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

doi:10.3233/SHTI230557

A Computational Infrastructure for Analyzing Tuberculosis Research Data in Brazil

Mariana MOZINI[a,1],Filipe BERNARDI[b], Ana Clara MIOTO[b],Victor CASSÃO[b], Afrânio KRITSKI[c] and Domingos ALVES[a]
[a] *Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil*
[b] *Bioengineering Postgraduate Program, University of São Paulo, São Carlos, Brazil*
[c] *Faculty of Medicine, Federal University of Rio de Janeiro, Brazil*
ORCiD ID: Mariana Mozini https://orcid.org/0000-0002-6235-7000

Abstract. Tuberculosis (TB) is one of the infectious diseases that currently causes the most deaths, with 6.4 million new cases recorded in 2021. Although it is a curable disease, drug-resistant strains emerge due to a lack of hygiene and low-quality or inappropriate medications, among other factors. With this in mind, the World Health Organization initiated the End TB Strategy campaign to improve the health system in the fight against tuberculosis. For this, reliable and high-quality health data is necessary to create effective public policies. However, despite technological advancements such as emerging concepts like Big Data and the Internet of Things, generating health information faces several obstacles. Therefore, the present work aims to describe a pipeline for TB research in Brazil to contribute to obtaining high-quality data.

Keywords. Tuberculosis, Data Science, Research data

1. Introduction

Tuberculosis (TB) is the second highest cause of death caused by a single infectious agent, surpassing HIV/AIDS. Despite being a curable and preventable disease, in 2021, 6.4 million new cases were recorded, resulting in 1.4 million deaths among HIV-negative individuals and 187,000 deaths among HIV-positive individuals [1]. The emergence of drug-resistant strains can be attributed to low-quality medication, poor hygiene, use of inappropriate medication, and delayed treatment approaches [2].

Drug-resistant TB, which is expensive and requires a long treatment period, needs new, affordable, and effective diagnostic tools that guarantee quality and proven efficacy to be rapidly implemented [3]. In addition, health services should incorporate new information systems to aid decision-making. The World Health Organization launched the Stop TB Strategy in 2006, followed by a new strategy in 2015 with more ambitious goals and a greater focus on research and innovation. Although both

[1] Corresponding Author: Mariana Mozini, E-mail: mtmozini@usp.br.

approaches have significantly reduced TB cases in high-burden countries, multidrug-resistant tuberculosis remains a global problem [4].

Data reliability is critical for improving health service quality and creating effective public policies. However, generating health information is faced with barriers such as problems with patient data documentation, data interpretation difficulties, and organizational issues [5]. In TB research, a large amount of complex data is produced from dispersed sources with low integration and varying accuracy levels. It impedes knowledge extraction and data analysis, making it challenging to provide decision support in operational and administrative processes and scientific research [6].

Currently, data coordination between TB stakeholders parties can be messy, prone to delays, subject to manipulation, and obscure. Thus, this work aims to describe a computational pipeline and the infrastructure needed for analyzing TB research data, assisting in establishing a high-quality data source in Brazil.

2. Methods

2.1. Brazilian Tuberculosis Research Network Ecosystem

The most used diagnostic methods for TB consider bacteriological, radiological, histopathological, and immunological approaches. The bacteriological tests consist of bacilloscopy and culture. Clinical materials such as sputum, bronchial and bronchoalveolar lavage, and other samples that can be taken from the respiratory tract are used for TB research [2].

The clinical laboratory has a fundamental role in the health system, given that most medical decisions are made using the information provided by laboratory processes. Quality assurance in a clinical analysis laboratory is built on all process stages, from the material collection (pre-analytical) to the result (post-analytical) delivery. Clinical samples sent to the laboratory for TB diagnosis must comply with a series of general conditions on which the quality and efficiency of the test results depend. It is essential to control the data quality from local centers belonging to the Brazilian TB research network, which covers 65 institutions and researchers [3].

2.2. Data Gathering, Infrastructure, Curation, and Analysis Pipeline

The TB Network conducts studies that require the collection and management of data in scientific, clinical, and managerial/epidemiological domains. The Research Electronic Data Capture (REDCap) platform is used for this purpose, which is a web-based application that allows the creation of case report forms, surveys, and research databases. REDCap is an open-source application that provides several tools for exporting data in multiple formats, including the CSV standard that facilitates compatibility with other third-party tools.

The collected data are categorized and segmented inside the network through the projects, and participants have restricted access based on the network policies. Data are automatically anonymized to ensure ethical, legal, and confidential issues, and their confidence rate level can be chosen. Before analysis, the data undergoes pre-processing to identify and treat missing and abnormal values and duplicate and redundant data. Validations are also carried out to ensure the types of variables and transformations, such as normalization and discretization.

