Heliyon 9 (2023) e15733

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

FAIR principles to improve the impact on health research management outcomes

Alicia Martínez-García [a], Celia Alvarez-Romero [a,*], Esther Román-Villarán [a], Máximo Bernabeu-Wittel [b], 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*
[b] *Internal Medicine Department, Virgen del Rocío University Hospital, Seville, Spain*

ARTICLE INFO

Keywords:
FAIR principles
Data management
Health research management
Time investment
Economic savings
Survey

ABSTRACT

Background: The FAIR principles, under the open science paradigm, aim to improve the Findability, Accessibility, Interoperability and Reusability of digital data. In this sense, the FAIR4-Health project aimed to apply the FAIR principles in the health research field. For this purpose, a workflow and a set of tools were developed to apply FAIR principles in health research datasets, and validated through the demonstration of the potential impact that this strategy has on health research management outcomes.

Objective: This paper aims to describe the analysis of the impact on health research management outcomes of the FAIR4Health solution.

Methods: To analyse the impact on health research management outcomes in terms of time and economic savings, a survey was designed and sent to experts on data management with expertise in the use of the FAIR4Health solution. Then, differences between the time and costs needed to perform the techniques with (i) standalone research, and (ii) using the proposed solution, were analyzed.

Results: In the context of the health research management outcomes, the survey analysis concluded that 56.57% of the time and 16800 EUR per month could be saved if the FAIR4Health solution is used.

Conclusions: Adopting principles in health research through the FAIR4Health solution saves time and, consequently, costs in the execution of research involving data management techniques.

1. Introduction

Open science is a movement that drives scientific research to be accessible at the level of researchers and professionals, including sharing data, software, publications, etc, and the corresponding diffusion, which is a real need in the field of health research [1]. The FAIR principles, formally published in 2016 by the Force11 community, are a set of guidelines with the aim of improving the Findability, Accessibility, Interoperability, and Reusability of digital data [2]. Although the FAIR principles are applicable to any scenario and kind of data, from 2016 the use of the FAIR principles in the health field is in an exponential growth. With the high-level objective to apply the FAIR principles in the health field, the FAIR4Health project [3] which started in December 2018 and ended in November

[*] Corresponding author.
E-mail address: celia.alvarez@juntadeandalucia.es (C. Alvarez-Romero).

https://doi.org/10.1016/j.heliyon.2023.e15733
Received 20 March 2023; Received in revised form 19 April 2023; Accepted 20 April 2023
Available online 3 May 2023

2021 aimed to encourage the European Union Health Research community to apply the FAIR principles, aiming to facilitate the health data sharing and reuse. This European project was founded by the European Commission Call 'SwafS-04-2018', in the topic 'Encouraging the re-use of research data generated by publicly funded research projects'. FAIR4Health project was coordinated by the Technological Innovation Group at Virgen del Rocio University Hospital as part of the Andalusian Health Service in Spain, and accounted for seventeen partners from eleven European and non-European countries, bringing together expertise from different domains like health research, data managers, medical informatics, software developers, standards and lawyers.

Sharing health research data involves many challenges in different areas such as technical, conceptual and legal. The FAIR principles aim to ensure that data is shared in a way that enables and enhances the reuse of information by humans and machines. Numerous researchers have analyzed the advantages of applying FAIR principles in the field of health [4] and especially in the research of patients with chronic diseases [5,6]. In fact, it is essential to refer to a 2018 European Commission report [7] on the costs of NOT having FAIR data. That report estimates that: i) the cost of NOT having FAIR data is approximately €10.2bn per year for the EU; ii) besides, the open data economy suggests that the impact on innovation of FAIR could add another €16bn to the minimum cost estimated; and iii) that would make a total of at least €26.2bn per year. That is not even counting related reproducibility problems. In addition, as Barend Mons states [8], on average, 5% of overall research costs should go towards data stewardship. With €300 billion (US$325 billion) of public money spent on research in the European Union, we should expect to spend €15 billion on data stewardship.

To make health data Findable, Accessible, Interoperable, Reusable, a FAIRificiation workflow was designed and developed based on the guidance proposed by the GO FAIR initiative [9], applying restrictions on existing steps and including new steps for specific requirements and needs of health data. Taking into account the specific aspects added by the data from Electronics Health Records (EHR) and other health data sources, as well as the sensitive nature of this kind of data, this new workflow was designed and published [10] to address the translation of raw data/metadata into FAIR data/metadata in the health research field. Fig. 1 shows the designed FAIRification workflow that was applied in the FAIR4Health project.

Once the FAIRification workflow was designed, two applications were developed to cover the steps of this workflow. On the one hand, the Data Curation Tool (DCT) [11] aims to extract, transform and load existing healthcare and health research data into HL7 FHIR repositories. On the other hand, the Data Privacy Tool (DPT) [12] was developed to handle the privacy challenges exposed by the sensitive health data through anonymisation and de-identification techniques.

