Knowledge Discovery in Databases: Comorbidities in Tuberculosis Cases

Isabelle Carvalho[1][0000−0002−1384−6166], Mariane Barros Neiva[1][0000−0002−0371−5859], Newton Shydeo Brandão Miyoshi[2][0000−0002−2335−371X], Nathalia Yukie Crepaldi[2][0000−0001−8011−868X], Filipe Andrade Bernardi[2,3][0000−0002−9597−5470], Vinícius Costa Lima[2,3][0000−0002−2467−358X], Ketlin Fabri dos Santos[3][0000−0001−5670−6588], Ana Clara de Andrade Mioto[3][0000−0003−4475−1984], Mariana Tavares Mozini[2][0000−0002−6235−7000], Rafael Mello Galliez[4][0000−0003−0348−8374], Mauro Niskier Sanchez[5][0000−0002−0472−1804], Afrânio Lineu Kritski[4][0000−0002−5900−6007], and Domingos Alves[2][0000−0002−0800−5872]

[1] Institute of Mathematical and Computer Sciences, University of Sao Paulo, 400 Trabalhador São Carlense Avenue, Sao Carlos/SP, Brazil
isabelle.carvalho@alumni.usp.br,marianeneiva@usp.br
[2] Ribeirao Preto Medical School, University of Sao Paulo, 3900 Bandeirantes Avenue, Ribeirao Preto/SP, Brazil
{newton.sbm,nathaliayc,mtmozini}@usp.br; quiron@fmrp.usp.br
[3] São Carlos School of Engineering - University of Sao Paulo, 400 Trabalhador São Carlense Avenue, São Carlos/SP, Brazil
{filipepaulista12,viniciuslima,anaclara.mioto}@usp.br
[4] Pontifical Catholic University of Minas Gerais, 500 Dom Jose Gaspar, 500, Belo Horizonte/MG, Brazil
ketlin.fabri@gmail.com
[5] Faculty of Medicine, Federal University of Rio Janeiro, 373 Carlos Chagas Filho Avenue, Rio de Janeiro/RJ, Brazil
galliez77@ufrj.br, kristskia@gmail.com
[6] School of Health Sciences, University of Brasilia, Campos Univ. Darcy Ribeiro, Brasilia/DF, Brazil
maurosanchez@unb.br

Abstract. Unlike the primary condition under investigation, the term comorbidities define coexisting medical conditions that influence patient care during detection, therapy, and outcome. Tuberculosis continues to be one of the 10 leading causes of death globally. The aim of the study is to present the exploration of classic data mining techniques to find relationships between the outcome of TB cases (cure or death) and the comorbidities presented by the patient. The data are provided by TBWEB and represent TB cases in the territory of the state of São Paulo-Brazil, from 2006 to 2016. Techniques of feature selection and classification models were explored. As shown in the results, it was found high relevance for AIDS and alcoholism as comorbidities in the outcome of TB cases. Although the classifier performance did not present a significant statistical difference, there was a great reduction in the number of attributes and in the number of rules generated, showing, even more, the high rel-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

evance of the attributes: age group, AIDS, and other immunology in the classification of the outcome of TB cases. The explored techniques proved to be promising to support searching for unclear relationships in the TB context, providing, on average, a 73% accuracy in predicting the outcome of the cases according to characteristics that were analyzed.

Keywords: Comorbidity · Data Mining · Knowledge Discovery · Public Health · Tuberculosis

1 Background

Tuberculosis (TB) is an infectious disease caused by a mycobacterium that mainly affects the lungs but can also appear in other organs of the body, such as bones, kidneys, and meninges (membranes that involve the brain). The infection has been treated for years as a public health problem and it has been the focus of many types of research. Besides the efforts, TB is still one of the main causes of death among infectious diseases, being the most important single infectious agent for mortality worldwide [27]. The problem is that TB compromises the patient's immune system, making them more susceptible to other diseases. Moreover, without adequate treatment, it might progress to more serious conditions as well as allowing for the development of drug resistance [8].

As mentioned before, TB is dangerous due to the ability to weaken the immune system, turning the body susceptible to other diseases. The term comorbidities define coexisting medical conditions, distinct from the primary condition under investigation, that influences patient care during detection, therapy, and outcome. The study of comorbidities associated with TB is extremely important to raise hypotheses about the relationship of other nosologies with the disease in order to help prevent and treat these patients [5].

