Check for updates

BRIEF REPORT

REVISED FAIR4Health: Findable, Accessible, Interoperable and Reusable data to foster Health Research

[version 2; peer review: 1 approved, 2 approved with reservations]

Celia Alvarez-Romero[1], Alicia Martínez-García[1], A. Anil Sinaci[2], Mert Gencturk[2], Eva Méndez[3], Tony Hernández-Pérez[3], Rosa Liperoti[4], Carmen Angioletti[4], Matthias Löbe[5], Nagarajan Ganapathy[6], Thomas M. Deserno[6], Marta Almada[7], Elisio Costa[7], Catherine Chronaki[8], Giorgio Cangioli[8], Ronald Cornet[9], Beatriz Poblador-Plou[10], Jonás Carmona-Pírez[10], Antonio Gimeno-Miguel[10], Antonio Poncel-Falcó[11], Alexandra Prados-Torres[10], Tomi Kovacevic[12,13], Bojan Zaric[12,13], Darijo Bokan[13], Sanja Hromis[12,13], Jelena Djekic Malbasa[12,13], Carlos Rapallo Fernández[14], Teresa Velázquez Fernández[14], Jessica Rochat[15], Christophe Gaudet-Blavignac[15], Christian Lovis[15], Patrick Weber[16], Miriam Quintero[17,18], Manuel M. Perez-Perez[17,18], Kevin Ashley[19], Laurence Horton[20], Carlos Luis Parra Calderón[1]

[1]Computational Health Informatics Group, Institute of Biomedicine of Seville, IBiS / Virgen del Rocío University Hospital / CSIC / University of Seville, Seville, 41013, Spain
[2]SRDC Software Research Development and Consultancy Corporation, Ankara, 06800, Turkey
[3]Dept. of Library & Inf Sci. Universidad Carlos III de Madrid, Getafe, 28903, Spain
[4]Department of Geriatric and Orthopedic Sciences, Catholic University of Sacred Heart, Roma, 00168, Italy
[5]Institute for Medical Informatics (IMISE), University of Leipzig, Leipzig, 04107, Germany
[6]PLRI Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, 38106, Germany
[7]Ucibio Requimte, Faculty of Pharmacy University of Porto. Porto4Ageing, Porto, 4050-313, Portugal
[8]HL7 Europe Foundation, Brussels, 1000, Belgium
[9]Amsterdam UMC, University of Amsterdam, Medical Informatics, Amsterdam Public Health, Amsterdam, 1105AZ, The Netherlands
[10]EpiChron Research Group, Aragon Health Sciences Institute (IACS), IIS Aragón, Miguel Servet University Hospital, Zaragoza, 50009, Spain
[11]EpiChron Research Group, Aragon Health Sciences Institute (IACS), IIS Aragón, Aragon Health Service, Zaragoza, 50009, Spain
[12]Medical Faculty University of Novi Sad, Novi Sad, 21000, Serbia
[13]Institute for Pulmonary Diseases of Vojvodina, Sremska Kamenica, 21204, Serbia
[14]J&A Garrigues, S.L.P., Seville, 41013, Spain
[15]University of Geneva and University hospitals of Geneva, Geneva, 1211, Switzerland
[16]Nice Computing SA Le Mont-sur-Lausanne, Le Mont-sur-Lausanne, 1052, Switzerland
[17]Atos Research and Innovation - ARI. Atos IT., Madrid, 28037, Spain
[18]Atos Research and Innovation - ARI. Atos Spain., Madrid, 28037, Spain
[19]Digital Curation Centre, University of Edinburgh, Argyle House, Edinburgh, EH3 9DR, UK
[20]Digital Curation Centre, University of Glasgow, Glasgow, G12 8QQ, UK

First published: 09 Mar 2022, **2**:34
https://doi.org/10.12688/openreseurope.14349.1

Open Peer Review

v2 **Latest published:** 31 May 2022, **2**:34
https://doi.org/10.12688/openreseurope.14349.2

Abstract

Due to the nature of health data, its sharing and reuse for research are limited by ethical, legal and technical barriers. The FAIR4Health project facilitated and promoted the application of FAIR principles in health research data, derived from the publicly funded health research initiatives to make them Findable, Accessible, Interoperable, and Reusable (FAIR). To confirm the feasibility of the FAIR4Health solution, we performed two pathfinder case studies to carry out federated machine learning algorithms on FAIRified datasets from five health research organizations. The case studies demonstrated the potential impact of the developed FAIR4Health solution on health outcomes and social care research. Finally, we promoted the FAIRified data to share and reuse in the European Union Health Research community, defining an effective EU-wide strategy for the use of FAIR principles in health research and preparing the ground for a roadmap for health research institutions.

This scientific report presents a general overview of the FAIR4Health solution: from the FAIRification workflow design to translate raw data/metadata to FAIR data/metadata in the health research domain to the FAIR4Health demonstrators' performance.

Keywords

FAIR principles, health research data management, HL7 FHIR, health data, data sharing, data reuse, health research, open science, privacy-preserving computing, machine learning.

This article is included in the Research on Research gateway.

This article is included in the Health Sciences gateway.

This article is included in the Horizon 2020 gateway.

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 31 May 2022	✓ view		
version 1 09 Mar 2022	? view	? view	? view

1. **Anupama Gururaj**, National Institutes of Health, Rockville, USA
2. **Gary Saunders**, EATRIS ERIC, Amsterdam, The Netherlands
3. **Salvador Capella-Gutierrez**, Barcelona Supercomputing Center (BSC), Barcelona, Spain

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Digital Health and Well-being collection.



This article is included in the Healthcare and Policy collection.

