Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 1373–1379

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022

Clinical Pathways and Hierarchical Clustering for Tuberculosis Treatment Outcome Prediction

Verena Hokino Yamaguti[a]*; Alberto Freitas[c]; Anderson Chidi Apunike[a]; Rui Pedro Charters Lopes Rijo[b]; Domingos Alves[a]; Antonio Ruffino Netto[a]

[a]*University of São Paulo, Ribeirão Preto, Brazil*
[b]*Polytechnic Institute of Leiria, Leiria, Portugal*
[c]*CINTESIS@RISE, FMUP, University of Porto, Porto, Portugal*

Abstract

Clinical pathways are chronological event series that happen throughout a patient's treatment. They can be extracted from the Electronic Health Record medical information and this can be used to correlate the pathway to possible healthcare outcomes. This can be applied to a wide variety of diseases to point pathways related to bad outcomes. These pathways can be audited and patients that start to follow such patterns can be put in special observation and care. Tuberculosis (TB) is one of the leading causes of death through infectious disease and its control is based on search for cases, accurate and premature identification, and treatment. The use of the aforementioned method can help in disease control and premature identifications of bad outcomes for ongoing treatments. Therefore, the current study goals are: 1) identify the existing clinical pathways; 2) group these pathways using hierarchical clustering; 3) create a classification model based on the generated clusters to predict bad outcomes. The dataset used consisted of 277,870 TB treatment cases from the state of São Paulo collected through TBWEB, a information system for monitoring and follow-up of TB cases. All cases with ongoing treatment were excluded from the study and the resulting dataset was splitted in training and test samples. To reduce bias due to imbalance the undersampling technique was applied to the training dataset resulting in a final sample size of 90,184. The test dataset had a size of 52,639 cases. Both datasets had 16 attributes describing the patient diagnosis and drug scheme evolution through the treatment. All attributes unique values were mapped and a representation character was assigned to each one. Later, these representation characters were concatenated in the chronological order of the events and diagnosis creating a representational string for the clinical pathway. The resulting pathways of the training dataset were used to build the clusters which were later used to build the classifier to predict the treatment outcome based on the test dataset clinical pathways. The final model overall accuracy is at 0.829. The model showed a significant improvement of accuracy from

* Corresponding author. Tel.: +55-16-98805-6756; fax: +55-16-3602-1526.
E-mail address: verena.yamaguti@usp.br

1877-0509 © 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022
10.1016/j.procs.2023.01.425

previous studies and had similar or better performance than others in the literature. We believe this model can be implemented to a informational system to further improve treatments management and tuberculosis control.

© 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022

Keywords: Tuberculosis; Clinical pathways; Process mining; Public health; Clustering; Machine Learning.

1. Introduction

Clinical pathways describe a series of clinical events that occur throughout a patient's treatment in a timely manner. These clinical interventions are decided by a medical staff and have specific goals. They are initialized at patient admission and end at patient discharge [1]. Clinical pathways can be designed to deliver a straight and structured healthcare to patients that reduces clinical variation and enhances operational execution and healthcare results [2]. Also, they serve as documentation which can improve overall teamwork and communication [3].

However, to ascertain which clinical pathways relate to a specific outcome, different permutations must be evaluated by a clinical team. This imposes a challenge due to the number of possible permutations in a clinical pathway. Therefore, the clinical pathway design would benefit from a more explicit design based on informational systems [4] using data mining and pattern recognition methods to speed up the process and clinical staff to determine possible pathways.

Clinical pathways can be extracted from the patient's Electronic Health Record (EHR) through the use of data mining methods [5]. EHRs have miscellaneous medical information about the patient profile and clinical events that happened throughout the patient's care. This clinical information can be used to correlate these events to possible healthcare outcomes [6].

The clinical pathway extraction process relies on determining the chronological events and attributing them to a single character which results in a representational string with all the events. These pathways can be visualized as flowchart or Petri's networks [7]. Their visualization provides an overview of most common pathways and enables the risk assessment and prediction through machine learning methods. Pathways that are mostly related to negative outcomes can be audited to verify if the clinical events follow the recommended guidelines and healthcare protocols.

Tuberculosis (TB) is one of the main leading causes of infectious disease death worldwide [8]. This disease control is based on search for cases, accurate and premature identification, and treatment [9]. Therefore, the use of clinical pathways for TB treatment can point pathways related to bad outcomes. These pathways can be audited and patients that start to follow such patterns can be put in special observation and care [10].

