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

doi:10.3233/SHTI230497

From Raw Data to FAIR Data: The FAIRification Workflow for Brazilian Tuberculosis Research

Filipe BERNARDI[a,1], Vinicius LIMA[a], Gabriel SARTORETTO[a], João BAIOCHI[a], Victor CASSÃO[a], Afrânio KRITSKI[b], Rui RIJO[c] and Domingos ALVES[a]
[a] *Ribeirão Preto Medical School, University of São Paulo, Brazil*
[b] *Faculty of Medicine, Federal University of Rio de Janeiro, Brazil*
[c] *School of Technology and Management, Polytechnic Institute of Leiria, Portugal*
ORCiD ID: Filipe Bernardi https://orcid.org/0000-0002-9597-5470

Abstract. Among the main factors that negatively influence the decision-making process, it is possible to highlight the low quality, availability, and integration of population health data. This study aims to highlight the difficulty of research based on tuberculosis data available in Brazil. The FAIR methodology is a solution for standardizing data and sharing information about the disease. All the main actors involved, including those who generate data and administrators of information systems, should be encouraged to know their strengths and weaknesses. Continuously fostering strategies to promote data quality is, therefore, a strong stimulus for strengthening national health information systems and can potentially benefit from recommendations on how to overcome the inherent limitations of these information systems. Data quality management in Brazilian tuberculosis information systems is still not carried out organized and systematically. According to the FAIR principles, the evaluation demonstrates only 37.75% of compliance.

Keywords. Tuberculosis, Data governance, Data quality

1. Introduction

Although tuberculosis (TB) is entirely curable and preventable, it affects individuals who lack sustainable living conditions and easy access to information. In Brazil, patients with TB infections who do not have private health plans can be treated by public hospital healthcare systems through the Unified Health System (SUS). The SUS has established a series of programs specifically aimed at controlling, dining, and supporting patients infected with TB. One is the National Program for TB Control, which aims to eliminate TB as a public health problem in Brazil. However, challenges must still be overcome to achieve this objective [1].

Most notably, there is a need to create, adapt, and introduce new diagnostic and treatment technologies and contribute to research on innovative technologies. Strengthening control and information systems (IS) for the disease in vulnerable populations, who are the most affected, is also necessary. Activities such as public health

[1] Corresponding Author: Filipe Bernardi, Ribeirão Preto Medical School, University of São Paulo, Brazil . Email: filipeandradebernardi@gmail.com.

program control and organization, health facility management, ensuring the availability of medical supplies, and the lack of instruments may hinder the quality of care from supporting the definition of management parameters [2].

Among the main factors that negatively influence the decision-making process, it is possible to highlight the low quality, availability, and integration of population health data. Several computerized systems are created to follow up information about patients with TB in practical and quick ways. These systems prioritize collecting data such as personal information from patients, medications, treatments, diagnostic tests, and control of routines. The IS addressed in this work were: SISTB (a local IS), TBWEB – Tuberculosis Patient Control System of São Paulo State; SINAN – Information System for Notifiable Diseases and GAL – Laboratory Environment Manager, both are national systems, but for data accessibility reasons, we used samples from the state of Rio de Janeiro.Despite the existence of such data, some reasons make it difficult for managers and healthcare professionals to access them, such as the lack of digitization of processes, heterogeneity, and duplication of data in health IS, and the existence of a large amount of isolated data accessible only in a particular context [3].

To address these difficulties, such as the lack of expansion of scientific discoveries and the discovery of new knowledge, low communication between obtained data, and non-easily reproducible data, the Global Open FAIR (Findable, Accessible, Interoperable, and Reusable) initiative was created. It aims to disseminate a large-scale recommendation that supports better management of research data in open science, access to research data, and scientific information by machines and humans. The FAIR methodology is a solution for standardizing data and sharing information about the disease [4]. So, this study highlights the difficulty in conducting research based on TB data available in Brazil.

2. Methods

Due to the information heterogeneity in the systems above, the FAIR principles can be considered weak and ambiguous enough to lead to different interpretations. Thus, it is necessary to clarify its meaning and define criteria to evaluate the data FAIRification process. Since the maturity of FAIR data depends on the capabilities of the ecosystem components, the FAIR assessment should include not only FAIR data as an outcome but also a specific assessment of criteria relevant to each essential element of every process involved in the flow of information from systems, taking into account the combination of data and services they [5]. To calculate the metrics centered on the data present in the TB IS, that is, on the variables they collected, the dimensions described by Bernardi et al. (2022) [6] and the instrument validated by the European project “Fostering FAIR Data Practices in Europe” (https://fairaware.dans.knaw.nl/).

3. Results and Discussion

The quality and relevance of the information produced to know the population's health conditions may be compromised when there are variables with inadequate completion. Incomplete data make it impossible to assess other quality dimensions and use techniques that allow the crossing of information. Juxtaposed, overcoming such challenges can improve the filling quality and expand the scope of use of this information in

epidemiological studies and decision-making. The SATIFYD questionnaire was designed to understand the compliance of the datasets evaluated against the FAIR principles. In this sense, the higher the score, the more synchronized the dataset is with the FAIR Guidelines. Table 1 describes the TB IS analyzed and their compliance with the FAIR principles.