Corresponding author: Celia Alvarez-Romero (celia.alvarez@juntadeandalucia.es)

Author roles: Alvarez-Romero C: Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation; **Martínez-García A**: Investigation, Methodology, Validation, Visualization, Writing – Review & Editing; **Sinaci AA**: Investigation, Software, Writing – Review & Editing; **Gencturk M**: Investigation, Software, Writing – Review & Editing; **Méndez E**: Conceptualization, Investigation, Writing – Review & Editing; **Hernández-Pérez T**: Conceptualization, Investigation, Writing – Review & Editing; **Liperoti R**: Investigation, Validation, Writing – Review & Editing; **Angioletti C**: Investigation, Validation, Writing – Review & Editing; **Löbe M**: Investigation, Writing – Review & Editing; **Ganapathy N**: Investigation, Writing – Review & Editing; **Deserno TM**: Investigation, Writing – Review & Editing; **Almada M**: Investigation, Validation, Writing – Review & Editing; **Costa E**: Investigation, Validation, Writing – Review & Editing; **Chronaki C**: Investigation, Writing – Review & Editing; **Cangioli G**: Investigation, Writing – Review & Editing; **Cornet R**: Investigation, Writing – Review & Editing; **Poblador-Plou B**: Investigation, Validation, Writing – Review & Editing; **Carmona-Pírez J**: Investigation, Validation, Writing – Review & Editing; **Gimeno-Miguel A**: Investigation, Validation, Writing – Review & Editing; **Poncel-Falcó A**: Investigation, Validation, Writing – Review & Editing; **Prados-Torres A**: Investigation, Validation, Writing – Review & Editing; **Kovacevic T**: Investigation, Validation, Writing – Review & Editing; **Zaric B**: Investigation, Validation; **Bokan D**: Investigation, Validation, Writing – Review & Editing; **Hromis S**: Investigation, Validation, Writing – Review & Editing; **Djekic Malbasa J**: Investigation, Validation, Writing – Review & Editing; **Rapallo Fernández C**: Conceptualization, Investigation, Writing – Review & Editing; **Velázquez Fernández T**: Conceptualization, Investigation, Writing – Review & Editing; **Rochat J**: Investigation, Validation, Writing – Review & Editing; **Gaudet-Blavignac C**: Investigation, Validation, Writing – Review & Editing; **Lovis C**: Investigation, Validation, Writing – Review & Editing; **Weber P**: Investigation, Writing – Review & Editing; **Quintero M**: Investigation, Software, Writing – Review & Editing; **Perez-Perez MM**: Investigation, Writing – Review & Editing; **Ashley K**: Writing – Review & Editing; **Horton L**: Writing – Review & Editing; **Parra Calderón CL**: Conceptualization, Investigation, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 824666 (project FAIR4Health). Also, this research has been co-supported by the Carlos III National Institute of Health, through the IMPaCT Data project (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'.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Alvarez-Romero C *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Alvarez-Romero C, Martínez-García A, Sinaci AA *et al.* **FAIR4Health: Findable, Accessible, Interoperable and Reusable data to foster Health Research [version 2; peer review: 1 approved, 2 approved with reservations]** Open Research Europe 2022, **2**:34 https://doi.org/10.12688/openreseurope.14349.2

First published: 09 Mar 2022, **2**:34 https://doi.org/10.12688/openreseurope.14349.1

REVISED **Amendments from Version 1**

The second version has really few differences from the first version. Mainly, the comments proposed by the reviewers have been addressed and the improvements indicated have been included throughout the article. Likewise, the description of the Method and Results sections has been extended through more clarifications.

Besides, improvements to some statements in the text and the inclusion of an image that is better visualised have also been addressed.

In addition, new references of real relevance have been included in the Methods and Results sections, as well as more specifications of the two use cases.

Finally, a new reference has been added in the Discussion section to compare this solution with similar initiatives.

In conclusion, the added value in the second version of this article has been mainly the inclusion of references that are really relevant for the article, as well as improvements in the description of the text.

Any further responses from the reviewers can be found at the end of the article

Plain language summary

Health research organizations work more and more with health data. The reuse of health data has significant benefits for society, both financially and for our well-being. However, there are significant barriers to data sharing, which our project has taken steps to overcome. The FAIR principles of Findability, Accessibility, Interoperability and Reusability are intended to influence such institutions to support collaborative use of data. FAIR4Health promotes the application of FAIR principles in health research data derived from public projects. FAIR-4Health developed a workflow and tools to support the FAIR principles, and applied these to two case studies, extending across multiple health care sites, which confirmed feasibility.

Introduction

One of the more significant challenges of data-intensive science is to facilitate the breakthrough of knowledge by assisting humans and machines in the discovery, access, integration, and analysis of task-appropriate scientific data and their associated algorithms and workflows, facilitating reproducibility of the research.

The FAIR guiding principles describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse. Likewise, FAIR principles describe a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable[1]. Furthermore, the FAIR principles ensure that data are shared to enable and enhance reuse by humans and machines. Although FAIR emerged from a workshop for the life science community, the principles are intended to be applied to data and metadata from all disciplines.

Since their formal release via the FORCE11 community, FAIR principles have been adopted by several funders and governments worldwide. The European Commission data management guidelines were updated in 2017 to introduce the notion of FAIR. The European Open Science Cloud (EOSC) Declaration and recent EOSC Strategic Research and Innovation Agenda (EOSC SRIA) both emphasise the central role of FAIR data.

In addition, it is essential to refer to the report issued by the European Union about the costs of NOT having FAIR data[2]. The main conclusions of that report are that: i) the cost of NOT having FAIR data is approximately €10.2bn per year for the EU; ii) in addition, 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.

A diverse range of research disciplines are adopting FAIR principles. Several groups have been assessing FAIR uptake to date and the challenges being encountered. In the same way, the FAIR4Health project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824666, promotes the application of FAIR principles to health research data.