1.1 The TB scenario in Brazil and in the world

Tuberculosis continues to be one of the 10 leading causes of death globally nowadays. Since 1997, the World Health Organization (WHO) has been monitoring tuberculosis cases annually. It is estimated that TB caused 1.5 million deaths in 2020, including 208.000 deaths among HIV-positive people. Most of these cases occurred in emerging or underdeveloped countries. The combination of social and economic factors has contributed greatly to the reduction of these rates since effective treatment for TB already exists [27].

In Brazil, there are approximately 70,000 new TB cases per year, and it is one of the countries most affected by this issue [16]. Annually, the decrease in TB incidence does not exceed 2% while the ideal would be approximately 8%, to reach the goal of eliminating tuberculosis as a public health problem in Brazil [19]. In addition, treatment indicators do not reach 75% of cases depending on the region. Adherence to treatment is important because default contributes to disease transmission. According to the recommendation of the Brazilian Ministry of Health, cure rates below the 85% target and default rates above 5%

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

demonstrate the need to increase the quality of treatment coverage [10, 20]. All these conditions increase the vulnerability of Brazilian patients to be affected by comorbidities in TB and enhance the continuous necessity of research in this area.

1.2 The comorbidities in TB cases

The concern around the disease is even more relevant since TB shares socio-economic determinants with several other diseases. Thus, there are many studies that map the role of comorbidities in patients with tuberculosis [26]. Some of the tuberculosis-related comorbidities include HIV, diabetes, alcoholism, and drug addiction. The probability of developing TB disease is much higher among people infected with HIV and affected by risk factors such as diabetes, smoking, and alcohol consumption [21].

The relation between TB and HIV is one of the most studied topics in this context. Tuberculosis is one of the most common diseases among people with HIV, although the physical pathogenesis of TB acting as an immunosuppressant is not yet established. An HIV patient is 16 to 27 times more likely to develop TB than a person without HIV [12]. In the case of diabetes, the same relationship exists. Diabetes triples the risk of TB. This problem is still greater in emerging countries because of the increasing number of diabetes cases when compared to developed countries [11].

There is also evidence of alcohol use and the increased risk of developing tuberculosis. This risk increases in the case of alcoholic patients and/or people who drink more than 40g of alcohol per day [22]. Patients with TB also have conditions related to smoking. Population with a high smoking rate also presents a higher incidence of tuberculosis since it increases the risk of developing 2 to 3 folds [26]. In the case of drug abuse, users are considered risk groups for TB. This relationship has already been identified in several countries. One of the major problems associated with TB and drug abuse is treatment default [4].

1.3 Objectives

In the attempt to improve the knowledge about the patterns in TB, this article uses the power of data mining techniques to analyze the relationship among comorbidities and the final outcome (cure or death) in cases of tuberculosis.

2 Methods

In order to continue the study and to understand even more the relationship between TB and its related comorbidities, this work uses the power of machine learning (ML) techniques to statistically find the main aspects that can deal to cure or death in a case of TB. To summary the approach applied in the study, Figure 1 presents the general steps of the research.

The data mining experiments were performed in three main steps (shown in Figure 1 as 2a, 2b, and 2c). Each step of Figure 1 is detailed as follows:

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

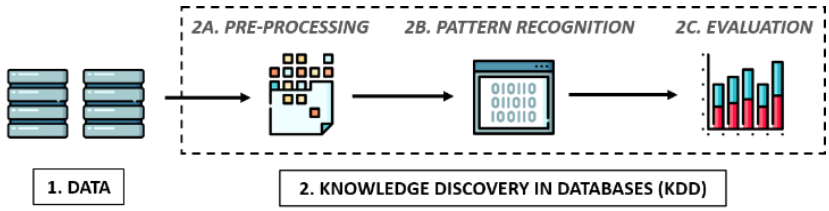

Fig. 1. Overview of methods.

1. DATA: The data used is provided by the Notification and Monitoring System of Tuberculosis of the Health Secretariat - São Paulo State Government (TBWeb) [2]. The database contains confirmed cases of tuberculosis in the territory of the state of São Paulo, from 2006 to 2016. In this step, a descriptive analysis of the data was explored.

2A. The key point of this step is to improve the representation and quality of raw data to provide effective analysis [25]. Two activities were explored in this step: class balancing and feature selection (FS).

