Analyses of public health databases via clinical pathway modelling: TBWEB

Anderson C. Apunike[1][0000−0003−2921−1437], Lívia Oliveira-Ciabati[1][0000−0002−7163−9456], Tiago L. M. Sanches[1][0000−0002−6240−8752], Lariza L. de Oliveira[1][0000−0002−5098−172X], Mauro N. Sanchez[2][0000−0002−0472−1804], Rafael M. Galliez[3][0000−0003−0348−8374], and Domingos Alves[1][0000−0002−0800−5872]

[1] University of São Paulo, Av. Bandeirantes n°3900, Ribeirão Preto – SP, Zip code 14040-900, Brazil
[2] University of Brasilia, Distrito Federal, Brasília – DF, Zip code 70910-900, Brazil
[3] Federal University of Rio de Janeiro, Av. Pedro Calmon n°550, Rio de Janeiro - RJ, Zip code 21941-901, Brazil

Abstract. One of the purposes of public health databases is to serve as repositories for storing information regarding the treatment of patients. TBWEB (TuBerculose WEB) is an epidemiological surveillance system for tuberculosis cases in the state of São Paulo, Brazil. This paper proposes an analysis of the TBWEB database with the use of clinical pathways modelling. Firstly, the database was analysed in order to find the interventions registered on the database. The clinical pathways were obtained from the database by the use of process mining techniques. Similar pathways were grouped into clusters in order to find the most common treatment sequences. Each cluster was characterised and the risk of bad outcomes associated with each cluster was discovered. Some clusters had an association with the risk of negative outcomes. This method can be applied to other databases, serve as a base for decision-making systems and can be used to monitor public health databases.

Keywords: Clinical Pathways · Process Mining · Public health · Clustering

1 Introduction

Over the years, informatics has changed the way data is stored and retrieved. With the progress of informatics, electronic repositories known as databases came into existence. A database is a set of related data that is organised and stored in a way that facilitates access, manipulation and control [10]. Interventions carried out on patients are stored in electronic health records and these records are stored in databases. Each patient follows a treatment sequence over the course of treatment. The interventions that make up the treatment sequence forms the clinical pathway. A clinical pathway is defined as a temporal sequence of clinical interventions established by a specialist or by a multidisciplinary team

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_41

to treat certain patients or reach certain objectives [13]. The development of clinical pathways involves setting up goals, clinical practice revision and the establishment, application and analysis of the protocol created by the multidisciplinary team [6]. Clinical pathways serve as a way of documenting treatment, improving teamwork and facilitating communication [9].

It is possible to extract clinical pathways from electronic health records with the use of process mining techniques [7]. Process mining is defined as the use of event logs to discover, monitor, and improve the processes of an establishment [20]. This technique originated from business management and it has been increasingly applied to health [22]. Examples of the application of process mining can be found in chemotherapy [4] and in patients with Acute Coronary Syndrome [11]. To the best of our knowledge, there are no studies where this technique was applied to study Tuberculosis care. The modelling process of the clinical pathways begins with discovering the chronological sequence of the events of interest. After finding the correct sequence of events, each event is represented by a character or symbol. The order of events are followed and their representative characters or symbols are put together to form a string (or a sequence of symbols) that represents the clinical pathway. To visualize the clinical pathway, diagrams such as networks, flow charts or Petri nets can be used [20].

With the representation of clinical pathways with visual elements, it is possible to have a generalised view of the treatment sequences that exist in the database. Furthermore, risk assessment techniques are able to detect which treatment sequences that are associated with the risk of bad outcomes. Pathways with such characteristics can be identified and avoided in order to improve the outcome of the treatment. In clinical practice, it is common to follow guidelines, protocols or recommendations during treatment. The pathways obtained from process mining can be audited to verify if the treatment sequences follow such recommendations.

The primary objective of this work was to perform secondary analysis on the TBWEB database, which is the database of the system used for tuberculosis epidemiological surveillance in the state of São Paulo, Brazil. This analysis was carried out with the use of the clinical pathways extracted from TBWEB database with process mining. The other objectives were to characterize the clinical pathways and perform a risk assessment on the pathways and detect which pathways are associated with the risk of negative outcomes. The results obtained from this work can serve as a basis for clinical and programmatic decision making regarding the treatment of tuberculosis patients.

