Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230452

Desiderata for the Data Governance and FAIR Principles Adoption in Health Data Hubs

Celia ALVAREZ-ROMERO[a,1], Silvia RODRÍGUEZ-MEJIAS[a], Carlos Luis PARRA-CALDERÓN[a]
[a] *Computational Health Informatics Group, Institute of Biomedicine of Seville, IBiS / Virgen del Rocío University Hospital / CSIC / University of Seville, Seville, Spain*
ORCiD ID: Celia ALVAREZ-ROMERO https://orcid.org/0000-0001-8647-9515
Silvia RODRIGUEZ-MEJIAS https://orcid.org/0009-0009-9916-4729
Carlos Luis PARRA-CALDERÓN https://orcid.org/0000-0003-2609-575X

Abstract. The objective of this study, as part of the European HealthyCloud project, has been to analyse the data management mechanisms of representative data hubs in Europe and identify whether they comply with an adequate adoption of FAIR principles that will enable data discovery. A dedicated consultation survey was performed, and the analysis of the results allowed to generate a set of comprehensive recommendations and best practices so that these data hubs can be integrated into a data sharing ecosystem such as the future European Health Research and Innovation Cloud.

Keywords. Health data hubs, data governance, FAIR principles, discoverability, European Health Research and Innovation Cloud, patterns of governance, survey.

1. Introduction

This study has been carried out during the HealthyCloud (Health Research & Innovation Cloud) Coordination and Support Action, supported by the European Union's Horizon 2020 research and innovation programme (grant agreement no. 965345) [1] and whose objective 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]. The purpose of this study has been to define a set of data governance recommendations together with a good practice guide in compliance with the FAIR principles that facilitate integrating data hubs into the future HRIC that will become a fundamental part of the European Health Data Space (EHDS) [3], enabling 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

[1] Corresponding Author: Celia Alvarez-Romero, celia.alvarez@juntadeandalucia.es, Avenue Manuel Siurot S/N - 41013 Seville (Spain), +34 95501333.

framework that reinforces the trust of patients and citizens. The approach of this survey study is based on reaching the goal of data management by following data governance guidelines and good practices to follow the FAIR principles [4] for European health data hubs to achieve an open science model while considering the protection of sensitive data under the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) [5].

2. Background

The digitisation of health systems represents an important opportunity for health research activities. Healthcare systems generate and collect enormous amounts of health-related data [6, 7]. Likewise, health data are collected and preserved for multiple diseases in the research field. So, dedicated research infrastructures have long been harmonising the collection and preservation of health data to enable other researchers to reuse the results [8] and taking into account the challenges of the dispersed nature of data generation and the ethical and legal constraints for using sensitive data [9, 10]. In this sense, HealthyCloud execution includes capturing governance models behind data hubs across Europe and managing health data to analyse the existing initiatives related to domain-specific data hubs. That is, dedicated data infrastructures with the following minimal inclusion criteria [11]: (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.

3. Methods

To define data governance patterns and guidelines for the adoption of FAIR principles in health data hubs, a dedicated consultation survey was designed in an electronic tool (typeform.com) and was conducted to understand in-depth how health data hubs manage health data and analyse commonalities in compliance with the FAIR principles of existing health data hubs. The survey performed included questions on data hubs criteria, data hubs' main features, data management, governance and legal aspects, data quality aspects, metadata, data findability, data accessibility, data interoperability and data reusability. Existing initiatives and projects related to domain-specific data hubs at regional, national, European and Worldwide levels were analysed and identified bases on previous experiences and contacts, as well as literature and internet searchers. After, a list of 99 representative data hubs around Europe, including worldwide data hubs, was collected. After, the survey was sent to them in January 2022.

4. Results

4.1. Recommendations for health data governance

Regarding the survey questions on data governance patterns and after a great effort to contact the identified data hubs, 42 out of the 99 (42%) data hubs answered the survey.

General analysis and stratifications were carried out following the specific patterns of the data hubs (type of organisation (centralised or decentralised) and role (data controller or data processor)), resulting in a set of the most frequent aspects that concluded specific recommendations on data management and governance, considering the ethical requirements and legal aspects of sensitive data. An in-depth analysis of the survey questions revealed the following recommendations:

Table 1. Recommendations for health data governance

Recommendations for data governance
Configure your data hub in a centralised way (77%).
Complete and sign a Data Processing Agreement (DPA) (47%).
Apply mechanisms of quality control to the data (83%).
Define a formal procedure to find out who provides the data (89%).
Provide a catalogue of the different data sources (81%).
Apply anonymisation (65) and/or pseudonymised methods (80%).
Use any tool to check for errors and data integrity (61%).
Include in the data hub website a Data Governance section describing the used data governance model.