1. For class balancing, the method of sub-samples was applied to achieve equality in the number of instances for each class available in the set. This is a prime step in the process of knowledge discovery. The selected examples are the most relevant of the set, this is, the not selected examples are those with repeated or with little information and thus the least possible impairment in the analyzes.
2. In the FS step, two different studies were evaluated:
   i. the use of a filter method that ranks each feature according to a degree of importance within the database computed by the filter algorithm. In the experiments, three different methodologies were applied: Correlation analysis (CFS), Information gain analysis (Gain Ratio), and Chi-square test.
   ii. the use of a wrapper method that involves the selection of attributes based on the classification algorithm to be used [9]. This analysis explored three algorithms: K-Nearest Neighbor (KNN), Bayesian Network (BN), and Decision Trees (DT).

2B. KDD - PATTERN RECOGNITION: This step aims the application of algorithms for the identification of relations in the data and the construction of mathematical models based on these relations [25]. Two analysis were explored in this step:

1. For the FS filter method + rules extraction, three algorithms were applied: C4.5, PART, and RIPPER.
2. For FS wrapper approaches + classification were applied to the same algorithms used for the construction of the relationships: KNN, BN, and DT.

2C. KDD - EVALUATION: The evaluation of the predicted patterns was performed through the accuracy of the exploited algorithms (KNN, BN, and DT). Accuracy presents the proportion of correct predictions.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

This work is the first contact of the machine learning techniques with this dataset, which justifies the choice of classical algorithms and highly explored in the literature as objects of exploration. To initiate the discovery of knowledge in this context, we used a standard parameterization of the algorithms studied and the 10-cross-validation technique, to aid in the generalization of the results [25]. The tools Weka [9], Matlab [15], Python language [24] provided support in activities.

3 Results and Discussion

3.1 Data Characterization

The data set consists of 172,474 TB cases presented through 15 features (8 general features and 7 features about comorbidities). The general features are: id; race/color; age group; sex; are you pregnant?; naturalness; education and type of occupation. The comorbidities features are based on the presence or absence of AIDS, diabetes, alcoholism, mental disorders, drug addiction, other immunology, and tobacco use.

Figure 2 shows the distribution of TB cases with related comorbidities according to the output. We can observe that, proportionally, AIDS, and alcoholism are the two comorbidities that have greater cases of death compared to the other comorbidities.

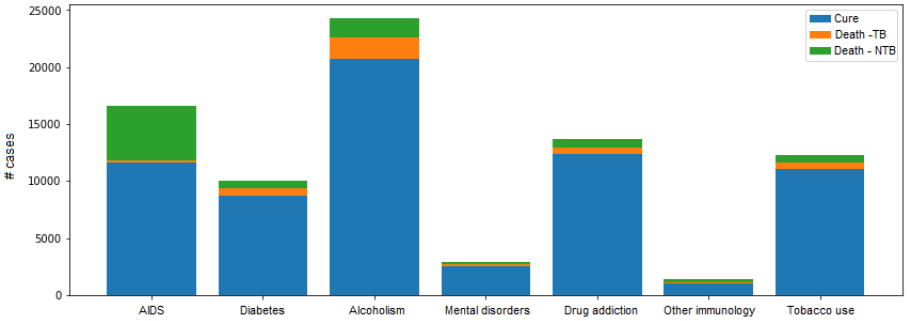

Fig. 2. Distribution of comorbidities in TB cases.

Figure 3 presents the presence or absence of comorbidities in TB cases, over the time studied. The first pattern that can be observed in Figures 3(c), 3(e), and 3(g), starting around 2011, is the increase in the presence of alcohol, drugs, and tobacco, personal habits that influenced the incidence of TB cases. Two main classes are found in the TB database: 156,184 (91%) of the cases are cured and while 16,290 (9%) were deaths (with or without TB as the main cause of death).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

