A web portal for real-time data quality analysis on the Brazilian Tuberculosis Research Network: A Case Study

Victor Cassão[1][0000−0002−8967−0267], Filipe Andrade Bernardi[1][0000−0002−9597−5470], Vinícius Costa Lima[2][0000−0002−2467−358X], Giovane Thomazini Soares[2][0000−0001−9273−3815], Newton Shydeo Brandão Miyoshi[2][0000−0002−2335−371X], Ana Clara de Andrade Mioto[1][0000−0003−4475−1984], Afrânio Kritski[3][0000−0002−5900−6007], and Domingos Alves[2][0000−0002−0800−5872]

[1] São Carlos School of Engineering, University of São Paulo, São Carlos/SP, Brazil
{victorcassao,filipepaulista12,anaclara.mioto}@usp.br

[2] Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto/SP, Brazil
quiron@fmrp.usp.br;viniciuslima@alumni.usp.br, newton.sbm@gmail.com, giovane.soares@usp.br

[3] School of Medicine, Federal University of Rio de Janeiro, Rio de Janeiro/RJ, Brazil
kritskia@gmail.com;

Abstract. Research projects with Tuberculosis clinical data generate large volumes of complex data, requiring sophisticated tools to create processing pipelines to extract meaningful insights. However, creating this type of tool is a complex and costly task, especially for researchers who need to gain experience with technology or statistical analysis. In this work, we present a web portal that can connect to any database, providing easy access to statistical analysis of the clinical data in real-time using charts, tables, or any other data visualization technique. The tool is user-friendly and customizable, reaching the project's needs according to its particularities. The developed portal in this work was used as a use case for the research project developed by the Federal University of Rio de Janeiro (UFRJ) for the validation and cost of performance of the Line Probe Assay 1 and 2 (LPA) as a method of diagnosing resistant Tuberculosis in Brazilian's reference centers. In the use case, the tool proved to be a valuable resource for researchers, bringing efficiency and effectiveness in analyzing results for quick and correct clinical data interpretation.

Keywords: Web Portal · Data Processing · Clinical Data · Statistical Analyzes.

1 Introduction

A web portal is a web-based repository of data, information, facts, results of analyzes, and also knowledge [1,2]. Portals can enable data search and filtering

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_24

[3], facilitating access to information of interest. In health, information is usually gathered from multiple sources and organized in a user-friendly way [4].

Web portals provide a unified interface for applications and databases that, without this approach, would be isolated elements [3]. Also, they are an efficient and effective form of communication and dissemination, allowing the production of knowledge and intelligence [2, 5]. Given the relevance of these products for planning processes [6], web portals in the health area [1] can assist in the planning of public health programs [7]. In this context, it is understandable that the content of web portals must be valid, trustworthy, coherent, representative, sensitive, comprehensive, and ethical [8, 5].

However, health web portals stand out for the lack of interoperability [2, 5], accessibility, and quality [9] that data in this area can present. Also, in the health domain, many web portals suffer from usability problems [10], undermining their effectiveness [11]. When we discuss health portals, we need to understand in which context and disease it is present. In Brazil, one of the diseases that is still a challenge, both in terms of control and cure, is Tuberculosis.

Tuberculosis (TB) is an infectious disease caused by the bacterium Mycobacterium tuberculosis, which can affect different organs, mainly the lungs. In most cases, it is a curable disease if the treatment is carried out correctly. In 2021, it was estimated that 10.6 million people globally were sick with TB, with 1.6 million deaths. The Brazilian scenario is also very challenging, with around 88099 cases reported in 2021, aggravated by the Covid-19 pandemic [12].

Several studies point to a higher disease prevalence in countries with low socioeconomic status, especially among vulnerable groups such as homeless people and people deprived of liberty [13]. The current cycle between poverty and illness is fed back between individuals and their social environment, influencing the outcome of TB treatment, which is the primary way to cure and reduce transmission of the disease.

It evidences problems like the low effectiveness and adherence to treatment, which is around 70%, the national average [14]. Between its causes, we can mention treatment abandonment (people that did not attend medical follow-ups and stopped taking all medications), incorrect use (people that use only prescribed medications), and irregular use of drugs (people that use in wrong periods). It is also worth highlighting that engagement problems can lead to treatment failure, drug resistance, and TB relapse [15].