The FAIR4Health project has been considered the first proposal to translate the FAIR principles to health research data in Europe [13]. Following the design of the new FAIRification workflow to convert health data into FAIR data, the FAIR4Health platform [14] was developed with the aim of applying Artificial Intelligence (AI) algorithms on the FAIR health research datasets. In this way, to validate the technical solution provided by FAIR4Health, two pathfinder case studies were carried out. Firstly, identification of multimorbidity patterns and polypharmacy correlation on the risk of mortality in elderly [5]. And second one, an early prediction service for 30-days readmission risk in patients with Chronic Obstructive Pulmonary Disease [6]. Concretely, the FAIR4Health platform included Privacy-Preserving Distributed Data Mining (PPDDM) methods to carry out a federated use of different AI algorithms, to identify association between de data (like the FP-Growth algorithm), and to perform predictions (like the Support Vector Machine, Logistic Regression, Decision Trees, Random Forest, Gradient Boosted Trees). In both cases, taking into account the federated nature of the FAIR4Health platform, no data was shared between the clinical sites connected to the platform, AI partial models were generated in the facilities of each health data owner, and the platform generated merged models using these partial models.

To validate and assess the FAIR4Health solution, after obtaining approvals from the local ethics committees of all clinical centres involved, a prospective observational study [6] was carried out between April 2021 and September 2021 related to the FAIR4Health pathfinder case studies.

This paper aims to analyse the impact on health research management outcomes of the FAIR4Health solution by analysing the impact of the use of FAIR4Health tools on the time and costs of health research. For this purpose, a survey was designed and carried out to study whether the use of the tools and the new FAIRification workflow developed during the FAIR4Health project have an impact on health research if used in a real environment. So, researchers who participated in the study would have worked in both situations (with and without FAIR4Health tools) to be able to compare them.

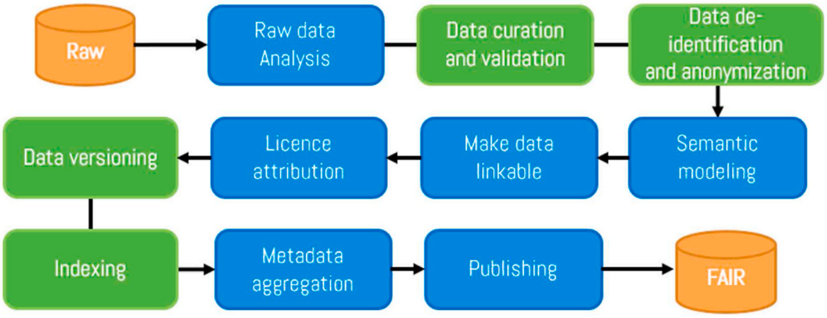

Fig. 1. FAIRification workflow implemented in FAIR4Health project.

2. Methods

2.1. Data collection

To understand how researchers manage their time during the research data management processes, a specific survey (Appendix 1) was designed to identify possible savings in time spent on scientists' data practices and efforts on research, thus bringing economic savings for the whole research community, and also how FAIR principles implementation could be enhanced. Privacy, consent and anonymisation issues related to the survey responses were indicated to each participant on the first page of the survey (and are shown in Appendix 1 of the manuscript, page 20).

The participants in the survey could identify what scientists are doing now in terms of data collection, use, storage, reuse, etc, aiming to achieve improvements in tools and processes for the application of FAIR principles into health research data. The interviewees answered the survey through a general email sent to tentative participants, representing academic and clinical institutions, among others.

The survey collected data on the amount of time researchers spend on specific research activities, such as administrative work, care practice or research. Within the time spent exclusively on research, questions related to the different steps of data management to make data FAIR were also included. In addition, questions related to the perception of different obstacles that may hinder the implementation of FAIR principles were included.

The survey contained four different sections. The first one (General Information) was aimed to get to know the interviewee profile with details such as age, sex, organisation they work with, percentage of time dedicated to research activities, number of projects they are usually working, typical dataset size they work with, etc. The second one (General questions regarding Data Science) was aimed to understand how researchers perceive existing difficulties to find and access appropriate data, and deal with it to make it suitable for their research, and the endorsement towards FAIR principles for the organisations they work in. Also, this section contained technical questions, such as AI techniques, standards to represent data or repositories to deposit metadata that are being used. The third section (Standalone Research without FAIR4Health) had the objective to inform about the actual time investment that the interviewees are currently making for a research project they are involved with, including questions per each task in a typical data management process needed to run a health research project. Finally, the fourth section (Research with FAIR4Health) informed about the same time investment needed to manage data when running health research projects, this time using FAIR4Health tools (DCT, DPT, and FAIR4-Health platform). That is, using a before and after comparison when the FAIR4Health solution becomes available.

Before disseminating the survey to potential respondents, the questionnaire was validated by established researchers currently involved in the consortium who are experts in data science.

In order to quantify research impact, and taking into account all data management processes, researchers were asked to think of a research project they had recently worked on. Questions were then asked about the time spent on different data management tasks in their projects, i.e. without using the FAIR4Health FAIRification workflow tools (stand-alone research) and using the FAIR4Health FAIRification workflow tools.