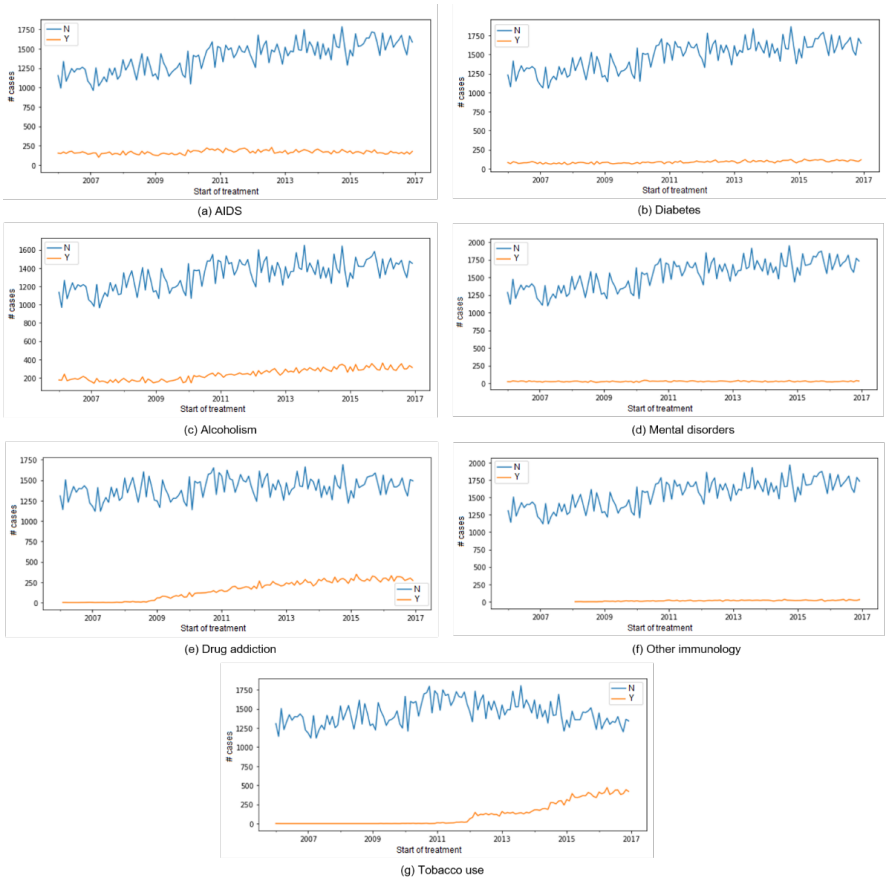

Fig. 3. Distribution of TB cases with comorbidities according to the outcome.

3.2 Pre-processing

As seen from the description of the classes above, the set is highly unbalanced. Therefore, the data set was balanced according to the number of instances of death set, which has lower samples. This resulted in 32,580 records, creating a 50/50 number of samples from each class in the dataset.

Furthermore, as the focus of the study is the relations between comorbidities and outcome of TB cases, four attributes were excluded from the analysis leaving us with 11 attributes composed the set for exploration: race/color, age group, sex, education, AIDS, diabetes, alcoholism, mental disorders, drug addiction, other immunology, and tobacco use. Feature selection results are to be described in the analysis sections below.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

3.3 Analysis: FS Filter Method + Rules Extraction

Table 1 presents the most relevant characteristics selected in each FS method explored. The largest reduction was 73% in the number of attributes, from the original set to match set, i.e., the set with features selected by all methods.

Table 1. Description of selected characteristics - FS filter methods.

Dataset	#	Description
CFS	4	age group, AIDS, mental disorders, other immunology
GainRatio	6	age group, education, AIDS, alcoholism, mental disorders, other immunology
ChiSquare	6	age group, sex, education, AIDS, alcoholism, other immunology
Match set	3	age group, AIDS, other immunology

With the analysis of these ranks, a total of five subsets of features were analyzed in a first experiment (including the original set). Notice that age group, AIDS, and other immunology are presented as main features for all feature selection methods applied, showing the importance of these characteristics. For quantitative analysis, Table 2 presents the performance (accuracy) of the rule extraction models applied to the subsets. As presented in the table, one can see that there is no significant difference between the models, on average, 72% of correct predictions were maintained.

Table 2. Accuracy - Rule extraction models.

Dataset	Classifier		
	C4.5	PART	RIPPER
Original set	73.06%	72.96%	72.70%
CFS	71.76%	71.71%	70.99%
GainRatio	73.00%	73.21%	72.90%
ChiSquare	73.10%	73.16%	72.73%
Match set	71.83%	71.79%	71.35%

Table 3 shows the number of rules generated in each model and subset explored. As the objective is to evaluate the relationships among characteristics, the lower the set of characteristics, without affecting the performance of the classifier, the better the result. With this exploration, on average, there was an 85% reduction in the number of rules generated in relation to the original set.

Although the classifier performance did not present a significant statistical difference, there was a great reduction in the number of attributes and in the number of rules generated, showing, even more, the high relevance of the attributes: age group, AIDS, and other immunology in the classification of the outcome of TB cases.

The finding of the variable AIDS as relevant comorbidity for the outcome of cases TB is in agreement with discussions in the literature that directly correlate

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

Table 3. Number of generated rules.