Methods and implementation of tools are presented in this manuscript, as well as the results obtained in two use cases.

Methods

First of all, we performed a comprehensive analysis of current barriers, facilitators and potential overcoming mechanisms in the EU to implement a FAIR data policy in health research institutions. Information from different perspectives (technical, ethical, security, legal, cultural, behavioural and economic) was gathered to generate guidelines providing an optimal strategy for implementing this policy in EU health research institutions. Concretely, a FAIR4Health public deliverable[3] provided an analytical overview of the main considerations addressed to identify, report and overcome the key barriers that could prevent Health Research Performing Organizations (HRPOs) from opening, sharing and FAIRifying their research data.

Then, FAIR4Health designed a workflow[3] to apply the FAIR principles to health research data, as well as to Electronic Health Record data, based on the FAIRification process of GO FAIR[4], but addressing the ethical, legal and technical aspects that health data include due to their sensitive nature by adding new steps in the workflow.

As shown in Figure 1 and Figure 2, new steps were included (in green) in the FAIR4Health FAIRification workflow to address these additional aspects through curation, validation and anonymization of sensitive health data. Adapted GO FAIR steps (in blue) define general actions for raw data analysis, license attribution, linking, semantic modeling, metadata management, and publishing to achieve FAIRness of existing (meta)data.

The requirements of health data were analysed in-depth, and FAIRification tools, based on the use of the HL7 FHIR standard, were developed to obtain FAIR data from raw data resulting

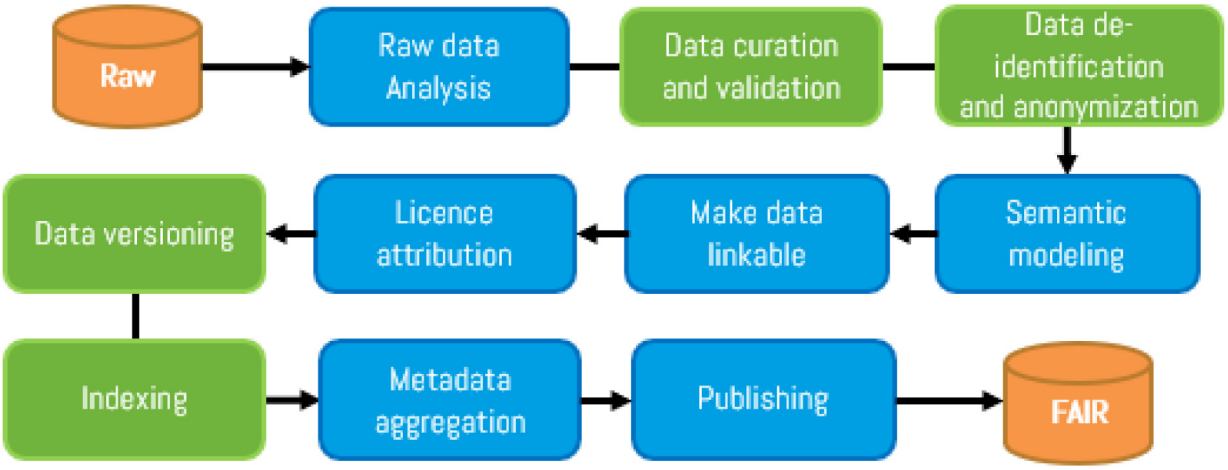

Figure 1. FAIR4Health workflow to apply FAIR principles in health research data (I).

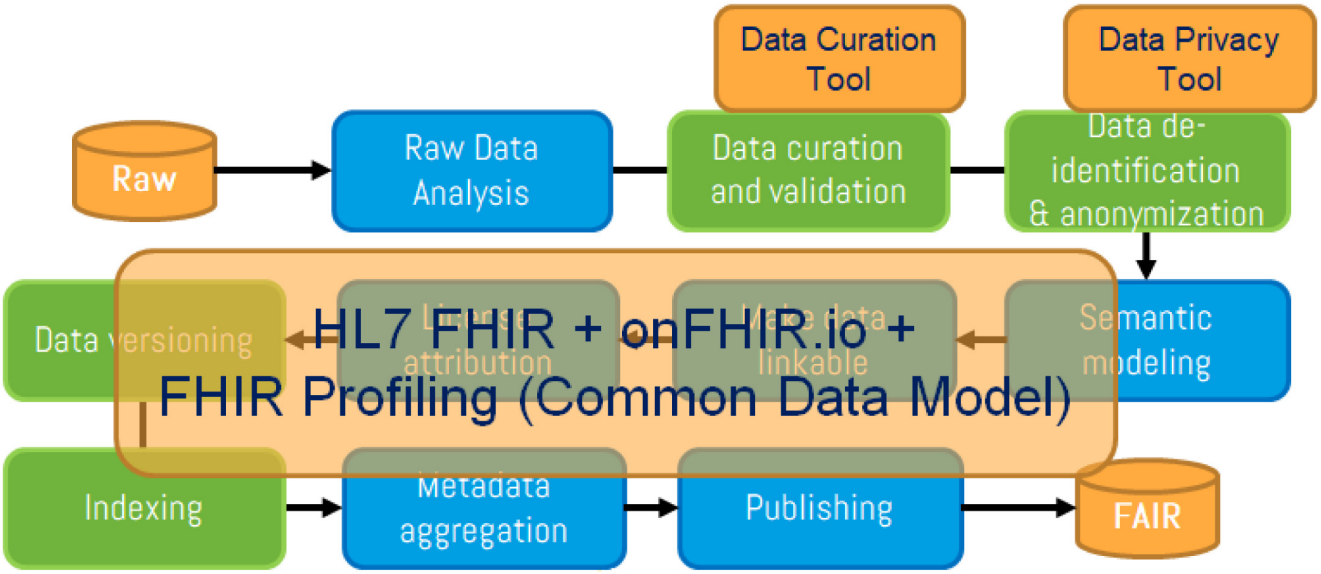

Figure 2. FAIR4Health workflow to apply FAIR principles in health research data (II). The FAIR4Health FAIRification workflow, based on the GO FAIR process (steps in blue), includes new steps (in green) to address the additional considerations for health data through curation, validation and anonymization of sensitive health data. Then, FAIRification tools, based on the use of the HL7 FHIR standard, were developed to obtain FAIR data from raw data.

from biomedical research. In the FAIR4Health project, the use of standards to facilitate the application of FAIR principles was studied, and the conclusion was that HL7 FHIR standard can support the FAIRification process and facilitate the representation of the FAIR data object conceptual components.