The current study goal is to develop a model for multi label classification using clinical pathways to predict the TB outcome. This model will use all clinical events recorded in the patient's EHR throughout the TB treatment. Also, the proposed model uses unsupervised hierarchical clustering for predicting a treatment outcome. The records have all health related events and information about the patient's diagnosis and medication through the treatment.

2. Materials and methods

2.1. Dataset and Software

The initial dataset had 277,870 TB treatment cases from the state of São Paulo. Data was collected through the TBWEB system between 2006 and 2019. TBWEB [11] is a system for notification and tuberculosis treatment follow-up in the state of São Paulo which belongs to the State Health Secretariat of São Paulo. It performs the role of a centralized database for all TB treatments. All data and information used in this study was previously anonymized. The state of São Paulo is one of 27 federal states of Brazil and is located in the southeast region of the country. It has

645 cities and an area of 248,219,491 km^2 [12]. It has the greatest population in the country, with around 45.9 million inhabitants [12]. All code implemented and used in this study was developed in *Python* 3.9.10.

2.2. Study and Preparation of the Data

The first step in this study was to analyze the dataset to understand its structure and variables. The dataset was prepared by selecting the variables of interest and defining the exclusion and inclusion criteria. The inclusion criteria considered treatment only with either a good or bad outcome. Therefore, excluding any ongoing treatment (14,677), which totaled in 263,193 cases. Later, the dataset was split in test (52,639) and training (210,554) sets. To avoid bias from class imbalance we applied the undersampling method to balance good and bad outcomes in the training dataset to 1-1 ratio. The original training dataset had only 21.4% of bad outcomes. The final training set had 90,184 instances.

The initial dataset consisted of 115 attributes which described the patient's demographic information, medication, interventions and clinical diagnosis throughout the treatment. The final dataset used only 16 attributes that described the diagnosis and drug scheme evolution through the patient's treatment. These attributes had been used in a previous study [10]. The order of these attributes is relevant for evaluating the patient treatment evolution in chronological order and is later used to assemble the string that will represent the clinical pathway.

2.3. Setup and Filtering of the Clinical Pathways

In order to determine the clinical pathway for a patient all attributes unique values were mapped and a representation character was assigned to each one. Later, these representation characters were concatenated in the chronological order of the events and diagnosis creating a representational string for the clinical pathway. For this study we considered all clinical pathways that had a good (Cure) or bad outcome (Death, Loss to follow up or Drug resistance). Pathways for ongoing treatments or with a different outcome were filtered and removed.

2.4. Hierarchical clustering

After pathways were identified for the training set, the distance matrix was calculated using the Levenshtein distance [13]. The Levenshtein distance is a relevant string metric for measuring the difference between two-character sequences and is used in various domains such as information retrieval, pattern recognition, error correction, and molecular genetics [14].

The generated distance matrix was used to apply the Weighted Pair Group Method with Arithmetic Mean (WPGMA) [15] clustering hierarchical algorithm and generating a dendrogram from the training set. The cutoff point of the dendrogram was determined by measuring the clusters mean sum of squared distances to their centers using the Elbow-curve method. Fig. 1 shows the Elbow curve on which the optimal k is 7.