It is still a worrying disease across the national and global scene, so WHO has developed the End TB Strategy. The main objectives of this mission are to zero the number of deaths and people living with the disease through 3 pillars: the first - integrated, patient-centered care and prevention; the second - bold policies and supportive systems and the third - intensified research and innovation [12].

2 Objectives

This paper aims to present the development, evaluation, and usability validation of a new web portal to integrate tuberculosis databases to a data processing pipeline for clinical data analysis using initially statistical methods and support for future use of machine learning based approaches. The results of this study will provide insights into the potential of web-based tools for clinical data analysis on the REDE-TB project and contribute to the development of more efficient and effective data analysis methods for improving patient outcomes.

3 Methods

In this section, the context in which this research was developed will be detailed, highlighting the dataset selection as a use case for this work. The data pre-processing phase and the tools and technologies will also be presented. Figure 1 shows the detailed workflow of how the application works, from the data acquisition to the pre-processing and the tools to make information available to the user. These steps will be detailed next.

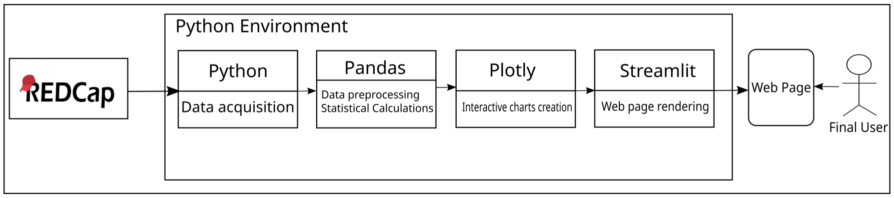


Fig. 1. Software workflow.

3.1 Scenario: Research Data on TB in Brazil

Created in 2001 by an interdisciplinary group of Brazilian researchers, the Brazilian Tuberculosis Research Network (REDE-TB) aims to promote interaction between government, academia, health service providers, civil society, and industry in developing and implementing new technologies and strategies to improve tuberculosis control across the country. REDE-TB is organized into ten research areas, comprising around 320 members from 65 institutions representing 16 of the 27 Brazilian states [16].

Data collection and management of studies carried out by REDE-TB are done through the Research Electronic Data Capture (REDCap) platform. Developed by Vanderbilt University and available through an open-source license, REDCap is a web application that manages case report forms (CRF), surveys, and research databases online.

REDCap also provides tools for exporting data in different formats. The standard used in developing this project was the Comma Separated Values (CSV), which is easy to handle through other third-party software.

Currently, the REDE-TB REDCap instance has 21 projects in progress and 261 registered users, which has a mechanism for sharing and tracking data captured through the software. [17].

3.2 Use Case: Line Probe Assay (LPA) Project

The LPA project started in 2019 and aims to evaluate the performance and cost of the Line Probe Assay 1 and 2 techniques in reference centers in Brazil to diagnose drug-resistant tuberculosis compared to the Xpert and MGIT 960 methods. It is an analytical, prospective, and multicenter study that aims to incorporate the method in the diagnostic routine. Reference centers with sputum samples and MTB cultures from patients with TB processed in 9 (nine) local reference laboratories for mycobacteria in the following Brazilian states participate in the study: Rio de Janeiro, São Paulo, Amazonas, and Bahia.

The study population is adults (18 years or older) with symptoms of pulmonary TB whose clinical specimens are positive on Xpert MTB/RIF and who have not received treatment for sensitive, drug-resistant, or multidrug-resistant TB for more than seven days in the past 30 days. Thus, it is expected to estimate, compare and analyze the diagnostic accuracy of the LPA compared to the gold standard adopted in health services. Secondary results aim to identify barriers and facilitators for implementing a quality system in participating laboratories and analyze the time elapsed between screening and a) detection of resistant TB; b) initiation of patient treatment.