FAIRification tools are standalone, desktop applications developed by the FAIR4Health project to perform "Data curation and validation" and "Data de-identification and anonymization" steps of the FAIRification Workflow in an easier way:

- Data Curation Tool[5] is a highly specialized Extract-Transform-Load tool that can extract data from relational databases and spreadsheets, apply user-defined transformations, and load the transformed resources into an HL7 FHIR repository.

- Data Privacy Tool[6] is responsible for handling the privacy challenges on sensitive health data by applying several data de-identification and anonymization techniques. After the curation process, the Data Manager uses the Data Privacy Tool to de-identify data before making it available to other systems/components as FAIR data. This tool reads and writes de-identified resources back to the HL7 FHIR repository.

Figure 3 shows the architecture implementing the FAIR-4Health FAIRification Workflow for health data. In the core of architecture, an HL7 FHIR Repository acts as the health data repository. That way, the FAIR4Health core architecture, including an FHIR Repository and based on a Common Data Model[7], is an enabling factor for implementing the steps of the FAIRification workflow in all aspects of FAIR principles. In FAIR4Health, onFHIR.io was utilized as the HL7 FHIR Repository deployed within the agents.

On top of these, the FAIR4Health Platform was developed to apply a Privacy-Preserving Distributed Data Mining (PPDDM) framework enabling health research organizations to perform joint data mining operations without exposing any sensitive patient information to the outside world. To address the privacy-preserving mechanisms, the data mining framework[8] of the FAIR4Health project was implemented. In addition, the PPDDM Agent, which is responsible for running the data mining algorithms on top of the FAIRified data for the use cases defined by the user through the FAIR4Health Platform, was developed for training, validation and testing of models for the use cases defined. To achieve its objectives, the PPDDM Agent communicates with the onFHIR.io FHIR Repository within the data source boundaries, and the FAIR4Health Platform to exchange the results and predictive model information in a distributed manner.

The overall architecture of the FAIR4Health solution is shown in Figure 4.

Results

The main objective of FAIR4Health was to facilitate and encourage the European Union Health Research community to FAIRify, share and reuse their datasets derived from publicly funded research initiatives through the demonstration of the potential impact that such a strategy has on health outcomes and health and social care research.

The FAIR4Health solution was validated with the two pathfinder case studies based on FAIRified data through the PPDDM framework.

> Use case 1. Identification of multimorbidity patterns and polypharmacy correlation on the risk of mortality in elderly.
>
> Use case 2. Early prediction service for 30-days readmission risk in patients with Chronic Obstructive Pulmonary Disease (COPD).

The goal of these case studies was to test the developed tools in the project. The prototypes were developed making use of federated machine learning methodologies and algorithms implemented upon the FAIR4Health Platform. First, each health research dataset was FAIRified using the FAIR4Health FAIRification tools. Then, the federated machine learning algorithms were trained and validated with retrospective datasets in both case studies. Finally, a prospective study was performed in the second use case to validate the developed model for prediction.

Concretely, the main goal of the pathfinder case study #1 was to analyze the impact of multimorbidity patterns and polypharmacy on the six-month mortality rate and cognitive impairment among elderly individuals in different health care settings. As a result, a multicentric retrospective observational study was designed in which data were collected from

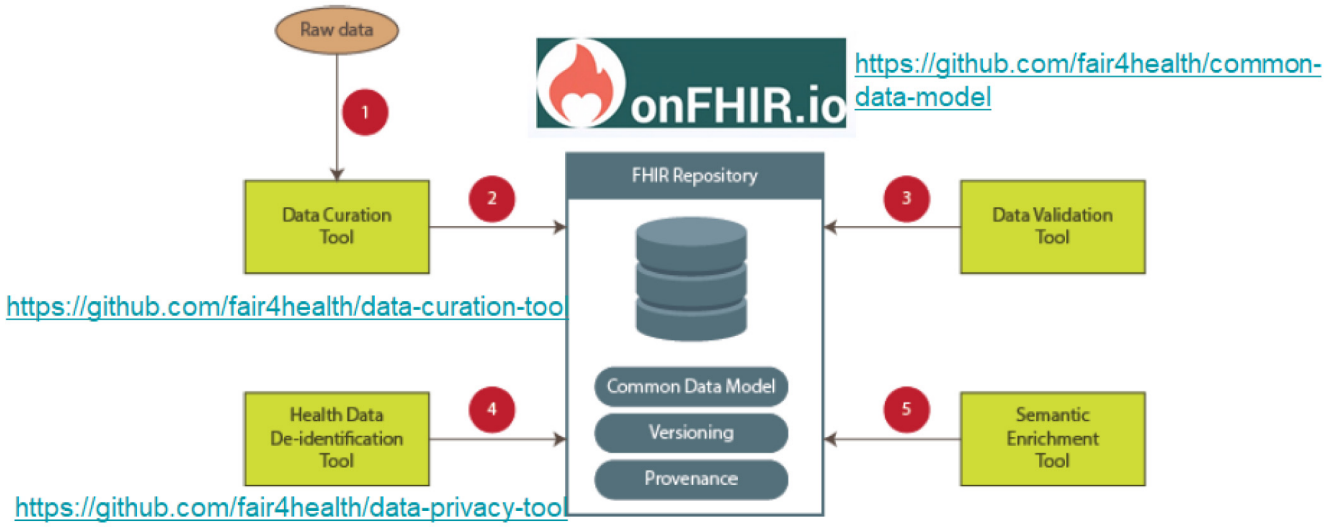

Figure 3. FAIR4Health architecture implementing the FAIRification Workflow for health data. At the core of architecture, an HL7 FHIR Repository acts as the health data repository. The FAIR4Health core architecture, which includes an FHIR Repository and is based on a Common Data Model, is an enabling factor for implementing the steps of the FAIR4Health FAIRification workflow in all aspects of FAIR principles.