2 Methods

2.1 Study dataset

The TBWEB system originated in 2004 and it belongs to the State Health Secretariat of São Paulo State [12]. The system serves as a platform for registering and monitoring tuberculosis cases in São Paulo State [1]. The dataset that was analysed was comprised of tuberculosis cases whose treatment began from 2006 to 2016. There were no age restrictions. The exclusion criteria was latent tuberculosis cases because such cases are monitored in another system [19]. In total, the dataset had 212,569 cases over a ten-year period. The TBWEB dataset was obtained from the Centro de Informação e Informática (CIIS) of Ribeirão Preto Medical School (FMRP) of the University of São Paulo, Ribeirão Preto, Brazil.

2.2 Statistical software

The statistical software developed for clinical pathway modelling and subsequent analyses was written in R language. RStudio version 1.0.136 [3] served as the Integrated Development Environment for writing the codes and executing commands. The following R packages were used over the course of the analyses: Stringr [21] for carrying out operations on strings, openxlsx for saving files in spreadsheets and igraph [8] for drawing networks. Risk assessment was performed with the use of epiR package [18] in order to calculate the relative risk associated with the pathways.

2.3 Analysis plan

The analyses performed with the TBWEB dataset followed a plan of four phases: (1) study and preparation of the data, (2) setup and filtering of the clinical pathways, (3) classification and clustering of the clinical pathways, (4) characterization and analysis of the clinical pathways.

Study and preparation of the data. This phase consists of studying the dataset, understanding its structure and the variables contained in the dataset. The preparation of the data involved choosing the variables of interest and implementing the exclusion criteria on the dataset, if there are any.

In the TBWEB dataset, variables related to treatment regimen, bacilloscopy results, the states of the patient (registered on a monthly basis) and treatment outcomes were chosen to set up the clinical pathways. Records with missing data or with values out of a plausible range were not considered for analyses. More details on the chosen variables are shown on the table below (see Table 1).

Table 1: Description of the selected TBWEB variables.

Type	Variable	Description
Treatment regimen	RHZ	rifampicin + isoniazid + pyrazinamide
Treatment regimen	RHZE	rifampicin + isoniazid + pyrazinamide + ethambutol
Treatment regimen	SZEEt	reptomycin + pyrazinamide + ethambutol + ethionamide
Treatment regimen	MR	Treatment regimen for Multi-Drug Resistant tuberculosis

DOI: 10.1007/978-3-030-50423-6_41

Table 1. Description of the selected TBWEB variables.(continued)

Type	Description	Values and representations
Treatment regimen	OTHERS	None of the above
Treatment regimen	No Info	No information.
Bacilloscopy results/states (BAC)	Positive	Positive BAC result
Bacilloscopy results/states (BAC)	Negative	Negative BAC result
Bacilloscopy results/states (BAC)	Progress	BAC exam in progress
Bacilloscopy results/states (BAC)	No	No BAC test
Bacilloscopy results/states (BAC)	No Info	No information on Bac test
Patient states or Treatment outcome	Default	Patient took medicine for more than 30 days and interrupted treatment for more than 30 consecutive days
Patient states or Treatment outcome	Primary Default	Patient took medicine for less than 30 days and interrupted treatment for more than 30 consecutive days or the diagnosed patient did not start treatment at all
Patient states or Treatment outcome	Inpatient treatment	Patient stays in hospital, receiving 24-hour care
Patient states or Treatment outcome	Outpatient treatment	Patient comes to the hospital, receives treatment and leaves the hospital
Patient states or Treatment outcome	Change diagnosis	Change in treatment regimen due to intolerance or toxicity
Patient states or Treatment outcome	TB Death	Death tuberculosis
Patient states or Treatment outcome	NTB Death	Death by other causes
Patient states or Treatment outcome	Transfer	Patient was transferred to another hospital, state or country.
Patient states or Treatment outcome	Others	Other states (none of the previously mentioned states or outcomes)
Patient states or Treatment outcome	No info	No information (only for patient states)

Setup and filtering of the clinical pathways. In this step, the exact sequence of events and interventions was followed, concatenating characters or

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_41

symbols in order to get the string that represents the clinical pathway. If there are certain patterns of pathways that fall into the exclusion criteria, such pathways were removed from the analysis, thus filtered out from other pathways.

