>Business processes</strong></td>
<td>Data quality control.</td>
<td>No peculiarities.</td>
</tr>
<tr>
<td></td>
<td>SOPs.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Procedure to know who provides data.</td>
<td></td>
</tr>
<tr>
<td><strong>ELSI aspects</strong></td>
<td>Pseudonymised data.</td>
<td>No peculiarities.</td>
</tr>
<tr>
<td></td>
<td>Require legal approval.</td>
<td></td>
</tr>
</tbody>
</table>
### Data hubs acting as data controller

<table>
<thead>
<tr>
<th>Actors</th>
<th>No peculiarities.</th>
</tr>
</thead>
</table>
| Data aspects | Managed centrally.  
Pseudonymised data.  
Receive and store data from: single source, multiple sources.  
Data from "General population".  
Use "Text".  |
| Business processes | Procedure to keep track of datasets versions.  
Procedure to know who provides data.  |
| ELSI aspects | SOPs.  
Catalogue of data sources.  |

### Data hubs acting as data processor

<table>
<thead>
<tr>
<th>Actors</th>
<th>No peculiarities.</th>
</tr>
</thead>
</table>
| Data aspects | Managed centrally.  
Receives and stores the data.  
Functional authorisation.  
Data accessibility mechanism in accordance with existing regulations.  |
| Business processes | Procedure to know who provides data.  |
| ELSI aspects | SOPs.  
Pseudonymised data.  |

### Discussion

Recent advances in big data are expected to expand our knowledge to test new hypotheses about disease management, from diagnosis to prevention to personalised treatment. However, the rise of big data also poses challenges in terms of privacy, security, data ownership, data stewardship and governance [16]. Besides, the wide availability of data has led to the need for additional attention to the health research field, where the number of studies seeking to leverage data to improve healthcare has grown significantly. Healthcare data are increasingly complex and are obtained in a variety of ways, from a variety of sources, contexts and technologies, and their nature can impede proper analysis. Any analytical research must overcome these obstacles to extract data and produce meaningful insights. Hence the importance of investigating the main challenges, data sources, techniques and technologies, as well as future directions in the field of big data analytics in healthcare [17].

The basis of the study performed has been that accommodating data hubs with different governance models is essential to enable the decentralised ecosystem for health research across Europe. Health data reuse is widely used in healthcare, research, government and business settings. Studying the benefits, the barriers to use with large clinical databases, the policy frameworks that have been formulated, and the challenges makes the study of data management and governance essential to promote the data sharing and reuse [18, 19]. To address the purpose of this study, a definition of health data management [20] and health data hub patterns of data governance was covered through the analysis in-depth of a dedicated survey of a representative list of National, European, and Worldwide health data hubs.
On the one hand, a general pattern of data governance for data hubs was defined using the findings obtained in the analysis of the 41 survey responses, detailing the most frequent aspects of health data hubs analysed and identifying actors and business processes involved in the data hubs’ governance. On the other hand, specific patterns of data governance were generated through the stratifications in terms of the kind of organisation (centralised vs. decentralised), and role (controller or processor). Specific profiles were defined including the actors involved, data and ELSI aspects, and business processes.

In addition, this is the first study that presents relevant recommendations on data management and governance taking into account the information provided by health data hubs, through the evaluation of the survey responses.

Particular attention was paid to understanding the potential limitations and constraints of existing governance models that resulted in a number of breakthroughs in the medical field [21, 22, 23]. 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 [24, 25, 26]. 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 the 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 error checking, completeness, version tracking, and legitimacy. Not all data hubs cover these kinds of mechanisms.

In terms of limitations in the study execution, 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. Additionally, the participation of the data hubs through a survey was not easy due to availability matters (41% of the contacted data hubs answered the survey). In terms of analysing the responses, in the case of non-mandatory questions, some data hubs did not fulfil some questions, 35 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).

Conclusion

The findings gathered in the in-depth analysis of the survey responses facilitated a list of recommendations proposed for health data hubs. Specific profiles generating specific patterns of data governance for health data hubs were defined in this study. Furthermore, it is relevant to highlight that the governance models discovered in this study were validated with the health data hubs interviewees, involving them in the review phase of the governance patterns.
The most relevant recommendations on data management and governance that must be considered together with the constraints of sensitive data were identified by analysing the list of the most frequent aspects of data hubs interviewees. They are drafted in Table 3.

Table 3. The most relevant recommendations on data governance for health data hubs.

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Description/Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configure your data hub in a centralised way</td>
<td>That is, it requires a connection process for whom the data hub receives and stores the data directly. For example, a specific data hub has the control of the data stored and can receive and store data from a single source and/or from multiple sources.</td>
</tr>
<tr>
<td>Complete and sign a Data Processing Agreement (DPA)</td>
<td>The DPA includes the data use policy and contracting situations, as well as the agreed terms between the data access provider and data processor in terms of processing.</td>
</tr>
<tr>
<td>Apply mechanisms of quality control to the data</td>
<td>For instance, a data hub can include data only if it reaches a certain quality level or performs data quality controls for internal use.</td>
</tr>
<tr>
<td>Define a formal procedure to find out who provides the data</td>
<td>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).</td>
</tr>
<tr>
<td>Provide a catalogue of the different data sources</td>
<td>For example, that catalogue is really useful in the case of a data hub that connects to several data sources.</td>
</tr>
<tr>
<td>Apply anonymisation and/or pseudonymised methods</td>
<td>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 [27].</td>
</tr>
<tr>
<td>Use any tool to check for errors and data integrity</td>
<td>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.</td>
</tr>
<tr>
<td>Include in the data hub website a Data Governance section describing the used data governance model</td>
<td>Important information related to the data governance model or data management can be provided by data hubs through their websites.</td>
</tr>
</tbody>
</table>

**Abbreviations**

DAA: Data Access Agreement
DPA: Data Processing Agreement
DPIA: Data Protection Impact Assessment
EHDS: European Health Data Space
EHR: Electronic Health Records
ELSI: Ethical, Legal, Societal Impact
GDPR: General Data Protection Regulation
HealthyCloud: Health Research & Innovation Cloud
HRIC: Health Research and Innovation Cloud
SOPs: Standard Operating Procedures

**Declarations**

**Funding**

This work was supported by the European Union’s Horizon 2020 research and innovation programme Coordination and Support Action HealthyCloud (grant agreement no. 965345) [1]. Also, this research has been co-supported by the Carlos III National Institute of Health, through the IMPaCT-Data programme (code IMP/00019) [28], and through the Platform for Dynamization and Innovation of the Spanish National Health System industrial capacities and their effective transfer to the productive sector (code PT20/00088), both co-funded by European Regional Development Fund (FEDER) ‘A way of making Europe’.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**References**
1. HealthyCloud website: https://healthycloud.eu/
13. Glossary of commonly used terms in the field of health data research - developed by the EU project HealthyCloud: https://doi.org/10.5281/zenodo.5997584


28. IMPaCT-Data website: https://impact-data.bsc.es/

Figures
Figure 1

Map with contacted data hubs.
Figure 2

Map with number of received responses from contacted data hubs.
Figure 3

The most frequent aspects (simple-choice questions with percentages)
Figure 4

The most frequent aspects (multiple-choice questions with absolute values)