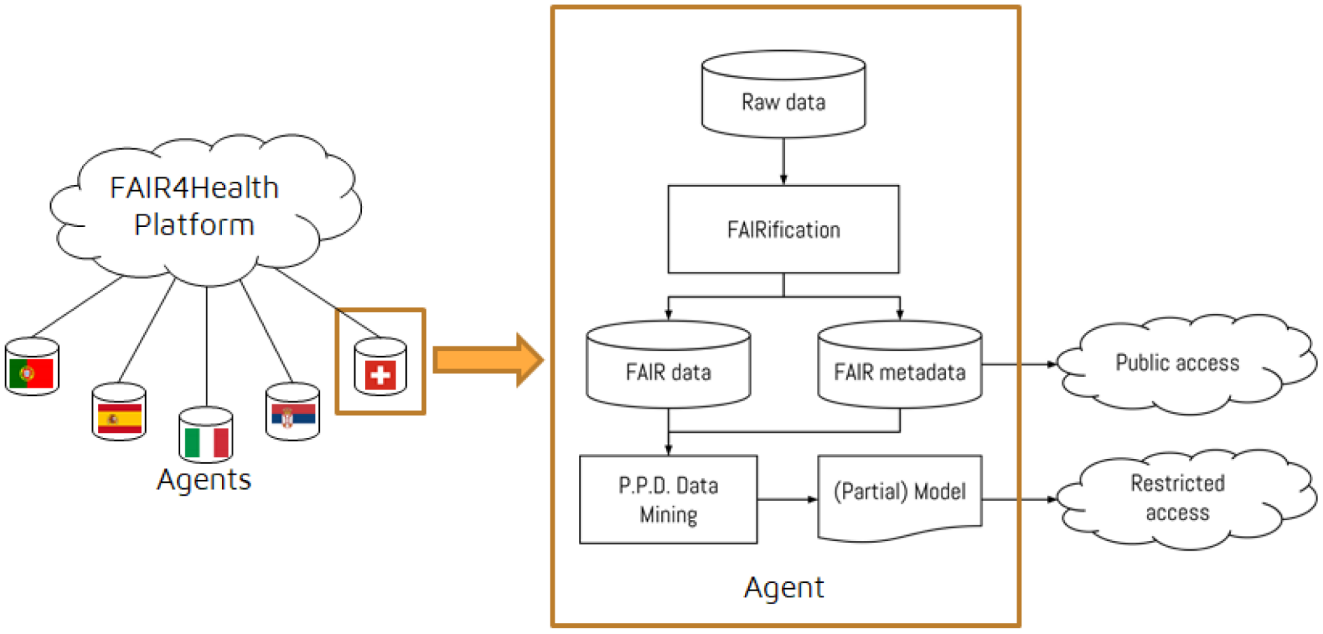

Figure 4. The overall architecture of the FAIR4Health solution. FAIR4Health Platform was developed to apply Privacy-Preserving Distributed Data Mining (PPDDM) models enabling health research organizations to perform joint data mining operations without exposing any sensitive patient information to the outside world. PPDDM Agents, which are responsible for running the data mining algorithms on top of the FAIRified data for the use cases defined by the user through the FAIR4Health Platform, were developed for training, validation and testing of models for the use cases defined. To achieve its objectives, the PPDDM Agents communicate with the onFHIR.io FHIR Repository within the data source boundaries, and the FAIR4Health Platform to exchange the results and predictive model information in a distributed manner.

5 different European cohorts. In this case, the sample size was 11486 patients. The population studied consisted of individuals aged 65 years or older with at least two chronic diseases. We used a frequent pattern tree association algorithm[9] implemented in the FAIR4Health Platform to identify the most frequent patterns in five different scenarios. The multimorbidity patterns obtained were consistent with previous studies[10,11], which show the clinical potential of this method. We could also estimate a strong association between multimorbidity and polypharmacy and each of them with mortality. The results of the first use case were published as Open Access scientific publication[12].

COPD is one of the most prevalent chronic diseases. It has been associated with high morbidity and mortality and a high rate of readmission/rehospitalization and therefore associated with high healthcare costs. Thus, the main goal of the pathfinder case study #2 was to develop, validate and assess the accuracy of a clinical decision support tool for predicting 30-day readmission risk in patients suffering from COPD at discharge. In this line, the pathfinder case study #2 was composed of two phases to reach the main objective. The first one included a retrospective multicenter observational study, including the training and generation of prediction models in the FAIR4Health Platform. Concretely, the prediction model for the 30-days hospital readmission risk was trained using the retrospective data of 4944 COPD patients. In the second phase, a prospective observational study with a 30-day follow-up was performed, from April 2021 to September 2021, to evaluate the accuracy of this tool by collecting data from a selected sample of subjects. The study population consisted of individuals aged 18 and older with a diagnosis of COPD who were admitted to the hospital for this disease. Finally, to assess the prediction risk accuracy associated with the early prediction service for 30-days readmission risk in COPD patients, predictions generated by the FAIR4Health Platform were compared with real-world data. The clinical assessment concluded that from 100 recruited patients, the prediction was correct in 87% of cases (that is, in real-life, the patient was readmitted and the algorithm predicted that there was early 30-days hospital readmission risk; or the patient was not readmitted and the algorithm predicted that there was not early 30-days hospital readmission risk). The results and main findings of the second use case are been published (Open Access paper accepted and is currently in production)[13].