Table 1. Brazilian Tuberculosis Information Systems (BTIS) compliance with the FAIR principles.

BTIS	Informational Level	Total Records	Timeframe (2000's)	FAIR Element %	Total FAIR %
TBWEB	State (São Paulo)	208.624	06 - 19	F:55 A:50 I:25 R: 22	38%
SINAN	National (State Sample)	104.541	01 - 14	F:27 A:50 I:25 R: 22	31%
GAL	National ((State Sample)	11.651	10 - 17	F:66 A:50 I:25 R: 22	41%
SISTB	Municipal (Ribeirão Preto)	4.065	00 - 22	F:66 A:50 I:25 R: 22	41%

The validation and adequacy of filling out collection instruments and databases of TB IS requires monitoring and evaluating completeness, which helps identify data weaknesses and strengths and recommend strategies to improve information quality. Completeness is still a little-explored quality dimension in TB IS in Brazil, and data often needs to be standardized or transformed. Compliance check is commonly implemented from pre-established guides and reference tables by organizations to ensure conformance to policies and standards.

These standards ensure that tasks are performed correctly and provide order within the organization. Complying with the same standards will prevent many errors, such as data duplication, capturing incorrect spellings, and using inaccurate formats. Despite this, the absence of standards observed in TB IS makes it impossible to compare their databases. The impossibility of reaching databases also reveals a barrier to data reliability, represented by precision. Precision is the degree to which the data show the truth about the described event. The sine qua noncondition for the existence of accurate information is preceded by complete and correctly represented data.

Coding errors and poor documentation are the primary reasons for data non-validation in information systems. While post-submission curation can improve data quality retrospectively, implementing practical data management solutions early in project design is crucial. FAIR principles, which promote better data availability and reuse, address interoperability and harmonization issues with health information systems. These requirements aim to improve clinical decision-making and are summarized by Kodra et al. [7] in their perspective on 'quality informatics' using health data collected from IS.

This study suggests that studies on improving data quality in information systems can help identify problems such as inadequate data collection instruments, lack of researcher training, and technical document review. Strategies that will enhance data quality, including integrated databases, can enable the recovery of incomplete or inconsistent data in Brazilian IS for TB. The study notes the use of metrics to support data evaluation according to the FAIR principles, where a compliance rate of 37.75% was achieved.

These findings indicate no significant gap between awareness of the need to implement the FAIR guidelines in health IS and the incorporation of data quality promotion activities based on the FAIR guidelines. Therefore, all the main actors involved, including those who generate data and administrators of IS, should be encouraged to know their strengths and weaknesses. Continuously fostering strategies to promote data quality is a potent stimulus for strengthening national health IS and can

potentially benefit from recommendations on how to overcome the inherent limitations of these IS.

The FAIRfication process for tuberculosis data in Brazil faces data quality, privacy issues, access difficulties, technological limitations, and different data sources. Future work to address these limitations includes improving data quality, developing ethical guidelines for data availability, enhancing technological infrastructure, standardizing data formats, and investing in research. These initiatives can ensure interoperability, availability, and reuse of tuberculosis data and promote disease control.

4. Conclusion

Data quality management in the Brazilian TB IS is not yet systematically carried out. The evaluation of only some parts of the information production cycle relies solely on specific and independent metrics. This necessitates further research in this field to develop systematic methods that consider the specific characteristics of each computerized scenario and identify their potential contributions to improving the quality of TB information.

Acknowledgments

This work was developed within the Bioengineering Postgraduate Program of the University of São Paulo. The São Paulo Research Foundation (FAPESP) - grant number 2020/01975-9, also supported this work, coordinated by author DA.

References

[1] Do Carmo IA, Maia JC, De Novaes JV, Almeida Lde, Pereira NA, da Costa GV, et al. Os Desafios para o controle da tuberculose no Brasil. Brazilian Journal of Health Review. 2022;5(6):23969–78. doi: 10.34119/bjhrv5n6-168

[2] Makeleni N, Cilliers L. Critical success factors to improve data quality of electronic medical records in Public Healthcare Institutions. SA Journal of Information Management. 2021;23(1). doi: 10.4102/sajim.v23i1.1230.

[3] Thapa C, Camtepe S. Precision Health Data: Requirements, challenges and existing techniques for data security and privacy. Computers in Biology and Medicine. 2021;129:104130. doi: 10.1016/j.compbiomed.2020.104130

[4] Courtot M, Gupta D, Liyanage I, Xu F, Burdett T. BioSamples database: Fairer samples metadata to accelerate research data management. Nucleic Acids Research. 2021;50(D1). doi: 10.1093/nar/gkab1046

[5] Jacobsen A, Kaliyaperumal R, da Silva Santos LO, Mons B, Schultes E, Roos M, et al. A generic workflow for the data fairification process. Data Intelligence. 2020;2(1-2):56–65. doi: 10.1162/dint_a_00028

[6] Bernardi FA, Alves D, Crepaldi NY, Yamada DB, Lima VC, Rijo RPCL. Data Quality in health research: an integrative literature review. medRxiv, 2022-05. doi: 10.2196/preprints.41446.

[7] Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier S, van Enckevort D, et al. Recommendations for improving the quality of rare disease registries. International Journal of Environmental Research and Public Health. 2018;15(8):1644. doi: 10.3390/ijerph15081644.