Classification and clustering of the clinical pathways. After discovering the clinical pathways, the next step is to find groups of related pathways. If there is any classification system to classify patients based on their clinical conditions, such a system can be applied to the pathways in order to obtain groups of pathways with related conditions. If there is no classification system based on clinical conditions, the pathways can be clustered directly.

To cluster the pathways, hierarchical clustering methods were taken into consideration because of the way the results are displayed in the form of a tree (or dendrogram). There are other methods of clustering but it was decided to limit the clustering of the pathways to hierarchical clustering methods. The objective of the clustering process was to form subgroups of pathways based on their level of similarity. Hierarchical clustering requires a distance metric in order to group the input data based on their similarity. Due to the fact that the clinical pathways are in the form of strings, two dimensional distance methods like Euclidean or Manhattan distances do not apply. Therefore, the Levenshtein distance [5] is applied to the pathways as a distance metric in this case. This distance method calculates the total number of transformations (insertions, deletions, substitutions) to transform one string to another. For example, the Levenshtein distance between “wafer” and “water” is one because only one transformation is required to switch from “wafer” to “water”. After finding the distances between the pathways, a hierarchical clustering method is applied to the pathways. Clustering methods like average-linkage, complete-linkage or mcquitty can be used. Average-linkage method is a clustering method where the distance between clusters is the mean distance between all the pairs of objects from each cluster [17]. In complete-linkage method, the distance between clusters is the maximum distance between two objects in each cluster [14]. The Mcquitty method is slightly different, as it calculates the distance between clusters based on the average distance of the newly formed cluster with one previously formed [16].

Hierarchical clustering generates a dendrogram that shows the clusters of pathways. In order to find the subgroups (or clusters) of pathways, the optimal number of clusters has to be found and depending on the configuration of the dendrogram, the optimal number of clusters can be hard to discover. The Elbow method [2] was used to find the optimal number of clusters. This method is comprised of four steps:

- Form clusters based on values. This process involves clustering data according to values known as “k”. The value k can vary from 1 to N clusters for example. In the dendrogram, clusters can be obtained by cutting the dendrogram at different heights. Therefore, in order to obtain different clusters according to the value k, the dendrogram was cut in various sections.
- Calculate the within-cluster sum of squares (wss) for each value. The within-cluster sum of squares involves calculating the sum of squares of all the

DOI: 10.1007/978-3-030-50423-6_41

pathways within each cluster. There are N clusters for every value of k. Therefore, Twss value which is the total within-cluster sum of squares for each k is calculated by adding the sum of squares of all the pathways within each cluster (wss). The formula below (1) illustrates the wss formula for k, where k is the number of clusters, N is the maximum number of clusters related to k, Mi and Mj are the number of pathways in clusters Pi and Pj respectively.

$$Twss(k) = \sum_{k=1}^{N} \left(\sum_{i=1}^{Mi} \sum_{j=1}^{Mj} Levenshtein_Distance(Pi, Pj)^2, i \neq j \right) \quad (1)$$

After calculating the within-cluster sum of squares for different numbers of clusters, the next step is to plot a graph of the Twss values against the number of clusters. The optimal number of clusters corresponds to the value that has the “elbow” which is the point where the within-cluster sum of values decreases in a lesser rate compared to previous values.

Characterization and analysis of the clinical pathways. After determining the optimal number of clusters, the final step is to characterize and analyze the clusters of pathways. This is done by calculating the relative risk of each cluster, performing descriptive analysis on all the clusters and finding the representative pathway of each cluster. In this work, relative risk is used to find an association between the clusters of clinical pathways and bad outcomes. Here, bad outcomes are referred to the occurrence of events such as death or default of treatment. In each case, the exposed group is the group that belongs to the cluster in question. The unexposed group refers to all the pathways that belong to other clusters Knowing the risk associated with each cluster, the next step of this phase is to perform a descriptive analysis of the clusters to know the characteristics of the patients that belong to each cluster. This descriptive analysis is centered around demographic variables, comorbidities and variables that describe the clinical condition of the patients. After describing the clusters, the final step is to discover the representative pathways of each cluster. In order to find the representative pathway of a cluster, the adjacency matrix of the cluster has to be found. An adjacency matrix is a matrix which rows and columns are the states or events that make up the pathways and is filled with the total number of transitions from one state or event to another. With the cluster’s adjacency matrix, a graph can be plotted to show all the pathways that form the cluster and the number of times the patients moved from one event to another. The representative pathway is the path with the highest transitions from one state to another, starting from the initial nodes (treatment regimens) to one of the final nodes (treatment outcome).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_41