Further details of the FAIR4Health pathfinder case studies can be found in the public report on the demonstrators' performance[14].

Conclusions/Discussion

FAIR4Health partners achieved the project's objectives and the FAIR4Health use cases were successfully carried out through to the correct implementation of the technologies and performance of the complex FAIR4Health technical solution.

The main aim of the FAIR4Health project was to test the developed tools in the project: 1) application of FAIR principles in health research through the FAIR4Health FAIRification tools; 2) use of federated machine learning techniques; and 3) clinical, technical and functional validation of the FAIR4Health Platform and agents.

Therefore, FAIR4Health partners got positive conclusions from the FAIR4Health use cases. In both use cases, significant cross-cutting data-related issues and challenges were identified and addressed. The task to extract data from EHRs and other kinds of healthcare sources aligning this extraction with a FAIR4Health Common Data Model was not trivial and required a lot of conceptual and technical efforts, because: (i) complexity of the raw data (the source EHRs are commonly very complex including information in several tables in the source databases); (ii) free text used in some fields in the raw data sources; and (iii) differences between the type of the raw data sources. To address the complexity of the raw data, each health research organization from different countries that participated in data extraction involved colleagues who were experts in each source data model. To address the information in free text fields, Natural Language Processing (NLP) techniques were assessed, and finally, in some cases, manual NLP to extract structured information from unstructured information was performed to apply the FAIR4Health Common Data Model. Due to the differences in the raw data sources, each raw dataset had to be analyzed in depth in collaboration between the clinical partners and the technical partners. This involved determining the required configuration of the FAIR4Health solution to enable FAIRification of all raw data. Finally, coordinated federated machine learning models were created using all sources.

Other large-scale efforts such as the Observational Health Data Sciences and Informatics (OHDSI)[15] initiative is community-led and have leveraged distributed analytics for answering scientific questions. Concretely, in order to compare similar initiatives, the OHDSI suite[16] is an open-source, modular solution that enables organizations to explore 360° patient journeys and turn data into evidence. The ecosystem provides a broad range of tools that cover all aspects of real-world data and evidence – from data characterization to a standardized data model (OMOP CDM). This enables large scale cross-database analytics with OHDSI.

It is relevant to add other significant conclusions as lessons learnt here, related to the application of the FAIR principles in health research:

- Implementation of FAIR principles allowed us to use larger and more heterogeneous datasets in FAIR4Health, increasing the variability of the data, the size of the datasets, and finally, more comprehensive and reliable results/outputs, compared to specific research studies without applying FAIR.
- We could reuse FAIR datasets from other clinical organizations in a secure way, ensuring compliance with General Data Protection Regulation (GDPR), and we could use the clinical datasets in the federated machine learning models. In the FAIR4Health project, we could also consider demographic, environmental, clinical and social information. We achieved greater variability of datasets and inclusion of more variables, compared to research where FAIR datasets are not reused.
- We obtained an increase in the scope of the research and improvements in health research, facilitating the discovery of scientific knowledge through data sharing and data reuse. Likewise, FAIR data reuse provided savings in data collection where much effort is currently invested.
- The implementation of FAIR principles facilitated the reproducibility of the study and access to large volumes of data to make the research more robust. Therefore, this study can facilitate the increase in secondary use of datasets once FAIR policies were implemented, related to the publication and sharing of FAIR datasets.
- Finally, it is essential highlight that a lot of manual effort and coordination was a part of the FAIR4Health project, and this concludes that improving the scalability of the proposed solution is a future work that can be addressed with the implementation of further use cases.

Data and software availability

Underlying data

No data are associated with this article.

Extended data

Along with the FAIR4Health software, FAIR metadata related to the FAIRified datasets generated in the FAIRification process, is published in the FAIR4Health GitHub. This is available to the scientific community, and the FAIR4Health consortium continues assessing the possibilities to open publish these metadata in other public repositories. Further information: https://github.com/fair4health/.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Author contributions

Celia Alvarez-Romero: Investigation + Methodology + Validation + Visualization + Writing – Original Draft Preparation

Alicia Martínez-García: Investigation + Methodology + Validation + Visualization + Writing – Review & Editing

A. Anil Sinaci: Investigation + Software + Writing - Review & Editing

Mert Gencturk: Investigation + Software + Writing - Review & Editing

Eva Méndez: Conceptualization + Investigation + Writing - Review & Editing

Tony Hernández-Pérez: Conceptualization + Investigation + Writing - Review & Editing

Rosa Liperoti: Investigation + Validation + Writing - Review & Editing

Carmen Angioletti: Investigation + Validation + Writing - Review & Editing

Matthias Löbe: Investigation + Writing - Review & Editing

Nagarajan Ganapathy: Investigation + Writing - Review & Editing

Thomas M. Deserno: Investigation + Writing - Review & Editing

Marta Almada: Investigation + Validation + Writing - Review & Editing

Elisio Costa: Investigation + Validation + Writing - Review & Editing

Catherine Chronaki: Investigation + Writing - Review & Editing

Giorgio Cangioli: Investigation + Writing - Review & Editing

Ronald Cornet: Investigation + Writing - Review & Editing

Beatriz Poblador-Plou: Investigation + Validation + Writing - Review & Editing

Jonás Carmona-Pírez: Investigation + Validation + Writing - Review & Editing

Antonio Gimeno-Miguel: Investigation + Validation + Writing - Review & Editing

Antonio Poncel-Falcó: Investigation + Validation + Writing - Review & Editing

Alexandra Prados-Torres: Investigation + Validation + Writing - Review & Editing

Tomi Kovacevic: Investigation + Validation + Writing - Review & Editing

Bojan Zaric: Investigation + Validation + Writing - Review & Editing

Darijo Bokan: Investigation + Validation + Writing - Review & Editing

Sanja Hromis: Investigation + Validation + Writing - Review & Editing

Jelena Djekic Malbasa: Investigation + Validation + Writing - Review & Editing

Carlos Rapallo Fernández: Conceptualization + Investigation + Writing - Review & Editing

