Achieving data FAIRification in a distributed analytics research platform for rare diseases

Anna Bernasconi[1,*], Cinzia Cappiello[1], Stefano Ceri[1] and Pietro Pinoli[1]

[1]*Department of Electronics, Information and Bioengineering – Politecnico di Milano, Milan, Italy*

Abstract

Data-driven medicine is fundamental to improving the accessibility and quality of the healthcare system. The availability of data is crucial for this purpose. In the context of a distributed analytics platform for analyzing healthcare data – employing the Personal Health Train paradigm – we propose to implement a solid data FAIRification infrastructure. This will allow us to achieve findability, accessibility, interoperability, and reusability of data, metadata, and results within a network of several medical centers participating in the BETTER Horizon Europe project, where the study of rare diseases (such as intellectual disability and inherited retinal dystrophies) will be targeted. Impacts will be visible to a large population of healthcare practitioners, prospectively influencing health policymakers.

Keywords

FAIR principles, healthcare, distributed analytics, rare diseases

1. Introduction

Data-driven medicine is a crucial research area for the achievement of a more and more high-quality accessible healthcare system. Typically, the more data available for the intended analysis, the higher the chance to achieve accurate results [1]. However, the amount of available patient data is critical, especially in the context of rare diseases; here, even more predominantly than in other diseases, data sets are available and usable only at single medical centers. Reasons for the lack of data sharing are connected to ethical, legal, and privacy aspects and rules. Data centralization is a viable option due to privacy concerns, particularly within the European Union, where the General Data Protection Regulation (GDPR) imposes stringent privacy standards.

The Horizon Europe project BETTER (Better rEal-world healTh-daTa distributEd analytics Research platform, https://www.better-health-project.eu/), started Dec. 1st, 2023, has proposed the design and implementation of a decentralized infrastructure that will allow us to exploit the full potential of large sets of multi-source health data. This will be achieved by using customized AI tools to compare, integrate, and analyze datasets in a secure as well as cost-effective fashion. The project will target various use cases involving 7 European medical centers; they will provide sensitive patient data, including, possibly, clinical reports, medical images, genomic data (whole-exome, whole-genome sequences), biological data (cellular and molecular pathways), metabolic,

Submitted to the 15th International Semantic Web Applications and Tools for Health Care and Life Science conference

*Corresponding author.

anna.bernasconi@polimi.it (A. Bernasconi); cinzia.cappiello@polimi.it (C. Cappiello); stefano.ceri@polimi.it (S. Ceri); pietro.pinoli@polimi.it (P. Pinoli)

0000-0001-8016-5750 (A. Bernasconi); 0000-0001-6062-5174 (C. Cappiello); 0000-0003-0671-2415 (S. Ceri); 0000-0001-9786-2851 (P. Pinoli)

CEUR Workshop Proceedings (CEUR-WS.org)

environmental and demographic data, patient interviews, forms, and therapies details. Only the secure information will be made available and analyzed with a GDPR-compliant mechanism via a Distributed Analytics paradigm called the Personal Health Train (PHT) [2].

As a technical partner of the project, Politecnico di Milano (i.e., the authors) will particularly focus on BETTER's objective to guide medical centers in collecting patients' data following a common schema in order to promote interoperability and re-use of datasets in scope. This includes legal/ethical data protection authorizations as well as data documentation, cataloging, and mapping to well-established ontologies. Attention will be devoted to the FAIRification of handled data. Legal and ethical implications will be duly considered, and data access and re-use procedures will be proposed. Data pseudonymisation will be performed as a default preprocessing step, mitigating the risk of personal data leaks; this will be followed by data quality and integrity assessment. A real-world large-scale data integration framework (based on well-established ontologies) will be demonstrated taking into account heterogeneous datasets, including those obtained by whole genome sequencing.

The platform will be tested primarily on two rare disease use cases, inherent to pediatric intellectual disability and inherited retinal dystrophies; the final goal is to keep a high generality, facilitating extension to various cases within the context of the European Data Space.

The rest of the manuscript briefly describes the PHT paradigm and illustrates our plan for ensuring that FAIR principles [3] (i.e., Findability, Accessibility, Interoperability, and Reusability) are guaranteed in the results of the project. Finally, we show how two use cases regarding rare diseases can benefit from such an outcome.