3 Results and discussions

This work is a pilot study on TBWEB and the analysis plan was applied to the dataset of tuberculosis cases in 2011. The choice to focus on one year and not on the entire dataset was a technical decision, as dealing with more than ten thousand pathways lead to memory issues when calculating the distances between the pathways. So the way out was to choose the yar with the highest number of pathways less than ten thousand. The 2011 dataset was chosen and it had 7321 pathways. The pathways were discovered and clustered with three methods; average-linkage, complete-linkage and mcquitty. Out of the three clustering methods, complete-linkage was chosen to define the clusters because when the elbow method was applied on all the methods, complete-linkage returned the highest number of clusters. This study focused on a hierarchical clustering method that has the highest number of clusters. The figure below (see Fig. 1) shows the dendrogram generated using complete-linkage method. Since complete-method was chosen to define the clusters, the dendrograms obtained from other hierarchical clustering methods are available at the link:
https://tinyurl.com/all-clusterings-tbweb.

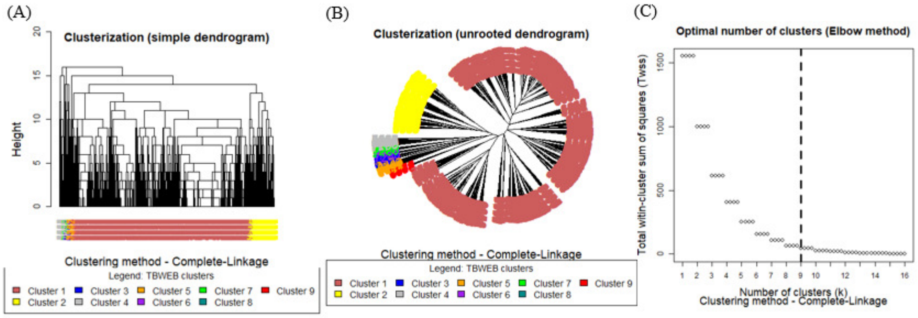

Fig. 1. Clusterization of TBWEB pathways with Complete-Linkage Method. (A) Simple dendrogram. (B) Unrooted dendrogram. (C) Application of Elbow Method to determine the optimal number of clusters.

The figure above illustrates the results obtained while clustering the pathways. According to the elbow method, 9 clusters was the optimal number of clusters. In Figure 2B, the total within-cluster sum of squares (Twss) decreases as the number of clusters increases. In this case, the bend is located at 9 because as from that point the value of Twss drops in a lesser rate compared to previous values. The results of the risk analysis performed on all the clusters are illustrated in the table below (see Table 2) .

Table 2: Risk assessment of the clusters.

Clusters	Number of Pathways	Good outcomes	Bad outcomes	Relative risk value	95% confidence interval
Cluster 1	6037	79 (1.3%)	5958 (98.7%)	0.1	[0.08, 0.13]
Cluster 2	846	45 (5.3%)	801 (94.7%)	0.7	[1.22, 2.29]
Cluster 3	89	7 (7.9%)	82 (92.1%)	2.3	[1.13, 4.80]
Cluster 4	149	5 (3.4%)	144 (96.6%)	1.0	[0.41, 2.34]
Cluster 5	106	98 (92.5%)	8 (7.5%)	25.6	[36.93, 51.47]
Cluster 6	29	0 (0%)	29 (100%)	-	-
Cluster 7	24	2 (8.3%)	22 (91.7%)	2.4	[0.64, 9.26]
Cluster 8	13	11 (84.6%)	2 (15.4%)	25.8	[19.81, 33.52]
Cluster 9	23	4 (14.3%)	24 (85.7%)	4.2	[1.69, 10.54]

Based on the risk analysis results, cluster 1 can be considered as protective, since the pathways contained in it poses a low risk of bad outcomes. On the other hand, the other clusters are associated with varying degrees of clusters, with clusters 5 and 8 having the highest risks of bad outcomes. The table below shows the results of the descriptive analysis performed on the clusters 1, 5 and 8 (see Table 3). The full table with all the clusters can be found at https://tinyurl.com/full-analysis-clusters-tbweb. Table 3 is divided into 14 sections and each section represents the proportions of a group of variables.