The study variables distributed in 14 CRFs were defined and divided into primary and secondary data. The primary data are gathered directly from health centers, where health agents collect data through interviews with eligible patients. Demographic, socioeconomic, recruitment, and patient eligibility data are collected in these cases. Secondary data are collected through the results obtained in tests performed by specialized laboratories participating in the study.

The research evaluates the accuracy of the test by using the Positive Predictive Value (PPV) and Negative Predictive Value (NPV). PPV represents the proportion of patients who tested positive, and NPV for the negatives, in a given exam result, e.g., rifampicin sensitive, isoniazid sensitive. Those metrics show how many patients were correctly diagnosed for both cases, with positives and negatives. Through this metric, we can compare the results of different exams and evaluate their accuracy to extract information about their performance and viability. Furthermore, the clinical impact between the time of sample collection and resistant TB detection is an important indicator of the research, especially, to reach the objective of screening time reduction and appropriate treatment initiation. This also impacts the reduction of mortality, treatment abandonment, and TB transmission in the community.

Regarding data quality, managing data at source and applying Findable, Accessible, Interoperable, and Reusable (FAIR) Guiding Principles for its transparency use are recognized as fundamental strategies in interdisciplinary scientific collaboration.A standard protocol designed for this study has been used based on these principles and in applied routinely algorithm data-driven monitoring. This algorithm is responsible for data verification according to the metrics mentioned before.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_24

3.3 Knowledge Discovery in Databases Methods

This work used the main processes and methods of Knowledge Discovery in Databases (KDD). KDD is a well-known knowledge extraction process from large databases that aims to improve data quality. It is divided into well-organized steps that aim to reach that finality. Figure 2 exemplifies the steps of the organization on KDD.

The first stage, data selection, was carried out with the support of a technical team that computerized the entire data acquisition process through electronic forms deployed on REDCap. Health agents are responsible for entering study data directly into the system. To increase the quality of this collected data, actions and procedures are adopted before and after data insertion in REDCap. Before the data entry phase, manuals and training are provided for users. The subsequent approach to data insertion is their validation, aiming to identify potentially false or inconsistent data with other information about the patients. Such validation is carried out by software developed by the team responsible for maintaining the REDCap platform and manually by the team members. After these procedures, data can be analyzed with more reliability.

The LPA data selected for analysis is divided into three main groups: socioeconomic data, exam results, and key dates. Socioeconomic data includes descriptive patient information, such as age, height, weight, BMI, skin color, gender, etc. These data are essential in profiling patients and identifying correlations with TB outcomes. Laboratory examination results provide information about the type of TB detected in the collected sample and for which medication was detected resistance. The selected variables present information on the results of gold-standard testing examinations such as Xpert, MGIT 960, LPA1, and LPA2. Auxiliary variables that discriminate the sample's resistance to certain medications, such as Rifampicin and Isoniazid, were also selected. The data belonging to key dates represent the beginning and end of two specific stages of the study. The purpose is to analyze the elapsed time between such dates to validate the time interval and its performance. The main selected dates are the date of collection, the date of release of examination results, the date of DNA hybridization, and other relevant dates.

The second stage of KDD, data preprocessing, has as its main purpose the improvement of data quality, making it possible to use it in machine learning algorithms, data mining, and statistical analysis [3]. It is a crucial stage in extracting knowledge from a clinical database because this type of data is stored in its raw format, often without any treatment, containing inconsistencies, noise, and missing data [4]. Another important process in this stage is data cleaning. The cleaning process aims to remove errors and inconsistencies. It is responsible for reducing noise, correcting or removing outliers, treating missing data, duplicate data, or any other routine that aims to remove anomalies in the database.

This stage focuses on correctly cleaning, treating, and standardizing invalid data and any noise or inconsistency. In the case of LPA use, the lack or inconsistency of data originates from various factors, such as errors in filling out the data on the form, delay in delivering the test results by the laboratories, or

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_24

some other unknown factor. Identifying these missing or inconsistent data cases is considered one of the software's functionalities so that the centers can request correct filling, avoiding problems for the patient and the study itself.

In the case of unavailable data, REDCap has, by default, specific data types that handle its absence and reason. Table 1 provides these different data types and their meaning.

Table 1. Codes for Missing Data.