2. A distributed analytics paradigm: The Personal Health Train

The intuition behind BETTER can be explained via a railway system analogy that includes *trains*, *stations*, and *train depots*. The trains use the network to visit different stations to transport several goods, which in this analogy correspond to analytical tasks. By adapting this concept to BETTER, the analytical task is brought to the data provider (i.e., a medical center), whereas the data instances remain in their original location (called *station*). Two mature implementations of PHT have been implemented in the past: 1) PADME (Platform for Analytics and Distributed Machine Learning for Enterprises, https://padme-analytics.de/) and 2) Vantage6 (priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange, https://distributedlearning.ai/). The first, PADME has been developed in the context of the University of Cologne and proven successful in several clinical use cases in Germany. Vantage6 [4] was released by Maastricht University and applied in many real-world healthcare use cases (e.g., oncology, cardiovascular diseases, diabetes type 2). These platforms represent the starting point for the BETTER platform, showing that federated learning in the healthcare domain is technically feasible. BETTER takes this paradigm to the challenging context of a European-level project.

3. Scalable data FAIRification

Methodology. A solid infrastructure that is able to organize and share the needed information at the central level (thus enabling pair-wise interchanges between the data providers' stations) is needed. For what regards metadata, we will be inspired by the DAMS approach [1], using a general metadata schema to allow distributed analytics infrastructures to comply with FAIR

principles. This will deal with business information (e.g., about social entities), technical information about the train, and data information about the data provider. Differently, clinical data types and related metadata are typically specific to the context of use, leveraging the characteristics of the disease, of patients, and relevant parameters for the problem at hand. BETTER is prepared to address the data management problem with an extremely general approach. As these data types are not covered in DAMS, their management will be inspired by our extensive previous work in the field (during the "Data-driven Genomic Computing" ERC AdG n. 693174, running 2016-2021). More specifically, four directions in the agenda of BETTER will be followed to guarantee the scalability of semantic/syntactic standards of clinical data types:

1. Interoperability at the level of the same pathology will be guaranteed by having the partners generating datasets agree upon the same standards.
2. We will employ a data schema that captures the main properties of a generic clinical context, keeping a high abstraction level to encourage maximum interoperability (successful examples are the Genomic Conceptual Model [5] and the COVID-19 Host Genetics Initiative Data Dictionary [6]). Typically, clinical data involve demographic (or static) information on the patient and longitudinal measurements related to medical encounters, treatment, or laboratory measurements.
3. We will use a key-value paradigm for that information that is not shared among different pathologies and that is specific to a given use case, thus creating a very flexible and expressive data model that allows storing all relevant information without dealing with integration and interoperability at the storage level (see [7]).
4. Semantic annotation will be made by using dedicated biomedical ontologies as we described in [8], sourcing them from BioPortal (https://bioportal.bioontology.org/) and Ontology Lookup Service (https://www.ebi.ac.uk/ols4). In this way, we will pursue complete semantic interoperability between the metadata associated with known ontology.

For genomic data, the BETTER project will initially acquire DNA and RNA sequencing data in both FASTQ and BAM formats. All submitted sequence data will be aligned using the latest human reference genome; variant and mutation calls will output VCF and MAF formats, whereas gene and miRNA expression quantification data will be kept in TSV format. Other genomic signals for tertiary data analysis will be homogenized according to guidelines of the Global Alliance for Genomics & Health (https://www.ga4gh.org/).

FAIRification will be achieved by researching and developing dedicated preprocessing and ETL (Extract, Transform, and Load) pipelines at each medical center where machine learning, NLP, and Human-in-the-Loop techniques will be used to automatize some of the steps. Importantly, data pseudonymization will be performed by default, and data quality and integrity will be assessed and monitored throughout the project. User-friendly FAIRification instruments will be preferred (see https://github.com/maastrichtu-cds/epnd-fairification). Finally, dedicated ETL processes will be developed to enable BETTER to interoperate with public health registries, European Health Data Space (EHDS), the 1+Million Genomes initiative (1+MG), and the European Open Science Cloud.

Working plan. As a first step, we will *catalog datasets available at each medical center.* Multiple focus groups will be organized with both technical and clinical stakeholders to understand

in depth the available datasets, more specifically: (1) dataset characteristics and size; (2) data types with their attributes and value ranges; (3) pathology-related interpretation; (4) examples of data usage in real-world scenarios.

Secondly, we will tackle *Data Pseudonymisation.* By default, data will be pseudonymized before joining the BETTER platform, which requires the implementation of modules for: (1) identifying personal data from images and text; (2) pseudonymization of reference ID to preserve leakage between same patient samples; (3) where applicable, defacing of face images.

Thirdly, we will *Design and develop a unified schema repository for medical centers' data and metadata integration.* A unifying global model will be designed to accommodate all the data formats and their describing metadata, and serve as a reference for the next analysis steps.

Finally, we will deal with *FAIRification of medical centers' datasets.* We will develop ad-hoc ETL pipelines to onboard health datasets to BETTER; this task will achieve data FAIRification by scheduling transformation functions to adjust the initial content into appropriate destination formats. Medical-center-specific data formats, protocols, and characteristics will be mapped to a standard schema, enabling interoperability and federated learning.