Table 3: Descriptive analysis of the clusters.

Attributes	Cluster 1	Cluster 5	Cluster 8
Number of Pathways	6037	106	13
Section 1: Sex			
Male	4196 (69.5%)	89 (84%)	10 (76.9%)
Female	1841 (30.5%)	17 (16%)	3 (23.1%)
Section 2: Age group			
Child (0-9 years)	101 (1.7%)	0 (0%)	0 (0%)
Teenager (10-19 years)	468 (7.8%)	7 (6.6%)	1 (7.7%)
Adult (20-59 years)	4885 (80.9%)	96 (90.5%)	12 (92.3%)
Senior adult (60 years and above)	580 (9.6%)	3 (2.8%)	0 (0%)
No information	3	0	0
Section 3: Ediucation			
No years in school	219 (4.1%)	2 (2.2%)	0 (0%)
1 to 3 years in school	630 (11.9%)	11 (12.1%)	2 (20%)
4 to 7 years in school	1943 (36.7%)	39 (42.9%)	5 (50%)

Table 3. Descriptive analysis of the clusters (continued)

Attributes	Cluster 1	Cluster 5	Cluster 8
8 to 11 years in school	2079 (39.3%)	33 (36.3%)	3 (30%)
12 to 14 years in school	275 (5.2%)	4 (4.4%)	0 (0%)
15 years in school and above	147 (2.8%)	2 (2.2%)	0 (0%)
No information	744	15	3
Section 4: Occupation			
Employed	3497 (62.1%)	67 (68.4%)	8 (61.5%)
Housekeeper	495 (8.8%)	4 (4.1%)	1 (7.7%)
Unemployed	593 (10.5%)	21 (21.4%)	1 (7.7%)
Retired	351 (6.2%)	2 (2%)	0 (0%)
Imprisoned	691 (12.3%)	4 (4.1%)	3 (23.1%)
No information	410	8	0
Section 5: HIV Test			
Positive	408 (6.9%)	15 (14.2%)	0 (0%)
Negative	4900 (82.6%)	76 (71.7%)	12 (92.3%)
HIV test in progress	33 (0.6%)	1 (0.9%)	0 (0%)
No HIV test	591 (10%)	14 (13.2%)	1 (7.7%)
No information	105	0	0
Section 6: AIDS			
AIDS (present)	366 (6.1%)	12 (11.3%)	0 (0%)
AIDS (absent)	5671 (93.9%)	94 (88.7%)	13 (100%)
No information	0	0	0
Section 7: Diabetes			
Diabetic	333 (5.5%)	5 (4.7%)	0 (0%)
Non-diabetic	5704 (94.5%)	101 (95.3%)	13 (100%)
No information	0	0	0
Section 8: Alcoholism			
Alcoholic	743 (12.3%)	27 (25.5%)	2 (15.4%)
Nonalcoholic	5294 (87.7%)	79 (74.5%)	11 (84.6%)
No information	0	0	0
Section 9: Mental disorders			
Mental disorder (present)	89 (1.5%)	0 (0%)	0 (0%)
Mental disorder (abesent)	5948 (98.5%)	106 (100%)	13 (100%)
No information	0	0	0
Section 10: Drug addiction			
Drug addict	495 (8.2%)	21 (19.8%)	4 (30.8%)
Non-drug addict	5542 (91.8%)	85 (80.2%)	9 (69.2%)

Table 3. Descriptive analysis of the clusters (continued)

Attributes	Cluster 1	Cluster 5	Cluster 8
No information	0	0	0
Section 11: Smoking			
Smoker	33 (0.5%)	0 (0%)	0 (0%)
Non-smoker	6004 (99.5%)	106 (100%)	13 (100%)
No information	0	0	0
Section 12: Case Type			
New case	5464 (90.5%)	85 (80.2%)	7 (53.8%)
Relapse	393 (6.5%)	7 (6.6%)	2 (15.4%)
Retreatment	180 (3%)	14 (13.2%)	4 (30.8%)
Section 13: Type of tuberculosis			
Pulmonary tuberculosis	5036 (83.4%)	89 (84%)	13 (100%)
Extrapulmonary tuberculosis	848 (14%)	13 (12.3%)	0 (0%)
Pulmonary and Extrapulmonary tuberculosis	142 (2.4%)	3 (2.8%)	0 (0%)
Disseminated tuberculosis	11 (0.2%)	1 (0.9%)	0 (0%)
Section 14: Pregnancy			
Pregnant	24 (1.3%)	2 (11.8%)	0 (0%)
Not pregnant	1829 (98.7%)	15 (88.2%)	3 (100%)
No information	4184	89	10
Section 15: Outcomes			
Cure	5958 (98.7%)	8 (7.5%)	2 (15.4%)
Default (Abandoned treatment)	57 (0.9%)	97 (91.5%)	11 (84.6%)
Death by tuberculosis	1 (0%)	1 (0.9%)	0 (0%)
Death by other causes	21 (0.3%)	0 (0%)	0 (0%)
No information	0	0	0