These tasks were as follows: (i) The 'data cleaning (pre-processing, curation and validation)' phase covered the processing and checks performed to ensure data quality in a specific investigation, taking into account the time needed to improve data and/or metadata quality or to transform data where necessary (e.g. identify and fix problems). (ii) In the "data normalisation, standardisation, semantic modelling, data integration and data interoperability" phase, data from different sources were consolidated into a homogeneous useable dataset, covering the transformation of two or more datasets into a machine-readable and harmonised format, considering the use of interoperability standards and/or semantic models. (iii) With the metadata record (link, licence, version, index and others), the data were adequately described accurately using all relevant data and/or metadata. In this case, consideration was given to the timing of recording the data and/or metadata and keeping this specific database up to date with a certain level of suitability for future research purposes. (iv) Publication of data and/or metadata in open public repositories covered the publication of the data and/or metadata so that it could be found. Consideration was given here to all the time in the publication processes that might vary between repositories, thus increasing the time. (v) Considering the sensitive nature of health data, one of the challenges for implementing FAIR principles in the health context is the legal one, related to compliance with the General Data Protection Regulation (GDPR). The FAIR4Health platform managed this challenge by using the PPDDM techniques explained in the introduction, and by offering the possibility to perform data anonymisation and/or de-identification. (vi) The activity "data processing and/or AI techniques" in the research cycle was ultimately aimed at extracting useful information to find insights and formulate conclusions and observations.

2.2. Statistical analysis

When doing the quantitative analysis, quantifying all costs of time spent by the research community when conducting standalone research (without using the FAIR4Health solution). Then, the same costs with using the FAIR4Health solution were aimed, thus obtaining an estimated measure of the potential savings. The revealed cost of Person/Month (PM) reported by each interviewee's institution was considered. Also, how much percentage of the time each interviewee dedicated fully to research efforts were considered.

Regarding the needed time to conduct a research without using the FAIR4Health tools, these were the times per average asked to the researchers, in order to cover all steps in the data management process: raw data extraction, data cleaning, make data interoperable, metadata registration, publication in open public repositories, make data anonymized and de-identified, and data processing.

On the other hand, those same questions with the scenario of using the FAIR4Health solution were asked to measure the percentage of time saved. Regarding using the FAIR4Health solution, all the steps concerning data cleaning, data normalisation, standardisation, semantic modelling, data integration and interoperability, metadata registration and data and/or metadata publication could be done using the DCT. By other hand, the DPT copes with the challenge of data anonymisation and/or de-identification. Finally, processing data and/or using AI techniques to find reproducible results is covered by the FAIR4Health platform, which is key to finding already made research efforts and leveraging them.

A data crossover plan was done, considering the previously mentioned tasks covered by each FAIR4Health tool in the whole data management process from a health research project. The time on average spent was asked in relation to the following.

According to that, answers for questions two (time spent on data cleaning), three (time spent on data normalisation, semantic modelling and data interoperability), four (time spent on metadata registration), and six (time spent on (meta)data publication in open/public repositories) in the section 'Standalone Research without FAIR4Health solution' were compared versus answers for the first question (time spent on those processes using the FAIR4Health DCT) in the section 'Research with FAIR4Health solution', to get to know time saved by applying the DCT.

Following the same logic, answers for questions five (time spent on data anonymisation and de-identification) in the section 'Standalone Research without FAIR4Health solution' and second one (time spent on anonymisation and de-identification applying the DPT) in the section 'Research with FAIR4Health solution' were compared against.

And the same occurred with questions related to the time on average spent on data processing and AI techniques. Answers from the questions seven in the section 'Standalone Research without FAIR4Health solution' and three in the section 'Research with FAIR4-Health solution' were also compared to measure times saved by the DPT and the FAIR4Health platform respectively.

Cost savings was calculated by multiplying the percentage of time saved by using the FAIR4Health solution with the revealed cost of Person/Month (PM) reported by each institution and the percentage of work time exclusively dedicated to research activities:

Cost savings = % Time Saved × Cost of PM × % Time aimed at research.

3. Results

The survey designed and fully included in Appendix 1, was sent to 56 researchers, and finally answered by 30 of them. The results of the statistical analysis defined is presented below, including the description of some interviewees' statistics, and the outcomes regarding saved time and costs.

The survey was answered by 30 researchers with knowledge in both data management techniques, and the use of the FAIR4Health workflow and tools. In fact, they participated in the design and functional requirements of the tools, as well as in the testing phase. Likewise, they hold meetings and training sessions with the FAIR4Health researchers who developed the tools and received user guides and training materials. These are the list of fourteen institutions for these 30 interviewees: Andalusian Health Service, Instituto Aragonés de Ciencias de la Salud, Universidad Carlos III de Madrid, Atos Spain SAE, HL7 Europe Foundation, Institut für Medizinische Informatik, Statistik und Epidemiologie - Universität Leipzig KöR, Academic Medical Center - University of Amsterdam, Université de Genève, Peter L. Reichertz Institute for Medical Informatics, University of Braunschweig, Università Cattolica del Sacro Cuore, European Federation for Medical Informatics Association, Software Research and Development Consultancy A. S., University of Porto, and the Institute for Pulmonary Diseases of Vojvodina.