Dataset	Classifier		
	C4.5	PART	RIPPER
Original set	135	266	23
CFS	22	19	8
GainRatio	57	49	10
ChiSquare	59	59	16
Match set	19	14	5

the two diseases [18]. TB is one of the most frequent opportunistic diseases in the HIV-infected patient and there is evidence that it is one of the main causes associated with death in this population [3].

3.4 Analysis: FS Wrapper approaches + Classification Algorithms

For the second analysis, Table 4 presents the most relevant characteristics selected in each FS wrapper approach. In this case, the largest reduction was 64% in the number of attributes, from the original set to the match set.

Table 4. Description of selected characteristics - FS wrapper approaches.

Dataset	#	Description
1NN	6	age group, education, AIDS, alcoholism, mental disorders, other immunology
BN	5	age group, AIDS, alcoholism, mental disorders, other immunology
DT	8	age group, sex, education, AIDS, diabetes, alcoholism, drug addiction, other immunology
Match set	4	age group, AIDS, alcoholism, other immunology

Once more, there was also no significant difference between the models explored (Table 5), 72% of correct predictions were maintained (on average).

Table 5. Accuracy - Classification models.

Dataset	Classifier		
	1NN	BN	DT
Original set	71.92%	72.27%	73.06%
1NN	73.28%	72.23%	73.00%
BN	72.43%	72.38%	72.38%
DT	73.03%	72.34%	73.09%
Match set	72.50%	72.37%	72.43%

However, in this exploration, the alcoholism appears as comorbidity, not mentioned until then, was evaluated as an attribute of high relevance for the outcome of cases TB. The percentage of TB patients who have problems with alcoholism

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

ranges from 10% to 50% [22, 14]. Furthermore, studies associate alcoholism with therapeutic failure, treatment abandonment, and death due to TB [13].

4 Conclusion

As a first approach of the KDD techniques to the Brazilian context of TB, the analyses performed showed to be promising when assessing the importance of comorbidities in TB cases. The findings are in agreement with recent studies in the health literature on TB, where AIDS and alcoholism comorbidities are being studied as important influencers for the treatment course of TB patients [4].

In future work, we intend to deepen the analyses, investigate the relationships found, investigate the differences among outcomes: death with main cause TB, death without main cause of TB, and treatment default, and also apply other methods of pattern recognition to the data set.

One aspect we intend to apply in this study is to use ontologies to contribute in every step of the KDD process: data selection, data preprocessing, data transformation, data mining, and interpretation and analysis of results. In the data selection step, ontologies can help to have a good understanding of the study domain and the data to be analyzed [1]. Data preprocessing can be done using restrictions and rules embedded in ontologies [7]. Feature engineering can be carried on using ontologies through semantic mapping enriching with additional knowledge [23, 17]. Obtained results can be analyzed using domain-specific ontologies. It is also possible to use ontologies to explain the results obtained from black-box ML algorithms [6].

Acknowledgements

We thank Freepik (www.freepik.com) to provide the icons used in the composition of Figure 1. DA would like to thank the São Paulo Research Foundation for financial support (Process numbers: 2022/00020-0 | 2021/01961 | 2020/01975-9).

References

1. Abhishek, K., Singh, M.: An ontology based decision support for tuberculosis management and control in india. International Journal of Engineering and Technology **8**(6), 2860–2877 (2016)
2. Apunike, A.C., Oliveira-Ciabati, L., Sanches, T.L., de Oliveira, L.L., Sanchez, M.N., Galliez, R.M., Alves, D.: Analyses of public health databases via clinical pathway modelling: Tbweb. In: International Conference on Computational Science. pp. 550–562. Springer (2020)
3. Bastos, S.H., Taminato, M., Fernandes, H., Figueiredo, T.M.R.M.d., Nichiata, L.Y.I., Hino, P.: Sociodemographic and health profile of tb/hiv co-infection in brazil: a systematic review. Revista brasileira de enfermagem **72**(5), 1389–1396 (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