4. Application to rare diseases

Intellectual disability (ID) is a rare pediatric disease, a common disorder characterized by significant limitations of cognitive functions and adaptive behavior, with onset before the age of 18. It is estimated that approximately 1-3% of the global population has some form of intellectual disability [9]. It occurs as a unique phenotype or in the context of rare forms of disease such as syndromic neurodevelopmental disorders (NDDs, e.g., Coffin-Siris syndrome) and inborn errors of metabolism (IEM, e.g., phenylketonuria) with neurological involvement. This use case aims to elucidate how different etiologies of rare diseases and ID complex diseases lead to convergent or divergent molecular mechanisms that underlie brain development and the mode of disease in children and adolescents with ID. Many data types will be integrated, including clinical data, brain images, genomic data (whole-exome, whole-genome sequences), and biological data (cellular and molecular pathways). The proposed use case will have a substantial impact on: (1) The health and social management of intellectual disability. From the knowledge and scientific point of view, recovering new data from patients will facilitate the generation of new emerging paradigms in ID. (2) The integration of genomic data in the newborn screening will have a strong impact on health care and public health—possibly, on the speed of the diagnosis and the possibility of diagnosing disease that is not possible to identify with the metabolic screening.

Inherited Retinal Diseases (IRDs) are a group of disorders characterized by the generally progressive death or dysfunction of photoreceptors and retinal pigment epithelium cells, leading to loss of visual function, sometimes leading to legal blindness. It is estimated that this group of diseases affects 1 in 3,000 people. IRDs are clinically very heterogeneous and can be classified according to multiple parameters; in addition, IRDs present a high allelic and genetic heterogeneity. An early molecular diagnosis is necessary to confirm the clinical diagnosis, offer adequate care to patients, give genetic and reproductive counseling to families, choose the most appropriate educational methods, and include in appropriate clinical trials based on genetic information. Different datasets will be employed, e.g., genomic data (gene panels, clinical exome, whole exome, whole genome), clinical reports, and images. This use case aims to develop

algorithms to increase the percentage of successfully diagnosed patients when compared with the success rate of targeted and/or whole exome sequencing or other pipeline analysis. The patient and society would greatly benefit from this approach because: (1) a higher probability of correct diagnoses would allow a more precise treatment plan for each patient; (2) accelerating the genetic confirmation of the disease will lessen clinical visits and, as a result, the disease's economic burden.

5. Conclusion

The BETTER project relies on "bringing computation to data" via incremental and federated learning, which avoids unnecessary data moving across medical centers while exploiting much of the information encoded in such data. The project leverages past expertise gained in the implementation of the PADME and Vantage6 projects, as well as in the health/genomic data integration expertise of the Data-driven Genomic Computing project. In accordance with the European Health Data Space (EHDS), BETTER will enable EU medical centers and beyond to make full use of the potential offered by a safe and secure exchange, use, and reuse of health data fostered by robust data FAIRification. In the context of intellectual disability and inherited retinal dystrophies, – with a large potential of expanding the same paradigm to other diseases – the generated analytical tools will help healthcare professionals become more proficient in cutting-edge digital technologies, data-driven decision support, health risk surveillance, and control activities, monitoring and management of healthcare quality levels, with positive repercussion also on health policymakers, and innovators in general.

Acknowledgements. The work is supported by BETTER, Grant agreement 101136262.

References

[1] S. Welten, et al., DAMS: A distributed analytics metadata schema, Data Intelligence 3 (2021) 528–547.

[2] O. Beyan, et al., Distributed analytics on sensitive medical data: the personal health train, Data Intelligence 2 (2020) 96–107.

[3] M. D. Wilkinson, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific data 3 (2016) 160018.

[4] A. Moncada-Torres, et al., VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange, in: AMIA annual symposium proceedings, volume 2020, American Medical Informatics Association, 2020, p. 870.

[5] A. Bernasconi, et al., Conceptual modeling for genomics: building an integrated repository of open data, in: ER 2017, Springer, 2017, pp. 325–339.

[6] A. Bernasconi, et al., A review on viral data sources and search systems for perspective mitigation of COVID-19, Briefings in Bioinformatics 22 (2021) 664–675.

[7] M. Masseroli, et al., Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying, Methods 111 (2016) 3–11.

[8] A. Bernasconi, et al., Ontology-driven metadata enrichment for genomic datasets, in: SWAT4HCLS 2018, volume 2275 of *CEUR Workshop Proceedings*, 2018.

[9] P. K. Maulik, et al., Prevalence of intellectual disability: a meta-analysis of population-based studies, Research in developmental disabilities 32 (2011) 419–436.