The descriptive analysis adds more information about the clusters and enables comparisons between clusters. From the diagram above, it is observed that clusters 5 and 8 have high default rates (table 3, section 14) and this reflected on table 2 where both clusters had the highest risk of bad outcomes (death or default). Cluster 5 has the highest proportion of pregnant women (table 3, section 13) and cluster 8 has the highest proportion of retreatment and relapse (table 3, section 11). This shows that factors such as pregnancy, relapse or retreatment can be considered as risk factors. Drug addiction (table 3, section 9) might be considered as a risk factor due to the that cluster 1 has is among the clusters with the lowest drug addiction proportions. Since clusters 1 and 5 are in stark contrast to one another in terms of risk, the representative pathways of both clusters are shown in the figure below (see figs. 2A and 2B) to have an idea of the most frequent procedures and patient states in the two clusters. Here the representative pathway shows the most travelled path in the cluster of clinical pathways. The representative pathways of all the other clusters can be found at https://tinyurl.com/all-pathways-tbweb.

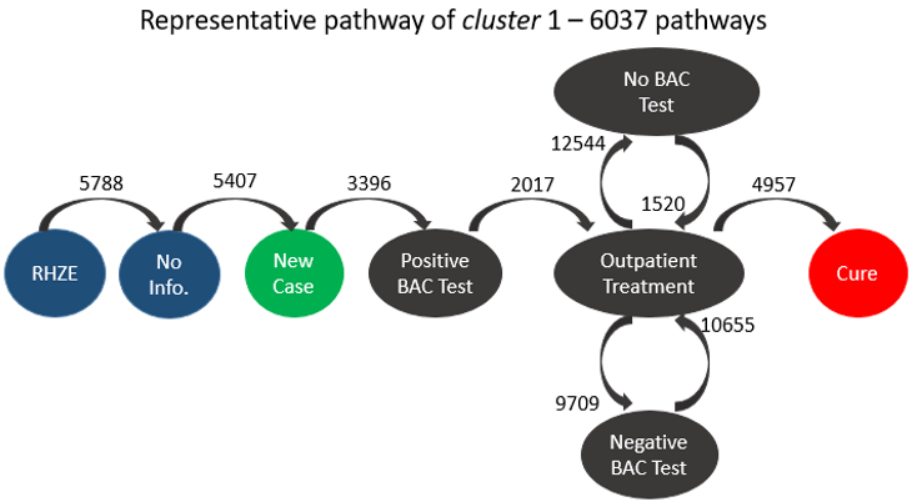

Fig. 2. Representative pathways - Cluster 1, lowest risk.

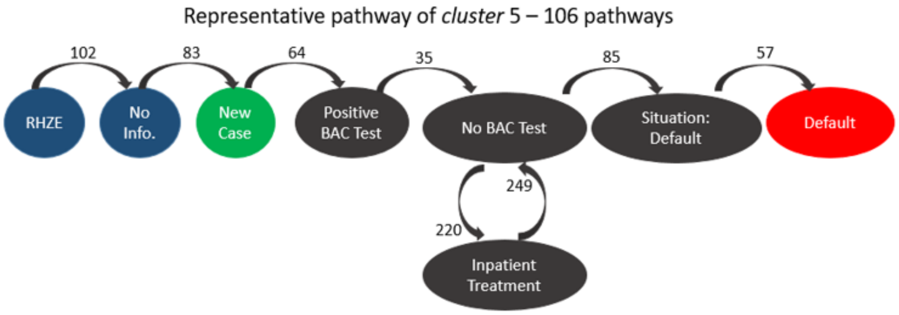

Fig. 3. Representative pathways - Cluster 5, highest risk.