Below, some statistics about the survey 30 interviewees characteristics are presented in Table 1.

On average, around a third of the work time of the interviewees was aimed at research efforts (32,2%), followed by teaching tasks (14,2%), projects management (13,9%), and healthcare practice (12,5%). Other responses were related to administration (9,0%), software development (7,6%), scientific divulgation (5,7%), and policy support (2,2%).

The research types were evenly distributed between categories not disjoint (Health Data Science, Other Medical Informatics topics and Other topics), being Health Data Science the category when more than half of interviewees were involved with (56,7%

Table 1
Characteristics of the 30 interviewees.

Sex	Nearly two thirds of the interviewees were male (63,33%), while the remaining were female (36,67%).
Age	While the participant's age was well distributed among different ranges, more than 80% were below fifty years old. The interviewees' age range were: 18–30 (20,0%); 31–40 (33,3%); 41–50 (30,0%); 51–60 (13,3%); and 61–70 (3,3%).
Primary country of employment	Variety from countries showed the cross cultural from researchers who actually participated in the survey: Spain (27%), Germany (17%), Italy (10%), Portugal (10%), Serbia (10%), Switzerland (10%), Turkey (7%), Belgium (3%), Netherlands (3%), and Other (3%).
Employment position [a] [b]	70% of interviewees to the survey were researchers at any stage of their professional career: R1 (13%), R2 (30%), R3 (17%), R4 (10%), Research Admin (10%), Others (17%), and Not applicable (3%).
Dataset size usually used	More than half of participants usually coped with datasets conformed by more than one thousand records. The dataset size usually used was well distributed among different ranges: <100 records (10%); 100–1000 records (20%); >1000 records (60%); and Not applicable (10%).

[a] R1 = First stage researcher (up to the point of PhD); R2 = Recognised researcher (PhD holders or equivalents who are not yet fully independent); R3 = Established researcher (researchers who have developed a level of independence); R4 = Leading researcher (researchers leading their research area or field); Others = Professions such as student, professor, project director, software engineer or senior consultant; NA = not answered.

[b] Research profiles descriptors extracted from: https://euraxess.ec.europa.eu/europe/career-development/training-researchers/research-profiles-descriptors.

respectively). 46,7% and 33,3% of interviewees were involved with "Other Medical Informatics topics" and "Other topics" categories, respectively. Those categories were not disjoint, so researchers were able to declare to be involved in projects in different categories.

Furthermore, half of our interviewees had less than or equal to three projects they were involved with. The number of health research projects involved per researcher was well distributed among different ranges: less than or equal to three projects (50,0%); between four and six projects (26,7%); and more than six projects (13,3%). This question was not applicable for 10,0% of interviewees.

3.1. Evaluation outcomes

The first step in the saved time analysis was to perform a quality control process to select compatible answers to consider for the quantitative results in terms of time invested. In this sense, 18 responses were considered valid from 30 final interviews to calculate time investment savings.

Analysing individual answers interviewees gave, it is estimated that 57,88% and 55,51% of the research time could be saved using DCT and DPT, respectively. These statistics were based on aggregated statistics from all valid answers considered. Also, interviewees stated that 57,30% of the time could be saved by using the FAIR4Health platform.

Putting all in perspective to cover all steps in a data management plan when conducting health research efforts, 56,57% of the time could be saved when using the whole FAIR4Health solution. This translated into economic savings of 16.800 EUR/month, representing 17,11% of the PM of the institution of our interviewees (Table 2). Considering the person/month cost revealed by each institution in their involvement in the FAIR4Health project, this calculus was made.

On the other hand, participants were asked to answer questions related to the relevance of the FAIR principles, using a scale from 1 to 5, when one is firmly not difficult and five is strongly difficult. On average, the participants answered with a 3,43, 3,36 and 3,33 to the questions regarding the difficulty of finding and accessing appropriate data, and making this data suitable, respectively. Comparing how answers to these questions differed according to a specific professional profile, established researchers (R3) are the group that is considered more difficult to find, access and make data suitable (Fig. 2).

When zooming on the techniques covered by the FAIR4Health solution, different professional profile groups were perceived at different times (Fig. 3). Established researchers (R3) were the group that used most of their time in using these techniques (37%).

When asked if the techniques they use allows them to reuse the data for future studies, 33,3% answered as "Yes, using the FAIR4Health solution", and 23,3% answered as "Yes, but not using the FAIR4Health solution". 6,7% answered "No" and 30% "I do not know". This question was not applicable for 6,7% of interviewees.

Nearly a half of our interviewees recognised that their organisations publish its data and/or metadata in public repositories or curate data and/or metadata, even though only a small percentage uses the FAIR4Health solution for that (Fig. 4).

4. Discussion

Many researchers and research organisations know the need and importance of data sharing and new practices to improve the knowledge discovery [15–17]. However, they are hesitant to share their datasets because of real or perceived costs, including time investment. After analysis of the survey results related to the impact of the FAIR4Health solution in terms of health research, the authors highlight the relevance and advantages of the practical use of the FAIR principles through the developed tools [18].