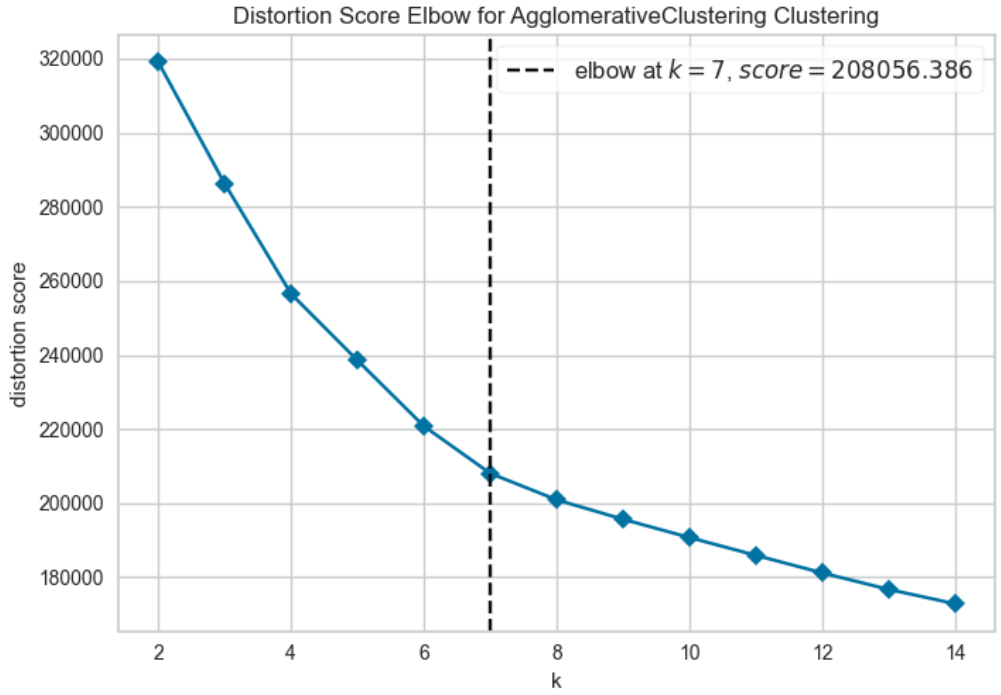

Fig. 1. Elbow curve for clusters between 2 to 14.

2.5. Model classification and performance evaluation

Once all clusters were built, we used the test instance to measure the model performance. Each cluster is evaluated and related to the outcome with the highest probability of the instances that belong to it. The test instances are compared to the clusters by calculating distance to cluster centroid. To simulate a real-world scenario all instances from the test dataset had the representational letter for the treatment outcome removed. Then, for each test instance the closest cluster gives a predicted outcome. To evaluate the model performance, we considered the average Precision, Recall and F1-score.

3. Results

The generated model produced k=7 hierarchical clusters. The uncut dendrogram for the hierarchical clustering is displayed in Fig. 2 in which the different colors denote the similarity between the nodes. Table 1 shows the clusters most related outcomes where it is observable their even distribution among good and bad outcomes. We stress that this even distribution among clusters is due to the undersampling applied to the dataset. Otherwise, we expected to see most clusters representing good outcomes.

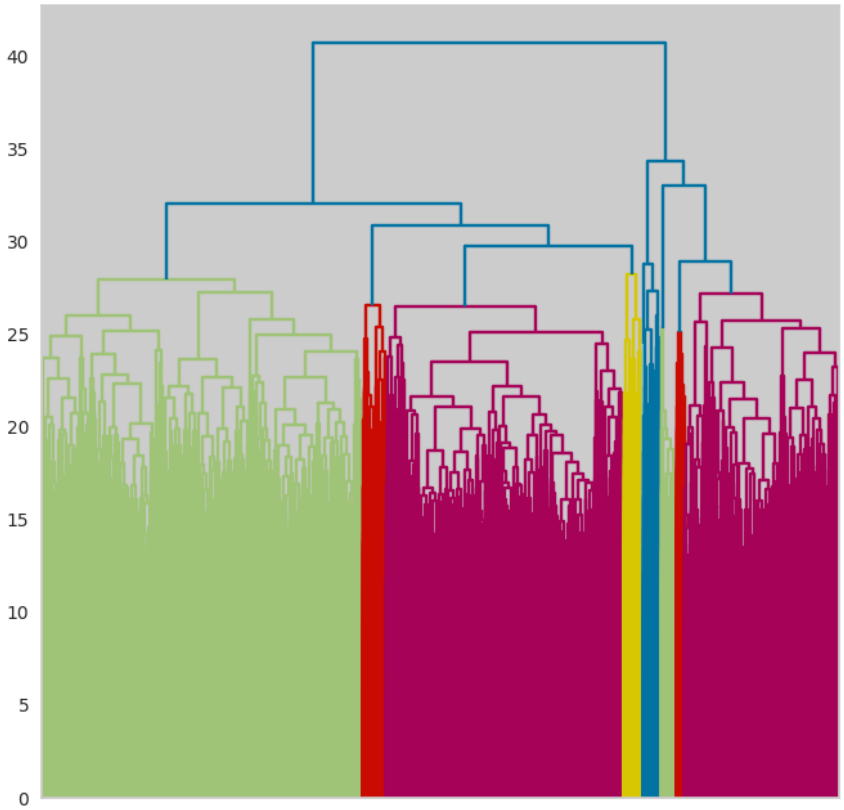

Fig. 2. Uncut dendrogram generated from the hierarchical clustering applied to the training dataset.

After the clusters were generated, we used them as a model for predicting each instance outcome in the test dataset. To predict an outcome for an instance we calculated the closest cluster to that instance pathway and defined the predicted outcome as the cluster most related outcome. Table 2 shows the Precision, Recall and F1-score for good and bad outcomes. We can observe that the model is precise to describe good outcomes rather than bad outcomes showing a precision of 0.883, recall of 0.866 and f1-score of 0.875. The model displays a weighted average precision of 0.832, recall of 0.748, and f1-score of 0.830. The overall model accuracy is at 0.829.