From the data hubs that responded to the survey and, in compliance with the FAIR principles and the ethical and legal requirements of the sensitive data, we have obtained as main conclusions the set of recommendations presented in Table 1. 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.

4.2. Best practices for the adoption of FAIR principles by data hubs

Regarding questions survey on FAIR principles (findability, accessibility, interoperability and reusability aspects of data), 42/99 (42%) answered. General analysis and stratifications were carried out following the specific patterns of the data hubs (type of organisation (centralised or decentralised), role (data controller or data processor) and level of aggregation (individual, aggregated and both)), resulting in a set of frequent aspects that concluded specific best practices on data FAIR principles adoption. An in-depth analysis of the survey revealed the following best practices:

Table 2. Best practices for the adoption of FAIR principles by data hubs

Best practices for FAIR principles adoption
Have a persistent, unique identifier for your data (40%) and associated metadata (26%).
Produce or collect metadata for all data, placing them in a resource connected to the data identifiers (82%).
Have a public metadata catalogue (57%), providing a metadata record API endpoint (16%).
Establish a clear and public protocol to access and reuse individual and/or aggregated data for third parties (76%), complying with GDPR and any other relevant regulations, identify the Data Protection Officer and data controller contact.
Offer the possibility to extract the data from the data infrastructure (82%), or a safe space to analyse the data.
Encrypt sensitive data (55%) using an approved encryption protocol when data is stored (13/23) and/or transferred (20/23).
Use a standard for metadata and use a commonly used format for data distribution (98%).
Require ethical, privacy and/or legal approval for secondary use of sensitive data from the requestor, based on GDPR templates (e.g. DPA, DPIA) (26/42).

Table 2 shows a set of best practices obtained from data hubs answers according to the FAIR principles and the ethical and legal requirements of the sensitive data. In summary, a data hub should have a unique and persistent identifier for its metadata and data, have a public metadata catalogue and protocol to access and reuse data, encrypt sensitive data, use standards for data distribution and require ethical, privacy and/or legal approval for secondary use of sensitive data.

5. Discussion

To start this study, there were difficulties in identifying 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. The data hubs' participation in the survey was difficult due to availability matters, although 42 of 99 contacted data hubs (42%) answered the survey.

6. Conclusion

This manuscript describes the first study to present relevant recommendations on data management and governance considering the information provided by health data hubs by assessing the survey responses. Regarding the best practices for adopting FAIR principles, the idea is to accelerate knowledge discovery while reducing biases and improving the robustness and quality of scientific evidence presented by members of the FAIR community. The set of recommendations that have been generated in this study will allow health data hubs to integrate into a data sharing ecosystem such as the future European Health Research and Innovation Cloud, contributing to improving the quality of translational, sustainable, and trustworthy research based on data sharing for its reuse safely and with the appropriate guidelines to join efforts in search of progress in health research. The governance models and good practices for applying the FAIR principles in health data hubs analysed in this study were validated with the interviewees from the health data hubs, involving them in the review phase of the recommendations generated.

References

[1] HealthyCloud website: https://healthycloud.eu/

[2] Aarestrup, F. M., Albeyatti, A., Armitage, W. J., Auffray, C., Augello, L., Balling, R., ... & Van Oyen, H. (2020). Towards a European health research and innovation cloud (HRIC). Genome medicine, 12(1), 1-14.

[3] European Health Data Space: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

[4] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

[5] Eva, G., Liese, G., Stephanie, B., Petr, H., Leslie, M., Roel, V., ... & Greet, S. (2022). Position paper on management of personal data in environment and health research in Europe. Environment international, 107334.

[6] Dinov, Ivo D. "Volume and Value of Big Healthcare Data." Journal of medical statistics and informatics vol. 4 (2016): 3.

[7] Feinleib D. Big Data Bootcamp. Springer; 2014. The Big Data Landscape; pp. 15–34.

[8] Data Management Task Force, e-Infrastructure Reflection Group, "e-IRG Report on Data Management" http://www.eirg.eu/images/stories/e-irg_dmtf_report_final.pdf

[9] Regidor, Enrique. "The use of personal data from medical records and biological materials: ethical perspectives and the basis for legal restrictions in health research." Social science & medicine 59.9 (2004): 1975-1984.

[10] Vlahou, A., Hallinan, D., Apweiler, R., Argiles, A., Beige, J., Benigni, A., ... & Vanholder, R. (2021). Data sharing under the general data protection regulation: time to harmonize law and research ethics?. Hypertension, 77(4), 1029-1035.

[11] 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