Several initiatives, organisations and European projects are working on this issue. On the one hand, the European Open Science

Table 2
Economic savings using the FAIR4Health solution.

#	% time aimed at research	% time saved using FAIR4Health tools	Total savings
1	40%	66,67%	1.188,53€
2	40%	23,53%	1.059,11€
3	40%	64,71%	1.187,22€
4	60%	82,35%	2.266,52€
5	10%	94,05%	120,95€
6	0%	73,19%	0,00€
7	20%	78,05%	868,53€
8	40%	66,67%	1.483,73€
9	55%	84,65%	2.095,12€
10	50%	80,00%	1.800,00€
11	25%	89,29%	1.375,00€
12	30%	62,50%	1.155,00€
13	20%	50,00%	1.125,30€
14	30%	86,67%	1.315,60€
15	35%	81,25%	1.438,94€
16	20%	0,00%	0,00€
17	25%	1,41%	14,44€
18	30%	−66,67%	−1.714,00€
Total Savings			16.799,98€
Total Cost of Payroll			98.047€
% Savings			17,11%

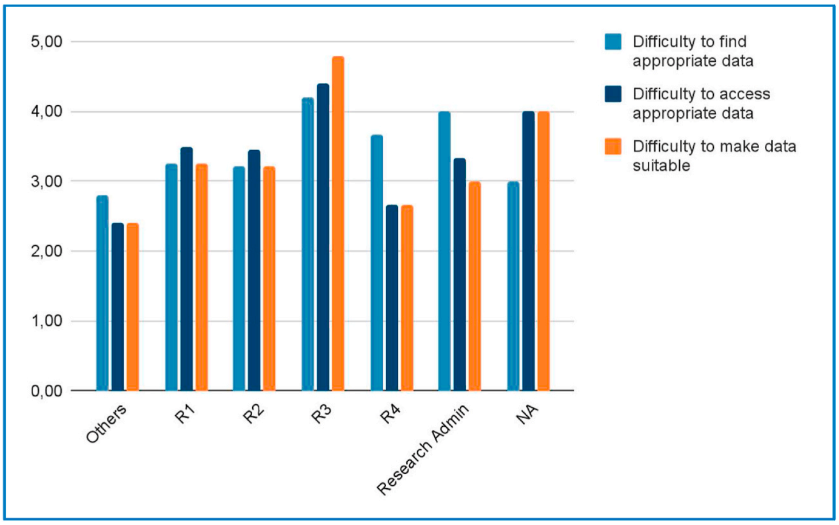

Fig. 2. Relevance of the FAIR principles, answers related to how difficult is finding, accessing, and dealing with the data for making them appropriate and suitable for your research, broken down by professional profiles.

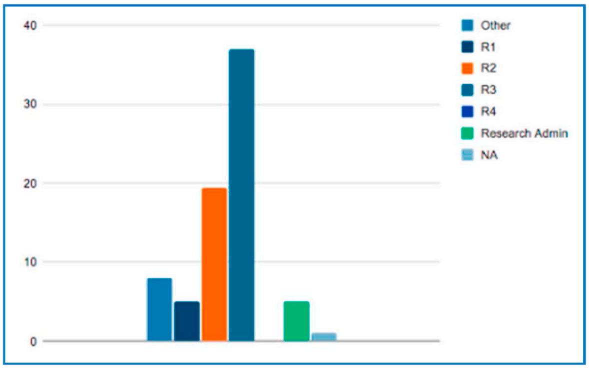

Fig. 3. Percentage of time used in data analysis/AI tasks is spent using these techniques: FP Growth, Support Vector Machine, Logistic Regression, Decision Trees, Random Forest, Gradient Boosted Trees.

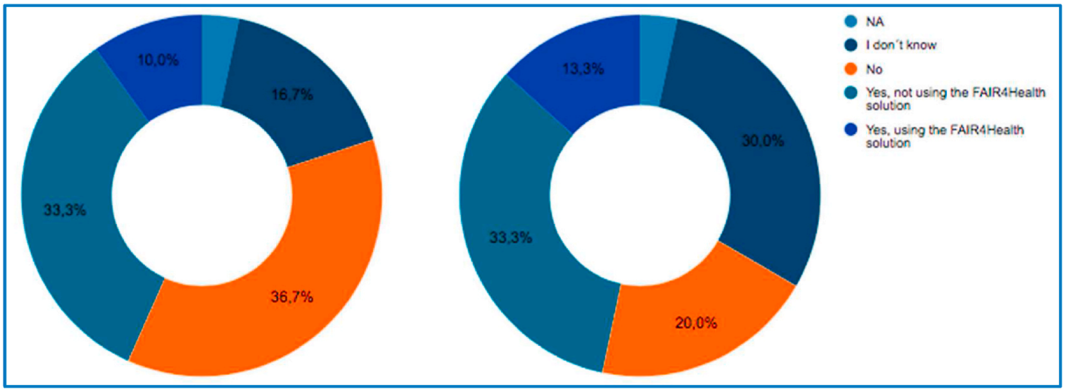

Fig. 4. Publishing data and/or metadata in public repositories (left)/Curation data and/or metadata (right).