Teresa Velázquez Fernández: Conceptualization + Investigation + Writing - Review & Editing

Jessica Rochat: Investigation + Validation + Writing - Review & Editing

Christophe Gaudet-Blavignac: Investigation + Validation + Writing - Review & Editing

Christian Lovis: Investigation + Validation + Writing - Review & Editing

Patrick Weber: Investigation + Writing - Review & Editing

Miriam Quintero: Investigation + Software + Writing - Review & Editing

Manuel M. Perez-Perez: Investigation + Writing - Review & Editing

Kevin Ashley: Writing - Review & Editing

Laurence Horton: Writing - Review & Editing

Carlos Luis Parra Calderón: Conceptualization + Investigation + Project Administration + Supervision + Writing – Review & Editing

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
2. European Commission: **Cost of not having FAIR research data - Cost-Benefit analysis for FAIR research data.** 2018.
Reference Source
3. **FAIR4Health Guidelines for implementing FAIR open data policy in health research.**
Reference Source
4. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, *et al.*: **From Raw Data to FAIR Data: The FAIRification Workflow for Health Research.** *Methods Inf Med.* 2020; **59**(S 01): e21–e32.
PubMed Abstract | **Publisher Full Text**
5. **FAIR4Health Data Curation Tool.**
Reference Source
6. **FAIR4Health Data Privacy Tool.**
Reference Source
7. **FAIR4Health Common Data Model.**
Reference Source
8. **FAIR4Health Privacy-Preserving Distributed Data Mining (PPDDM) framework.**
Reference Source
9. Han J, Pei J, Yin Y: **Mining frequent patterns without candidate generation.** *ACM sigmod record.* 2000; **29**(2): 1–12.
Publisher Full Text
10. Poblador-Plou B, Calderón-Larrañaga A, Marta-Moreno J, *et al.*: **Comorbidity**

of dementia: a cross-sectional study of primary care older patients. *BMC Psychiatry.* 2014; **14**(1): 84.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, *et al.*: **Multimorbidity patterns: a systematic review.** *J Clin Epidemiol.* 2014; **67**(3): 254–266.
PubMed Abstract | Publisher Full Text

12. Carmona-Pírez J, Poblador-Plou B, Poncel-Falcó A, *et al.*: **Applying the FAIR4Health Solution to Identify Multimorbidity Patterns and Their Association With Mortality Through a Frequent Pattern Growth Association Algorithm.** *Int J Environ Res Public Health.* 2022; **19**(4): 2040.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Alvarez-Romero C, Martínez-García A, Ternero-Vega JE, *et al.*: **Predicting 30-days Readmission Risk for COPD Patients Care through a Federated Machine Learning Architecture on FAIR Data: Development and Validation Study.** *JMIR Medical Informatics.* 2022; **1**.
Publisher Full Text

14. **FAIR4Health Report on the demonstrators performance.**
Reference Source

15. **Observational Health Data Sciences and Informatics (OHDSI).**
Reference Source

16. **Observational Health Data Sciences and Informatics (OHDSI) suite.**
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 2

Reviewer Report 24 June 2022

https://doi.org/10.21956/openreseurope.16044.r29410

© **2022 Gururaj A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

✔ **Anupama Gururaj**
National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, USA

The authors have appropriately addressed the concerns raised previously.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomedical data sharing and management, clinical informatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 27 Jul 2022

Celia Alvarez

Thank you very much for reviewing our article. All the best.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 19 April 2022

https://doi.org/10.21956/openreseurope.15486.r28796

© 2022 Capella-Gutierrez S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Salvador Capella-Gutierrez
Department of Life Sciences, Barcelona Supercomputing Center (BSC), Barcelona, Spain

Alvarez-Romero and colleagues provided a fascinating article on how FAIR data in health research can benefit researchers and patients. This effort is part of the recently finished FAIR4Health project and illustrates its progress.

I have to congratulate the authors on the extended FAIRification process for health research data. I appreciate the data versioning as an essential step toward reproducibility of any result.

I have some comments aiming to foster the discussion and clarify some aspects of this manuscript:

1. Is there any specific reason to focus on HL7 FHIR as an interoperability standard across participating centres? Despite HL7 FHIR, I'd expect at least a mention of other standards to facilitate health-related research, e.g. OMOP. This aspect is especially relevant when working with observational studies, e.g. cohort-based research. Thus, the interesting point is to know how extensible this work is and the technical implications of making such an effort.

2. Can you provide additional details on the privacy-preserving mechanisms? If it has already been published somewhere else, a couple of sentences summarising it with the reference should be enough.

3. I wonder about the use of onFHIR platform. I assume it has been developed by one of the FAIR4Health partners. Then it makes sense to use it. I appreciate the fact of using open-source software. Can similar tools/platforms replace it? How interoperable are the outputs of this platform? I'm thinking of interested parties whiling to use/leverageFAIR4Health outcomes having their solutions.

4. I'd suggest delineating/introducing the use-cases earlier in the text.

5. I find figure #3 very informative. However, it led me to ask myself about potential mechanisms for accessing data produced in the consortium. If there is any formal mechanism to access or request access to those datasets, I think they can be included in this figure.