Code	Description
NI	No information
INV	Invalid
UNK	Unknown
NASK	Not asked
ASKU	Asked but unknown
NA	Not applicable
NAVU	Not available

In addition to the types of invalid data provided by REDCap, there is also the possibility of data not existing. In these cases, this step is also responsible for transforming the data into the types provided by the platform. This process makes it possible to extract which patients have missing records in a standardized way, allowing for correct completion by the responsible center. Inconsistencies found in variables that store dates are also treated in this step. Errors in date insertion can cause inconsistencies in the elapsed time calculation(in cases where the end date is sooner than the start date). In these cases, the inconsistencies are not included in the treated database, avoiding anomalies in the analysis and requesting a correction.

The third step is data transformation. This step transforms the cleaned and pre-processed data into the correct format for data mining algorithms. The main need met by the data transformation step was the standardization of test results, making it possible to compare different tests. Some variables in laboratory test results are stored in different formats, making it impossible to compare them directly. An example of this transformation is the manipulation of data from the TSA and LPA1 test results to detect resistance to the drug Isoniazid. While the first is stored in a categorical data format, where each row represents the situation of the test result, the second is stored in boolean form, stored in several columns that indicate one of the possible test results as the column name, having true or false values in the rows to represent whether the condition is valid or not. The applied operation transforms this boolean data into categorical data, facilitating comparison with other tests and ensuring better analysis of their values through the available charts.

With all KDD steps performed, algorithms are applied to extract patterns and generate reports and dashboards. The results section provides more information on all analysis generated after data processing.

3.4 Tools and technologies

Python programming developed the dedicated web portal for the LPA project. Due to its versatility, all the particularities of the software were developed solely and exclusively using Python and auxiliary libraries.

The Pandas library was used for data manipulation, preprocessing, statistical calculations, and other operations on descriptive data analysis. It is currently one of the most famous libraries for data manipulation developed in Python.

Plotly is a library for creating different charts, with native integration with Pandas, making its use more straightforward in the project. Plotly can create interactive charts where the user can filter data by legend and periods, apply zoom in/out, download the chart as an image, and many other functionalities. These filtering options run in real-time, improving the user experience.

Streamlit is a Python library for web page creation, with no need for in-depth knowledge of front-end technologies. Due to its native rendering of tables, graphs, filters, and other features, it is an excellent tool for sharing data-oriented applications. It has a native integration with Plotly, rendering the graphics generated by it and all its functionalities directly on the web page.

4 Results

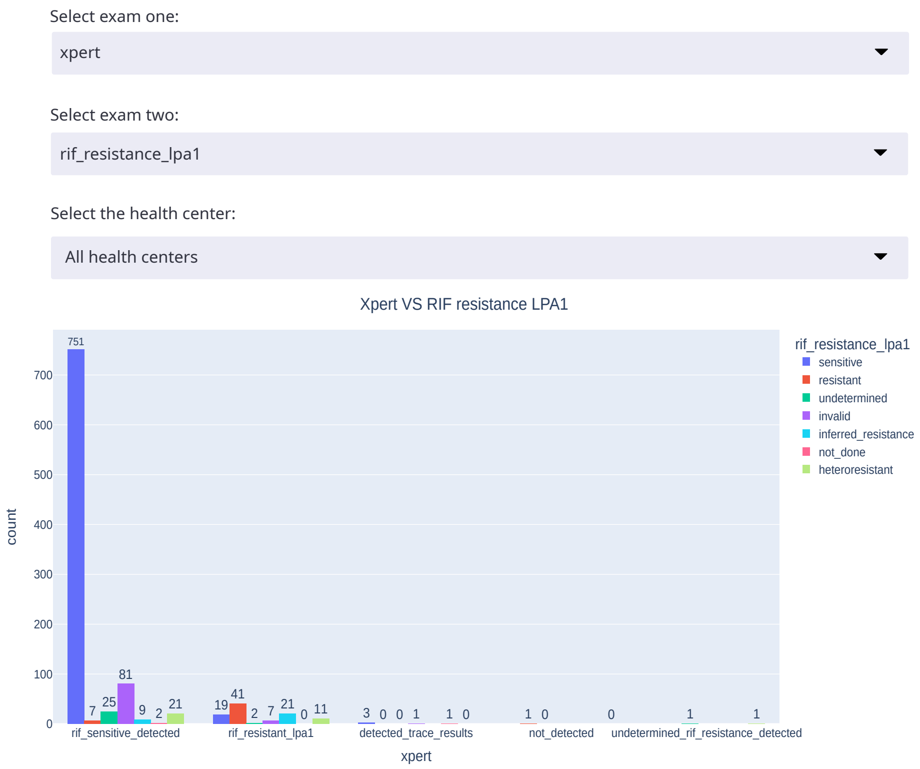
This section brings detailed information about the results of the analyzes available on the web portal and the insights extracted by the researchers who used the tool.