Cloud [19,20], the GO FAIR initiative [21], and the EOSC-Life project [22] (and their working group called "Turning FAIR into reality"). Then, the European Health Data Space (EHDS) [23], one of the strategic priorities for the European Commission, the European Health Research and Innovation Cloud (one piece of the EHDS), and the HealthyCloud European project [24]. Besides, the Research

Data Alliance [25,26], in both some Groups involving health data (Health Data Interest Group, Reproducible Health Data Services Working Group, Covid-19 WG), and some Groups applying explicitly the FAIR principles (FAIR Data Maturity Model WG, FAIR for Research Software WG, Raising FAIRness in health data and Health Research Performing Organisations WG). Also, the HL7 international through the 'FAIRness for FHIR' project [27] as part of the Services Oriented Architecture Workgroup, and their first main tangible result: the HL7 FHIR implementation guide titled 'FHIR4FAIR' [28]. And finally, other European projects like FAIRsFAIR [29] and FAIRplus [30].

The implementation of the FAIR principles remains challenging due to some obstacles in the way, be it the lack of awareness in the research community on how to share data (which format, what information or metadata has to be provided, etc) [31], lack of preparation of a data management plan/missing metadata [32,33] or different standards for research data used [34–36].

The survey authors indicated that the survey should be answered only by researchers working in some data management technique (i.e. cleaning, normalisation, standardisation, semantic modelling, integration, interoperability, metadata registration, anonymisation, publication, processing with AI techniques or others) with knowledge about the use of at least one of the tools as part of the FAIR4Health solution (DCT, DPT and FAIR4Health platform). However, some interviewees answered without including relevant information in these essential questions for this study. Besides, although the questions related to the time on average used without and using the FAIR4Health solution were mandatory, some respondents added that they spent 0% of their time on research (on average). But to answer the survey, they had to know about the FAIR4Health solution, which is part of a research project.

Another issue that arose when facing the answers was the need for more capacity to compare the times of researchers using the FAIR4Health solution for specific research and replicate this same research without using the FAIR4Health tools. It is because, in some cases, the interviewee did not include allocation of time in the corresponding sub-categories, or the researcher established time (hours) about a working week while in the scenario of using the FAIR4Health solution they did not include this same reference. For these two reasons, only eighteen answers were used as valid ones, leaving the remaining twelve out of this part of the analysis.

Finally, another limitation identified in relation with the survey. Identifying researchers with experience in data management techniques and with knowledge about the use of the FAIR4Health solution was difficult, this is the reason the survey was sent only to 56 researchers, and all of them were from the FAIR4Health consortium.

Based on the survey to analyse the impact in terms of health research management outcomes, some questions were included in the design to collect information related to future improvements of the FAIR4Health solution, such as the coverage of others health informatics standards in the FAIRification tools, and the inclusion of new AI algorithms in the FAIR4Health platform.

The FAIR4Health solution covers the use of the following health informatics standards: HL7 FHIR, Snomed-CT, LOINC, ICD. These standards were used commonly with the 50,0%, 43,33%, 40,0% and 73,33% of interviewees, respectively (Fig. 5). As next steps the FAIR4Health solution could be extended including the following health informatics standards: ISO/CNE standards, openEHR, HL7 CDA, the International Classification of Primary Care, Dublin Core, CDISC family of standards, epidemiological standards, W3C standards.

The FAIR4Health platform includes six AI techniques. To the question of how much time when doing data analysis/AI tasks they spent using any of the above mentioned techniques, 36.7% answered as "Do not use" or "Not answered", and 50,0% as "Less or equal than 40% of time" (Fig. 6).

In a subsequent question, the survey questioned what other techniques (different from the above mentioned) are used for data analysis in your research. So as the following steps, the FAIR4Health platform could be extended including the following AI algorithms: factor analysis, other kinds of cluster analysis, neural/others network analysis, linear regression, convolutional and recurrent neural networks -i.e., Long short-term memory-, content analysis, social network analysis, ontological and semantic analysis, rule-based analysis, knowledge-based analysis, deep learning techniques, etc.

Another issue under-valuing the savings on direct costs in terms of time investment by researchers is that the data extraction/recollection still needs to be covered by the FAIR4Health tools. So, in a new version of the FAIR4Health solution, some improvements could be included by further reducing the time used to cover the data management techniques, thus bringing more beneficial results in terms of economic savings as well.

Therefore, the FAIR4Health solution provides tools to facilitate the application of FAIR principles to health research datasets after the design of a specific FAIRification workflow [10]. Once a survey was designed to analyse the impact on health research that the FAIR4Heatlh solution can have, it was disseminated to researchers who know data management techniques and the FAIR4Health workflow and tools. Finally, the main results were presented in this manuscript, concluding that adopting the FAIR principles in health research through the FAIR4Health solution saves time and, consequently, costs in the execution of research involving data management techniques.

Author contribution statement

Carlos Luis PARRA-CALDERÓN: conceived and designed the experiments; contributed reagents materials, analysis tools or data; wrote the paper.