Table 1. Most representative outcome per cluster.

Clusters	Outcome
Cluster 1	Bad
Cluster 2	Good
Cluster 3	Good
Cluster 4	Bad
Cluster 5	Good

Cluster 6	Bad
Cluster 7	Bad

Table 2. Model accuracy per outcome.

Outcome	Precision	Recall	F1-Score
Bad	0.717	0.748	0.732
Good	0.883	0.866	0.875

4. Discussion

The produced clusters allow the prediction of an on-going treatment outcome. If incorporated into a health system this can provide means to alert health care professionals to perform additional actions and efforts to prevent bad outcomes. Additionally, this study enables us to evaluate the most common pathways for good and bad outcomes by analyzing the most representative pathway for each cluster.

Through this study the main problem faced for the prediction of TB treatments was the class imbalance in favor of good outcomes present in the dataset. This imbalance is mostly due to the National Tuberculosis Control Plan enforced by the Brazilian Health Ministry which among other policies implements the Directly Observed Therapy (DOT) which assists the country reducing the number of deaths due to TB and treatment loss to follow-up. In order to reduce this bias, we opted for the undersampling method instead of others because we would still have a significant sample even after applying the undersampling technique.

In general, the present model has improved the accuracy of previous studies [16,17,18] done using a similar dataset. Also, it shows better accuracy when compared to a logistic regression model for TB prediction [19]. We believe this improvement has come due to the temporal analysis provided by the clinical pathways. This allows the model to identify temporal patterns that affect the treatment outcome, something the previous model disregards. Additionally, it provides a more intuitive approach than other predicting models in the literature [20] due to the fact that the generated clusters can be defined according to their more representative clinical pathways. This gives to the involved healthcare professionals a guide to better understand problematic clinical pathways and identifies bad patterns through the patient care.

5. Conclusion

It was possible to build a predictive model of TB treatment outcome through an unsupervised learning model. The generated clusters of the unsupervised model can be used for prediction by calculating the closest cluster for an on-going treatment clinical pathway. Also, the clusters can be used to determine their most common pathway and this can be used to establish a guideline for healthcare professionals [10]. The model presented a significant improvement of accuracy from previous studies and had similar or better performance than others in the literature [19,20]. We aim in future studies to implement the model in an informational system for tuberculosis [21,22,23] treatment and management to help healthcare professionals to manage and monitor TB treatment.

Acknowledgements

We would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and São Paulo Research Foundation (FAPESP) (Grant Nos. 2018/23963-2, 2021/08341-8 and 2020/01975-9).

References

[1] Hunter, Billie and Jeremy Segrott. (2010) “Using a clinical pathway to support normal birth: Impact on practitioner roles and working practices.” *Birth* **37** (**3**): 227–236.

[2] Lin, Fu-ren, Shien-chao Chou, Shung-Mei Pan and Yao-mei Chen. (2001) “Mining time dependency patterns in clinical pathways.” *International Journal of Medical Informatics* **62** (**1**):11-25.

[3] Deneckere, Svin, Martin Euwema, Pieter Van Herck, Cathy Lodewijckx, Massimiliano Panella, Walter Sermeus and Kris Vanhaecht. (2012) “Care pathways lead to better teamwork: Results of a systematic review.” *Social Science and Medicine* **75** (**2**):264–268.

[4] Kempa-Liehr, Andreas, Christina Lin, Randall Britten, Delwyn Armstrong, Jonathan Wallace, Dylan A Mordaunt and Michael O'Sullivan. (2020) “Healthcare pathway discovery and probabilistic machine learning.” *International Journal of Medical Informatics* **137**:104087.

[5] Caron, Filip, Jan Vanthienen, Kris Vanhaecht, Erik Van Limbergen, Jochen Deweerdt and Bart Baesens. (2014) “A process mining-based investigation of adverse events in care processes.” *Health Information Management Journal* **43** (**1**):16–25.

[6] Huang, Zhengxing, Wei Dong, Lei Ji and Huilong Duan. (2016) “Outcome Prediction in Clinical Treatment Processes.” *Journal of Medical Systems* **40** (**1**):8.

[7] Van Der Aalst, Wil. (2011) “Process Mining: Discovery, Conformance and Enhancement of Business Processes.” *Springer*, London.

[8] World Health Organization (WHO). (2021) “Global Tuberculosis Report 2021.” Geneva.

[9] Ministério da Saúde. (2017) “Plano nacional pelo fim da tuberculose como problema de saúde pública.” Brasil.