The web portal developed to fit the needs of the LPA use case can compare 11 different laboratory exam results with charts and tables to make knowledge extraction easier from it. It also can compare five key dates for elapsed time checking between crucial steps from research. More than 20 variables related to the patient's socioeconomic data are also available in the web portal for review.

The analysis of laboratory test results is interactive, where the user can choose between two different options, filtering by center or making a general comparison between them. The results are calculated according to the filter and dynamically rendered as a bar chart. The primary purpose is an accuracy check between the studied test (LPA) and all the other gold-standard TB tests. An example, a comparison between Xpert and LPA1 to detect resistance against Rifampicin is shown in Figure 2.

Besides comparing test results, mismatching and missing data were seen and removed from the cleaned dataset. From that, contacting the research centers and asking for data correction is possible. Figure 3 shows a bar chart with the relationship between the test results and missing values.

DOI: 10.1007/978-3-031-36024-4_24



Select exam one:

xpert

Select exam two:

rif_resistance_lpa1

Select the health center:

All health centers

Fig. 2. Comparison between Xpert and LPA1 to detect resistance against Rifampicin.

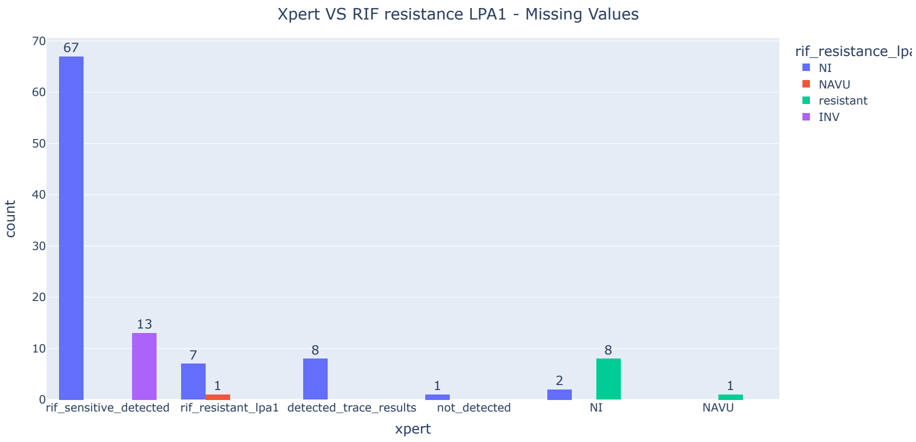
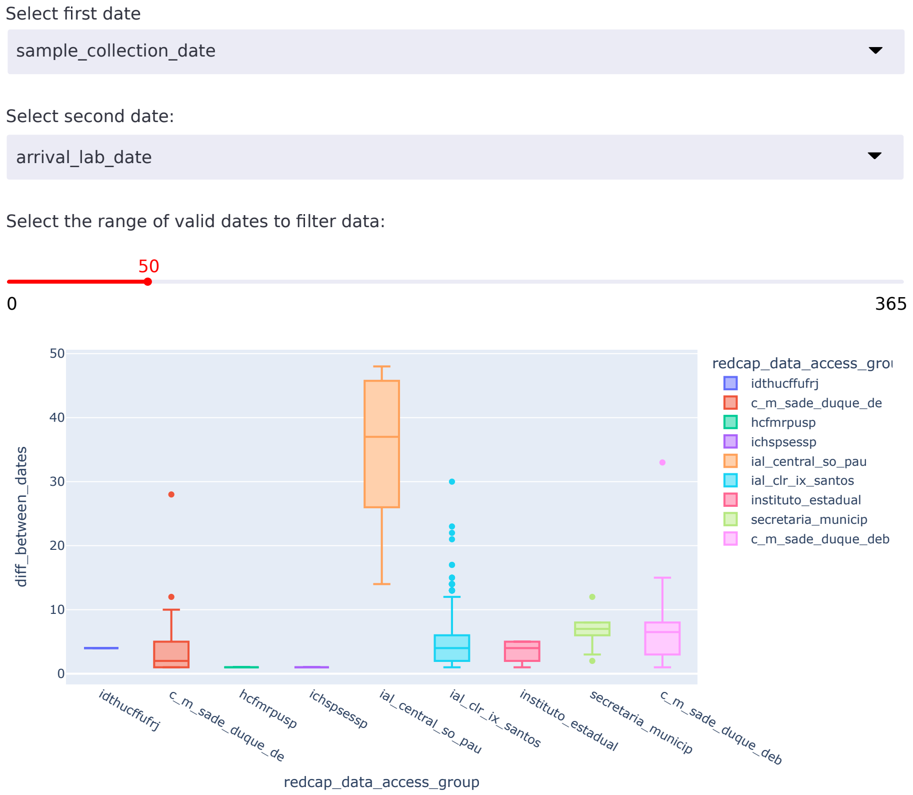