Alicia MARTÍNEZ-GARCÍA: conceived and designed the experiments; performed the experiments; analyzed and interpreted the data; wrote the paper.

Celia ALVAREZ-ROMERO: conceived and designed the experiments; performed the experiments; analyzed and interpreted the data; wrote the paper.

Esther ROMÁN-VILLARÁN: performed the experiments; analyzed and interpreted the data; wrote the paper.

Máximo BERNABEU-WITTEL: analyzed and interpreted the data; contributed reagents, materials, analysis tools or data; wrote the

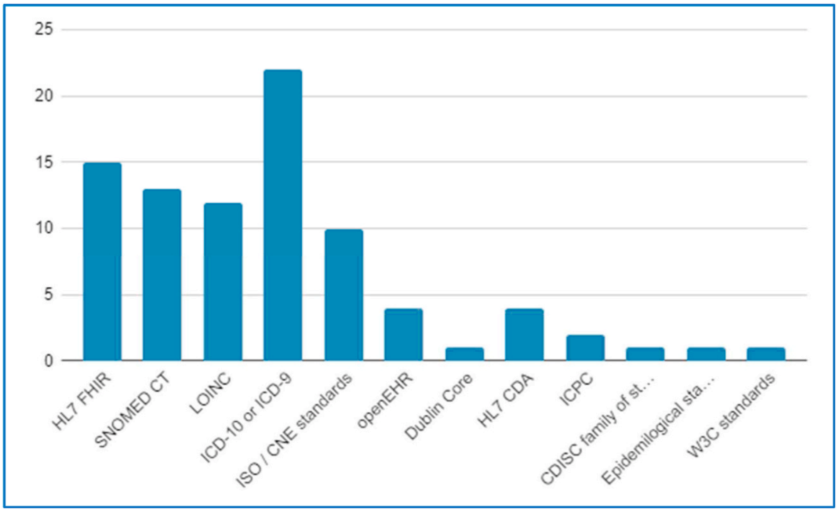

Fig. 5. Standards used to represent data or concepts.

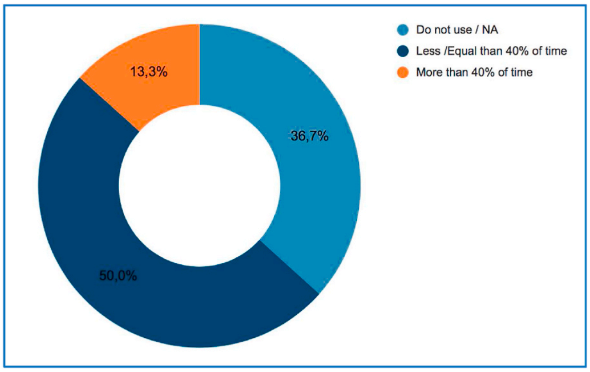

Fig. 6. Percentage of the time used in data analysis/AI tasks is spent using any of the following techniques: FP Growth, Support Vector Machine, Logistic Regression, Decision Trees, Random Forest, Gradient Boosted Trees.

paper.

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was co-supported by the FAIR4Health project [3], which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 824666. Also, this research has been co-supported by the Carlos III National Institute of Health, through the IMPaCT-Data program (code IMP/00019) 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'. Special acknowledgements to the FAIR4Health researchers who contributed to answering the survey to analyse the tentative impact on health research management outcomes.

Appendix B. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.heliyon.2023.e15733.

Abbreviations

AI	Artificial Intelligence
DCT	Data Curation Tool
DPT	Data Privacy Tool
EHDS	European Health Data Space
EHR	Electronic Health Records
FAIR	Findable, Accessible, Interoperable and Reusable
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
HL7	Health Level Seven
PM	Person/Month
PPDDM	Privacy-Preserving Distributed Data Mining
WG	Working Group.

References

[1] M. Woelfle, P. Olliaro, M.H. Todd, Open science is a research accelerator, Nat. Chem. 3 (10) (2011) 745–748.

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

[3] FAIR4Health project website. https://www.fair4health.eu/, 2016.

[4] M. Th, M. MartinGillian, S. Ugur, B. van OmmenGertJan, Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health, Biopreserv. Biobanking (2018).

[5] J. Carmona-Pírez, B. Poblador-Plou, A. Poncel-Falcó, J. Rochat, C. Alvarez-Romero, A. Martínez-García, A. Prados-Torres, Applying the FAIR4Health solution to identify multimorbidity patterns and their association with mortality through a frequent pattern Growth association algorithm, Int. J. Environ. Res. Publ. Health 19 (4) (2022) 2040.

[6] C. Alvarez-Romero, A. Martinez-Garcia, J.T. Vega, P. Díaz-Jimènez, C. Jimènez-Juan, M.D. Nieto-Martín, C.L.P. Calderón, Predicting 30-day readmission risk for patients with chronic obstructive pulmonary disease through a federated machine learning architecture on findable, accessible, interoperable, and reusable (FAIR) data: development and validation study, JMIR Medical Informatics 10 (6) (2022) e35307.