6. Looking at use-case #1, you mentioned that implemented algorithms are available as part of the platform in the text. Perhaps you can include a link to it, e.g. a specific repository in the FAIR4Health GitHub Organization.

7. Still looking at use-case #1, you mentioned: "... a strong association between multimorbidity and polypharmacy and each of them with mortality." Have you performed any statistical analysis here? Perhaps it is good to include them as part of the use-case #1

discussion.

8. I like the description of use-case #2 and appreciate that there is an extensive report (68 pages) describing it. However, readers would appreciate a short explanation of the main findings for this use-case. Otherwise, it seems a bit disconnected.

9. You have briefly discussed the possibilities of using FAIR data for distributed ML across different sites. Can you explain the minimal computational requirements for carrying on these analyses? I think it is essential for readers to realise that it is not enough to have FAIR data at their sites but also the computational capabilities to conduct such analyses.

10. I like the description of the efforts to FAIRify health research data. However, I'm missing the language axis when using NLP technologies. NLP models and resources are language-dependent, which means that the final results are partially affected by them. As you are working with data from 5 different European countries, can you share your experiences on these aspects?

Looking forward to your comments on those aspects.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: FAIR principles

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 May 2022

Celia Alvarez

Dear reviewer, All comments and recommendations have been addressed. Improvements and expanded descriptions of some parts of the article have been included in the new version, along with contributions suggested by the other reviewers.

Thank you very much for the opportunity to improve the manuscript and for all the comments, very timely to improve the scientific quality. Best regards.

Competing Interests: No competing interests were disclosed.

Reviewer Report 08 April 2022

https://doi.org/10.21956/openreseurope.15486.r28777

© 2022 Saunders G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Gary Saunders
EATRIS ERIC, Amsterdam, The Netherlands

This is a strong brief report of two use cases utilising the FAIR4Health project solution in addressing two different use cases of health data demonstrating the benefits in the application of the FAIR principles to such data that result in strong conclusions. Furthermore, the authors have published and made available all software that is referenced in the report so that it is available to the community.

I support the publication of this report, and have only two minor corrective suggestions:

1. Please can the authors include the total number of individuals that were analysed in use case 1? It would also be good to have this number broken down to the 5 distinct populations, however this may not be possible due to privacy concerns. As the use case discusses the application of federated ML it would be good to have some indication of sample size that was used in the analysis.

2. The NLP techniques that are discussed in the conclusion should be described earlier in the manuscript. There is some nice text in the discussion section not only of the NLP techniques but also of the complex application of the FAIR4Health Common Data Model to the use cases discussed in the report. It would be good to have this properly described in the Methods section of the manuscript as not only will this aid transparency in the processes, and therefore reproducibility of the results described, but also aid the community in the potential application of the CDM to future datasets.

In summary, this is a nice, neat and concise manuscript describing the work in an easy-to-read, follow, and digest manner.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: FAIR data, big data management, federated data access and analysis, health care data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 May 2022

Celia Alvarez

Dear reviewer, All comments and recommendations have been addressed. Improvements and expanded descriptions of some parts of the article have been included in the new version, along with contributions suggested by the other reviewers.

Thank you very much for the opportunity to improve the manuscript and for all the comments, very timely to improve the scientific quality. Best regards.

Competing Interests: No competing interests were disclosed.

Reviewer Report 04 April 2022

https://doi.org/10.21956/openreseurope.15486.r28776

© 2022 Gururaj A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

?

Anupama Gururaj
National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, USA

This paper describes a project focusing on the need to make healthcare data FAIR and reusable in the European Union context. The study addresses an emerging research need within the healthcare system and will be of interest to clinical research personnel in various healthcare settings. This team has performed two pilot studies that outline a general framework that can be adopted by other systems. The manuscript is concise and provides an overview of the project and project outcomes including tools developed and use cases addressed.

Major Criticisms:

1. In the abstract, the authors mentioned access to certified FAIR datasets. It is not clear in the rest of the manuscript how and who has provided such certification to the metadata of the two datasets that are shared. Please provide further details or remove mention of certification.

2. In parts, the paper overstates some aspects of the methods and conclusions. As an example, in the methods section, it is stated that the FAIR4Health provides an overview of all considerations to overcome all barriers to open research data sharing by HRPOs. Likewise, in the conclusions/discussions section, the authors state that they have seen an increase in secondary use of the datasets without providing any usage metrics. My recommendation is to tone down the absolute statements.

3. It would be useful to the reader to get a better understanding of the Common Data Model that is used to harmonize the data from the various sources. Please provide a description of the model in the paper.

4. Other large-scale efforts such as OHDSI are community-led and have leveraged distributed analytics for answering scientific questions. Please compare and contrast the FAIR4Health initiative to such efforts in the discussion.

5. It would be useful to the community to hear about lessons learnt and challenges that have not yet been solved through the life of the project. Also, please comment on the scalability of the approach since a lot of manual effort and coordination was a part of the project.

Minor comments:

1. Some of the sentences (such as the starting sentence of the second paragraph of the introduction (The FAIR guiding principles...) are very long and make the paper hard to read. It is recommended to simplify or split these sentences.

2. Figure 1 is hard to read and follow since there are overlapping boxes. Please fix the figure to increase clarity.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomedical data sharing and management, clinical informatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 May 2022

Celia Alvarez

Dear reviewer, All comments and recommendations have been addressed. Improvements and expanded descriptions of some parts of the article have been included in the new version, along with contributions suggested by the other reviewers.

Thank you very much for the opportunity to improve the manuscript and for all the comments, very timely to improve the scientific quality. Best regards.

Competing Interests: No competing interests were disclosed.