Fig. 3. Comparison between Xpert and LPA1 for missing values.

DOI: 10.1007/978-3-031-36024-4_24

The elapsed time between key dates is generated in a Box Plot chart type to make it easier to identify the mean times and data distribution from all over the centers. For filtering and manipulation, a slide bar is available for the user to configure a threshold to set a minimum value in days between those two dates. Figure 4 presents this functionality in the Box Plot chart.



Select first date

sample_collection_date

Select second date:

arrival_lab_date

Select the range of valid dates to filter data:

50

Fig. 4. Box Plot chart between gathering and release date.

Complementary to the specific analyzes of the test results and time intervals between key dates, a section was also developed for the patient's socioeconomic data. In general, charts represent the frequency of different patient indicators, which can help better understand the samples in the database. Figure 5 shows data regarding height, skin color, weight, and gender.

5 Discussion

During the initial stage, the tool demonstrated highly satisfactory results in meeting all the expected requirements for the LPA project use case. One of the

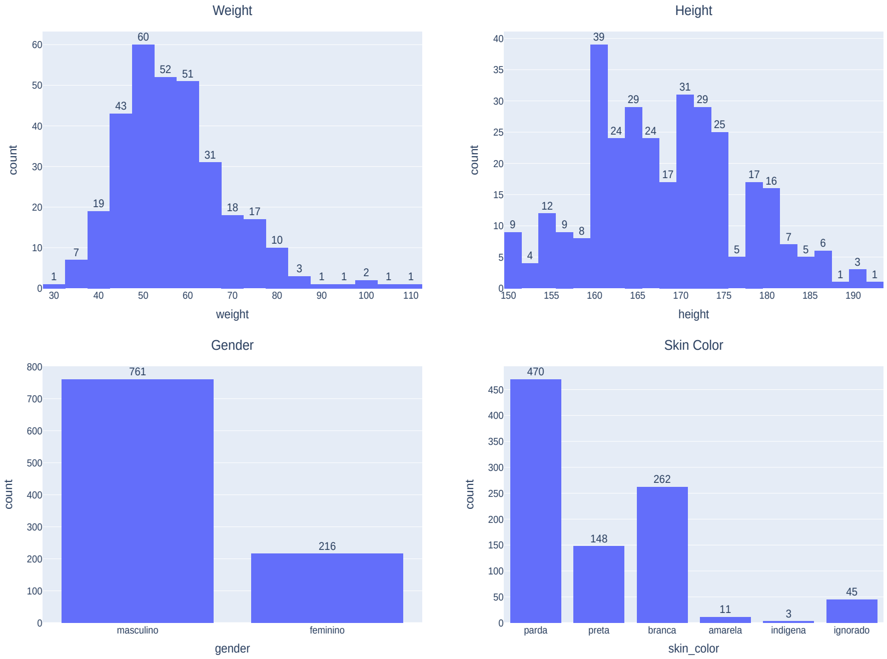


Fig. 5. Bar Charts with weight, height, gender and skin color histograms.

major challenges faced during the development process was to create an analysis of indicators and metrics to assess the accuracy of the tests being evaluated against the gold standard. The web portal provides up-to-date data and easy access to the values required for the statistical analysis and indicators as described on section 3.2.