4. Deiss, R.G., Rodwell, T.C., Garfein, R.S.: Tuberculosis and illicit drug use: review and update. Clinical Infectious Diseases **48**(1), 72–82 (2009)
5. Farley, J.F., Harley, C.R., Devine, J.W.: A comparison of comorbidity measurements to predict healthcare expenditures. American Journal of Managed Care **12**(2), 110–118 (2006)
6. Faust, K., Bala, S., van Ommeren, R., Portante, A., Al Qawahmed, R., Djuric, U., Diamandis, P.: Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. Nature Machine Intelligence **1**(7), 316–321 (2019)
7. Geisler, S., Quix, C., Weber, S., Jarke, M.: Ontology-based data quality management for data streams. Journal of Data and Information Quality (JDIQ) **7**(4), 1–34 (2016)
8. Glaziou, P., Floyd, K., Raviglione, M.C.: Global epidemiology of tuberculosis. In: Seminars in respiratory and critical care medicine. pp. 271–285. Thieme Medical Publishers (2018)
9. Johnston, A.H.: Practical machine learning: A beginner's guide to data mining with weka (2018)
10. Kritski, A., Andrade, K.B., Galliez, R.M., Maciel, E.L.N., Cordeiro-Santos, M., Miranda, S.S., Villa, T.S., Ruffino Netto, A., Arakaki-Sanchéz, D., Croda, J.: Tuberculosis: renewed challenge in brazil. Revista da Sociedade Brasileira de Medicina Tropical **51**(1), 2–6 (2018)
11. Lönnroth, K., Roglic, G., Harries, A.D.: Improving tuberculosis prevention and care through addressing the global diabetes epidemic: from evidence to policy and practice. The lancet Diabetes & endocrinology **2**(9), 730–739 (2014)
12. Organization, W.H., et al.: WHO policy on collaborative TB/HIV activities: guidelines for national programmes and other stakeholders. World Health Organization (2012)
13. Pelissari, D.M., Rocha, M.S., Bartholomay, P., Sanchez, M.N., Duarte, E.C., Arakaki-Sanchez, D., Dantas, C.O., Jacobs, M.G., Andrade, K.B., Codenotti, S.B., et al.: Identifying socioeconomic, epidemiological and operational scenarios for tuberculosis control in brazil: an ecological study. BMJ open **8**(6), e018545 (2018)
14. Pereira, J.d.C., Silva, M.R., Costa, R.R.d., Guimarães, M.D.C., Leite, I.C.G.: Profile and follow-up of patients with tuberculosis in a priority city in brazil. Revista de Saúde Pública **49**, 6 (2015)
15. Register, A.H.: A guide to MATLAB object-oriented programming. CRC Press (2007)
16. Reis-Santos, B., Shete, P., Bertolde, A., Sales, C.M., Sanchez, M.N., Arakaki-Sanchez, D., Andrade, K.B., Gomes, M.G.M., Boccia, D., Lienhardt, C., et al.: Tuberculosis in brazil and cash transfer programs: a longitudinal database study of the effect of cash transfer on cure rates. PloS one **14**(2), e0212617 (2019)
17. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: International Semantic Web Conference. pp. 498–514. Springer (2016)
18. Samuels, J.P., Sood, A., Campbell, J.R., Khan, F.A., Johnston, J.C.: Comorbidities and treatment outcomes in multidrug resistant tuberculosis: a systematic review and meta-analysis. Scientific reports **8**(1), 1–13 (2018)
19. Secretariat of Health Surveillance, Department of Surveillance of Communicable Diseases, Ministry of Health: National plan to end tuberculosis as a public health problem. Virtual Health Library of the Brazilian Ministry of Health (2020)
20. Secretariat of Health Surveillance, Department of Surveillance of Communicable Diseases, Ministry of Health: Recommendations Manual For The Control Of Tu-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1

berculosis In Brazil. Virtual Health Library of the Brazilian Ministry of Health (2020)
21. Silva, D.R., Muñoz-Torrico, M., Duarte, R., Galvão, T., Bonini, E.H., Arbex, F.F., Arbex, M.A., Augusto, V.M., Rabahi, M.F., Mello, F.C.d.Q.: Risk factors for tuberculosis: diabetes, smoking, alcohol use, and the use of other drugs. Jornal Brasileiro de Pneumologia **44**(2), 145–152 (2018)
22. Simou, E., Britton, J., Leonardi-Bee, J.: Alcohol consumption and risk of tuberculosis: a systematic review and meta-analysis. The International Journal of Tuberculosis and Lung Disease **22**(11), 1277–1285 (2018)
23. Unbehauen, J., Hellmann, S., Auer, S., Stadler, C.: Knowledge extraction from structured sources. In: Search computing, pp. 34–52. Springer (2012)
24. Van Rossum, G., Drake, F.L.: The python language reference manual. Network Theory Ltd. (2011)
25. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Record **31**(1), 76–77 (2002)
26. World Health Organization and others: Tb comorbidities and risk factors (2019)
27. World Health Organization and others: Global tuberculosis report 2020 (2020)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_1