[7] European Commission, Directorate-General for Research and Innovation, Cost-benefit analysis for FAIR research data: cost of not having FAIR research data, Publications Office. https://data.europa.eu/doi/10.2777/02999, 2019.

[8] B. Mons, Invest 5% of research funds in ensuring data are reusable, Nature 578 (7796) (2020) 491–494.

[9] Go Fair, FAIRification process, in: https://www.go-fair.org/fair-principles/fairification-process/, 2016.

[10] A.A. Sinaci, F.J. Núñez-Benjumea, M. Gencturk, M.L. Jauer, T. Deserno, C. Chronaki, G. Erturkmen, B L. From raw data to fair data: the fairification workflow for health research, Methods Inf. Med. 59 (S 01) (2020) e21–e32.

[11] Open source code for data curation tool. https://github.com/fair4health/data-curation-tool, 2016.

[12] Open source code for Data Privacy Tool. https://github.com/fair4health/data-privacy-tool, 2016.

[13] C. Alvarez-Romero, A. Martínez-García, A.A. Sinaci, M. Gencturk, E. Méndez, T. Hernández-Pérez, C.L.P. Calderón, FAIR4Health: findable, accessible, interoperable and reusable data to foster health research, Open Research Europe 2 (34) (2022) 34.

[14] FAIR4Health platform. https://portal.fair4health.eu/, 2016.

[15] T.T. Makovski, S. Schmitz, M.P. Zeegers, S. Stranges, M. van den Akker, Multimorbidity and quality of life: systematic literature review and meta-analysis, Ageing Res. Rev. 53 (2019) 100903.

[16] K. Barnett, S.W. Mercer, M. Norbury, G. Watt, S. Wyke, B. Guthrie, Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study, Lancet 380 (9836) (2012) 37–43.

[17] F.H. Iglesias, C.A. Celada, C.B. Navarro, L.P. Morales, N.A. Visus, C.C. Valverde, J.M.B. Simó, Complex care needs in multiple chronic conditions: population prevalence and characterization in primary care. a study protocol, Int. J. Integrated Care 18 (2) (2018).

[18] M. Gencturk, A. Teoman, C. Alvarez-Romero, A. Martinez-Garcia, C.L. Parra-Calderon, B. Poblador-Plou, A.A. Sinaci, End User Evaluation of the FAIR4Health Data Curation Tool, in: Public Health and Informatics, IOS Press, 2021, pp. 8–12.

[19] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O.B. da Silva Santos, M.D. Wilkinson, Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science Cloud, Inf. Serv. Use 37 (1) (2017) 49–56.

[20] K. Koski, K. Hormia-Poutanen, M. Chatzopoulos, Y. Legré, B. Day, Position Paper: European Open Science Cloud for Research. UDAT, LIBER, OpenAIRE, EGI, GÉANT, Bari, 2015.

[21] P.C. Henning, C.J.S. Ribeiro, L.O.B.D.S. Santos, P.X.D. Santos, GO FAIR and FAIR Principles: what do they represent for the expansion of data in open Science? Em Questão 24 (2) (2019) 389–412.

[22] Eosc-Life project website. https://www.eosc-life.eu/, 2016.

[23] A.E. Jauregui, Overview of the European health data Space (EHDS): goals and current challenges, Eur. J. Publ. Health (2021).

[24] HealthyCloud project website. https://healthycloud.eu/, 2016.

[25] A. Treloar, The Research Data Alliance: globally co-ordinated action against barriers to data publishing and sharing, Learn. Publ. 27 (5) (2014) S9–S13.

[26] Research Data Alliance Working Groups. https://www.rd-alliance.org/groups/working-groups, 2016.

[27] Fairness for FHIR project website. https://confluence.hl7.org/pages/viewpage.action?pageId=91991234, 2016.

[28] FHIR for Fair Implementation Guide. http://hl7.org/fhir/uv/fhir-for-fair/2022Jan/, 2016.

[29] FairsFAIR project website. https://www.fairsfair.eu/, 2016.

[30] Fairplus project website. https://fairplus-project.eu/, 2016.

[31] P.C. Henning, L. Sales, An Interview with Barend Mons. Liinc Em Revista; V. 15, N. 2, in: Dados de Pesquisa| Research Data| Datos de Investigación, 24(2), 2019.

[32] Z. Zahedi, S. Haustein, T. Bowman, Exploring Data Quality and Retrieval Strategies for Mendeley Reader Counts, in: SIG/MET Workshop, ASIS&T, Seattle, 2014.

[33] T. Parsons, S. Grimshaw, L. Williamson, Research Data Management Survey: Report, 2013.

[34] R. Johnson, T. Parsons, A. Chiarelli, J. Kaye, Jisc Research Data Assessment Support–Findings of the 2016 Data Assessment Framework (DAF) Surveys, 2016.

[35] H. Stehouwer, P. Wittenburg, Second Year Report on RDA Europe Analysis Programme, 2014.

[36] C. Tenopir, S. Allard, K. Douglas, A.U. Aydinoglu, E. Read, M. Manoff, M. Frame, Data sharing by scientists: practices and perceptions,'', PLoS One 6 (2011) 1–21.