[10] Apunike, Anderson Chidi., Lívia Maria de Oliveira-Ciabati, Tiago L. M. Sanches, Lariza Laura de Oliveira, Mauro N. Sanchez, Rafael Mello Galliez, and Domingos Alves. (2020) “Analyses of Public Health Databases via Clinical Pathway Modelling: TBWEB.” *International Conference on Computational Science: Computational Science – ICCS 2020* **12140**:550-552.

[11] Galesi, Vera Maria Neder. (2007) “Data on tuberculosis in the state of São Paulo, Brazil.” *Revista de Saúde Pública* **41** (**1**):121.

[12] Instituto Brasileiro de Geografia e Estatística (IBGE). (2020) “Estatísticas do estado de São Paulo.” Available online: <https://www.ibge.gov.br/cidades-e-estados/sp.html>. Access in: 16th may, 2022.

[13] Levenshtein, Vladimir Iossifowitsch. (1966) “Binary codes capable of correcting deletions, insertions, and reversals.” *Soviet Physics-Doklady* **10**(**8**):707–710.

[14] Will, Sebastian, Kristin Reiche, Ivo L Hofacker, Peter F. Stadler and Rolf Backofen. (2007) “Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering.” *PLoS Computational Biology* **3** (**4**): e65.

[15] Sokal, Robert Reuven. (1958) "A statistical method for evaluating systematic relationships." *University of Kansas Science Bulletin* **38**: 1409–1438.

[16] Yamaguti, Verena Hokino, Domingos Alves, Rui Pedro Charters Lopes Rijo, Newton Shydeo Brandão Miyoshi and Antônio Ruffino-Netto. (2020) “Development of CART model for prediction of tuberculosis treatment loss to follow up in the state of São Paulo, Brazil: A case–control study.” *International Journal of Medical Informatics* **141**:104198.

[17] Carvalho, Isabelle, Mariane Barros Neiva, Newton Shydeo Brandão Miyoshi, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Vinícius Costa Lima, Ketlin Fabri dos Santos, Ana Clara de Andrade Mioto, Mariana Tavares Mozini, Rafael Mello Galliez, Mauro Niskier Sanchez, Afrânio Lineu Kritski, and Domingos Alves. (2022) “Knowledge Discovery in Databases: Comorbidities in Tuberculosis Cases.” *International Conference on Computational Science:Computational Science – ICCS 2022* **13352**:3–13.

[18] Da Costa, Luana M.A., Filipe Andrade Bernardi, Tiago Lara Michelin Sanches, Afranio Lineu Kritski, Rafael Mello Galliez, and Domingos Alves. (2021) “Operational modeling for testing diagnostic tools impact on tuberculosis diagnostic cascade: A model design.” *Procedia Computer Science* **181**:650-657.

[19] Silva, Eveline de Almeida, Ulisses Umbelino dos Anjos and Jordana de Almeida Nogueira. (2014) “Modelo preditivo ao abandono do tratamento da tuberculose.” *Saúde em Debate* **38** (**101**):200–209.

[20] Kalhori, Sharareh Rostam Niakan and Xiao-Jun Zeng. (2013) “Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course.” *Journal of Intelligent Learning Systems and Applications* **05** (**03**):184–193.

[21] Pellison, Felipe Carvalho, Rui Pedro Charters Lopes Rijo, Vinícius Costa Lima, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Rafael Mello Galliez, Afranio Lineu Kritski, Kumar Abhishek, and Domingos Alves. (2020) “Data Integration in the Brazilian Public Health System for Tuberculosis: Use of the Semantic Web to Establish Interoperability.” *JMIR Medical Informatics* **8** (**7**):e17176

[22] Lima, Vinícius Costa, Filipe Andrade Bernardi, Michael Domingues, Afranio Lineu Kritski, Rui Pedro Charters Lopes Rijo, and Domingos Alves. (2022) “A computational infrastructure for semantic data integration towards a patient-centered database for Tuberculosis care.” *Procedia Computer Science* **196**:434–438.

[23] Crepaldi, Nathalia Yukie, Vinícius Costa Lima, Filipe Andrade Bernardi, Luiz Ricardo Albano dos Santos, Verena Hokino Yamaguti, Felipe Carvalho Pellison, Tiago Lara Michelin Sanches, Newton Shydeo Brandão Miyoshi, Antonio Ruffino-Netto, Rui Pedro Charters Lopes Rijo, and Domingos Alves. (2019) “SISTB: an ecosystem for monitoring TB.” *Procedia Computer Science* **164**:587–94.