The exploratory analysis is the next step in the analysis pipeline, where the data's distribution and amount are better understood. Descriptive statistics such as the mean, median, and standard deviation are calculated. Graphics such as bar and scatter charts and visual elements such as presentations and images are used to understand the results better. Regression and clustering calculations can also be used to find patterns and relationships in the data that are not easily identifiable.

2.3. Auxiliary Tools

The collaboration scripts for developing machine learning and artificial intelligence algorithms are developed in Google Colab, a Jupyter Notebook-based platform. It provides an accessible and readily available environment for developing Python programs. Python is popular due to its high-level nature and simplicity, which enables portability across different platforms, and it has a vast developer community and numerous libraries. Specific Python libraries used in programming tasks include Numpy for mathematical functions such as linear algebra, Pandas for data analytics functions, and Scikit-learn for machine learning algorithms. Streamlit, a Python-based data visualization platform that allows data transformation into shareable pages without prior knowledge of other programming languages, was adopted for data visualization. Streamlit can be easily integrated with Python applications, requiring minimal adaptation to receive Python outputs, and is an efficient way to display data without using large tables.

2.4. TBWeb Application for TB Analysis

A research network developed a web application to validate statistical analyses related to project data. The portal can connect to any database and offers real-time statistical analysis of clinical data through data visualization techniques. Researchers can monitor all updates in the data in real time.

3. Expected Outcomes

The expected primary outcomes of this work are tools and a curated knowledge basis for frameworks that can promote clinical data quality within the Brazilian tuberculosis stakeholders through a national research network. Over the medium time, we expect to also encourage new models of data sharing (e.g., safe data havens, data lakes, data hubs) and innovative privacy-preserving and processes analytical methods. Also, it is expected, in the long term, to obtain positive health outcomes, such as developments in public health indicators, a better understanding of health services processes, improved research outcomes, and new approaches to ethical and legal issues.

Although the final goal is to implement this pipeline to the national network in Brazil, in phase 1, it will be first deployed and validated in 7 research centers, leading to different data types. Data from several government sources will be gathered to build a test database for the computational tool and validate data sharing among peers. Our success indicators will be based on the following parameters: adoption of the tool by the TB Network entities; the volume of data available and used; and perception of

usefulness for academic, clinical, and managerial audiences. After defining these parameters, a validation process for an official implementation as a national tool for TB will be carried out with the TB experts independently committee.

Based on a well-established process and robust evidence, we hope we can compare our results with the source documents of each center. It will allow checking the validity of a sample of the data entered on the form and define intervention, such as visiting the local centers to improve the data culture and promote continuous education. We expect also to be enabled to compare our solution with consolidated approaches like the Observational Health Data Sciences and Informatics (OHDSI) tools and the FAIR Principles.

4. Conclusions

It is expected to successfully build a high-quality data source to provide a basis for developing new decision-support tools. We hope to advance scientific research and establish new diagnosis algorithms and optimized operational models toward better patient care and managerial decisions. In the long term, it is expected to achieve positive health outcomes such as improved public health indicators, a better understanding of health service processes, and new approaches to ethical and legal issues.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP) - grant number 2020/01975-9, coordinated by author DA.

References

[1] Global tuberculosis report 2022 [Internet]. World Health Organization. World Health Organization; [cited 2023Mar28]. Available from: https://www.who.int/publications-detail-redirect/9789240061729

[2] Kanabalan RD, Lee LJ, Lee TY, Chong PP, Hassan L, Ismail R, Chin VK. Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. Microbiological research. 2021 May 1;246:126674. Available at: https://doi.org/10.1016/j.micres.2020.126674.

[3] Kritski A, Andrade KB, Galliez RM, Maciel EL, Cordeiro-Santos M, Miranda SS, Villa TS, Ruffino Netto A, Arakaki-Sanchéz D, Croda J. Tuberculosis: renewed challenge in Brazil. Revista da Sociedade Brasileira de Medicina Tropical. 2018 Jan;51:02-6. Available at: https://doi.org/10.1590/0037-8682-0349-2017.

[4] Uplekar M, Raviglione M. Who's end TB strategy: From stopping to ending the global TB epidemic. Indian Journal of Tuberculosis. 2015;62(4):196–9.

[5] Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. BMC health services research. 2017 Dec;17:1-0.Available at: https://doi.org/10.1186/s12913-017-2697-y.

[6] Bernardi FA, Alves D, Crepaldi NY, Yamada DB, Lima VC, Rijo RP. Data Quality in health research: an integrative literature review. medRxiv. 2022:2022-05. doi:10.1101/2022.05.31.22275804.