Before the adoption of the web portal, this process was completely manual, expending a lot of human resources on a repetitive task, resulting in overdue reports delivery. This tool automated this entire process, greatly reducing the report creation time. In a few seconds, the researchers have all the analysis available, always updated with the most recent data from the dataset, with several auxiliary results to improve the study database maintenance and guarantee data quality.

The main contribution of this work is to demonstrate, through a web portal, a pipeline for processing and analyzing clinical data on tuberculosis using statistical methods. The results raise important questions about the need to provide researchers with a tool capable of extracting information from the data collected by their research, presenting them briefly, clearly, and objectively.

The main positive point was the accuracy of developing a tool capable of meeting the needs of the traditional clinical trial use cases. Transitioning from paper records to electronic records and reports is a complex task that requires an advanced information technology infrastructure and a significant investment in human resources. However, the ease provided by such tools used with increasing frequency in clinical research is counterbalanced by limitations in data quality. Successful implementation of an electronic data collection system must be accompanied by a routine review of data and regular feedback to researchers to improve the quality of data captured by the system[18].

The tool's implementation also supports a quick response to data quality tasks by bringing several short-term benefits to all parties involved in the project, such as local and national coordinators, diagnosis laboratories, and researchers. In many studies, researchers perform this type of analysis manually, taking hours or even days to complete, detect, and prevent mistakes.

Although many surveys that use electronic resources provide a dictionary of data from their sources, the units of measurement are often neglected and adopted outside established worldwide standards. Both the form of data collection and data entry impact the expected result of a data set. Extracting information to identify actionable insights by mining clinical documents can help answer quantitative questions derived from structured clinical research data sources[19].

This work demonstrates two main implications of using data quality metrics to lookout TB data in clinical trials. First, the sub-optimal completeness and concordance of the data can make it difficult for health services to make informed decisions in the health facility and present challenges in using the data for TB research. Second, some cases of TB service records were not found in the different screening forms analyzed, suggesting that some reports needed to be transcribed into some computerized systems or were transcribed imprecisely. It may be overcome through an institutional inventory study to understand patient follow-up better.

Although using quantitative research methods is more frequent to assess data accuracy, consistency, completeness, and availability, subjective evaluation can help design effective strategies to improve data quality. Critical and non-critical variables and multiple audits (before, during, and after) with quality improvement feedback should be included in the quality monitoring activities of survey data. This combination is considered a cost-effective solution to visualizing project issues and ensuring data quality.

6 Conclusion

The development of the web portal for statistical analysis delivery on Tuberculosis research has provided a powerful and user-friendly tool for researchers to analyze and interpret clinical data, even in the early stage of development. The portal has been developed to collect data from third-party databases, apply extract-load-transform pipelines, and provide a range of visualizations (such as box plots, bar plots, tables, and data frames) to help researchers to visualize their data and extract meaningful insights.

Going forward, we plan to continue to develop and refine our web portal to attempt the evolving needs of clinical research in the REDE-TB project. Specifically, we aim to enhance the researcher's experience by introducing new functionalities and improving data visualization capabilities. We also plan to create a machine learning section to make available prediction models to identify bad outcomes, resistant TB, or any other future need of the project. Overall, we believe that our web portal has the potential to make a significant impact on the Tuberculosis clinical research community and contribute to the advancement of medical science.

7 Acknowledgments

The study was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) - grant numbers 440758/2018-1, 22411-2019, coordinated by author A.K., and by the São Paulo Research Foundation (FAPESP) - grant number 2020/ 01975-9, coordinated by author D.A.

References

1. Rodrigues, R.J., Gattini, C.H.: Chapter 2 - National Health Information Systems and Health Observatories. In: Global Health Informatics: How Information Technology Can Change Our Lives In A Globalized World, 14-49. Elsevier (2016).
2. Yoshiura, V.T.: Desenvolvimento de um modelo de observatório de saúde baseado na web semântica: o caso da rede de atenção psicossocial. Ph.D. thesis, University of São Paulo, Brazil (2020).

3. World Health Organization - Guide for the establishment of health observatories, https://apps.who.int/iris/handle/10665/246123. Last accessed 03 Mar 2023.
4. Xiao, L., Dasgupta, S.: Chapter 11 - User satisfaction with web portals: An empirical study. In: Web systems design and online consumer behavior, pp. 192-204. Igi Global (2005).
5. Bernardi, F. A., et al.: A proposal for a set of attributes relevant for Web portal data quality: The Brazilian Rare Disease Network case. Procedia Comput. Sci. (in press).
6. Oliveira, D.P.R.: Planejamento estratégico: conceitos, metodologia e práticas. 23nd edn. Editora Atlas, São Paulo (2007).
7. Alves, D., et al.: Mapping, infrastructure, and data analysis for the Brazilian Network of Rare Diseases: protocol for the RARASnet Observational Cohort Study. JMIR Research Protocols 10(1), e24826 (2021).
8. World Health Organization. Regional Office for Africa. Guide for the establishment of health observatories. World Health Organization. Regional Office for Africa. https://apps.who.int/iris/handle/10665/246123. Last accessed 03 Mar 2023.
9. Pereira, B.S., Tomasi, E.: Instrumento de apoio à gestão regional de saúde para monitoramento de indicadores de saúde. Epidemiol. Serv. Saúde 25(2), 411-418 (2016).
10. Nahm, E.S., Preece, J., Resnick, B., Mills, M.E.: Usability of health Web sites for older adults: a preliminary study. CIN: Computers, Informatics, Nursing 22(6), 326-334 (2004).
11. Saeed, M., Ullah, S.: Usability Evaluation of a Health Web Portal. Master's thesis, Blekinge Institute of Technology, MSC-2009:16 Ronneby (2009).
12. World Health Organizatio n (WHO). Global Tuberculosis Report 2018. Geneva: WHO. 2018, https://apps.who.int/iris/handle/10665/274453. Last accessed 03 Mar 2023.
13. Macedo, L.R, Maciel, E.L.N, Struchiner C.J.: Populações vulneráveis e o desfecho dos casos de tuberculose no Brasil. Ciênc saúde coletiva [Internet]. (Ciênc. saúde coletiva) (2021). https://doi.org/https://doi.org/10.1590/1413-812320212610.24132020.
14. Rabahi, M.F, et all. Tuberculosis treatment. Jornal Brasileiro de Pneumologia. (2017). https://doi.org/https://doi.org/10.1590/S1806-37562016000000388
15. Sales, O.M.M, Bentes, P.V.: Tecnologias digitais de informação para a saúde: revisando os padrões de metadados com foco na interoperabilidade. Rev Eletron Comun Inf Inov Saúde, https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1469. Last accessed 03 Mar 2023.
16. Kritski, A. et al.: The role of the Brazilian Tuberculosis Research Network in national and international efforts to eliminate tuberculosis. Jornal Brasileiro Pneumologia, 44:77–81. https://doi.org/https://doi.org/10.1590/s1806-37562017000000435.
17. Bernardi F, et al. Blockchain Based Network for Tuberculosis: A Data Sharing Initiative in Brazil. Stud Health Technol Inform, 262:264-267 https://doi.org/10.3233/SHTI190069.(2019).
18. Sharma A, Ndisha M, Ngari F, et al.: A review of data quality of an electronic tuberculosis surveillance system for case-based reporting in Kenya. Eur J Public Health, 25:1095–1097. https://doi.org/https://doi.org/10.1093/eurpub/ckv092 (2015).
19. Malmasi S, Hosomura N, Chang L-S, et al.: Extracting Healthcare Quality Information from Unstructured Data. AMIA Annu Symp Proc, 1243–1252 (2017).