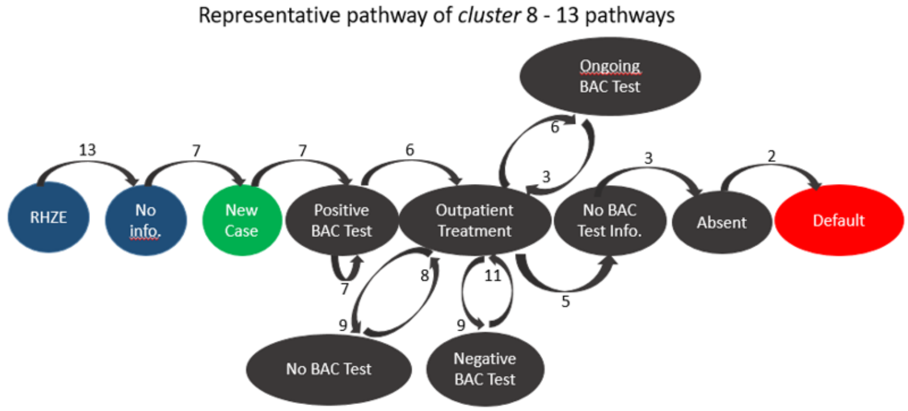

Fig. 4. Representative pathways - Cluster 8, second-highest risk.

The figure above brings some insights on the clinical pathways, for example, on health recommendations. The Brazilian Health Ministry recommends that tuberculosis treatment should commence with the use of RHZE [15] and over the nine representative pathways of the clusters, RHZE was the first treatment regimen. In addition, factors such as patients' presence during treatment and the conduction of bacilloscopy tests can interfere on risk of bad outcomes. In cluster 1 (no risk), there is no patient absence in the representative pathway and although at certain points bacilloscopy exams were not carried out, at some point the exam was performed and it gave a negative result and eventually led to cure. On the other hand, cluster 5 indicates that the most travelled pathway did not have bacilloscopy test after the first positive result and there was absence over the course of treatment which led to default. Cluster 8 which had the second-highest risk also had no baciloscopy tests or no information regarding the test at some points and the representative pathway ended in default. This leads to the belief that carrying out of regular tests, the availability of test results and the patient's continued presence during the course of treatment has an impact in avoiding negative outcomes. This can be confirmed by comparing the structure of the representative pathway of clusters 1 and 5 which poses the lowest and highest risks of negative outcomes respectively.

4 Conclusions

This work has shown that clinical pathway modelling can be used as a method of analyzing public health databases. It was possible to discover the treatments that were registered on the database and their respective outcomes. This was a pilot study and a specific subset of the entire database was selected to perform the analyses. In this work, the whole analyses were conducted on tuberculosis treatments that were registered in 2011. In other words, the whole TBWEB database was not analyzed. The idea of classifying and clustering pathways and finding the representative pathways gives a general view of the different treatment regimens that exist in a public health database. Also, risk analysis helps in measuring the risk associated with a cluster of clinical pathways.

Future applications of this analysis can be in the form of extending this method of analysis to the entire TBWEB database or extend the analyses to other public health databases. Also, this study was limited to hierarchical clustering methods, so future works can focus on repeating this analyses with other clustering methods. In addition, the discovered pathways can be checked if they are in accordance to the real treatment procedure of tuberculosis. This serves as a conformance checking process which is one of the core techniques in process mining. Furthermore, another way of continuing this study is to verify of the pathways influence outcomes. This can be done by excluding the final values regarding the treatment outcome in the pathway strings and repeating the same analyses (clustering and analyzing death rate and/or default rate among clusters). Such a procedure can prove if the clinical pathways influence outcomes. Moreover, depending on the volume of data, alternatives to calculate the distance between the pathways before clustering need be discovered. A divide and conquer strategy, calculating parts of the distance matrix and joining the distances or running the distance calculations on a server instead of a personal computer can be ways to overcome potential Big-Data problems.

Finally, predictive or decision-support systems can be created by combining clinical pathway modelling and artificial intelligence. In this way, the future treatment processes of a patient's treatment can be predicted or the medical team can be alerted if a specific treatment pathway poses a high risk of bad outcome to the patient, allowing the medical team to change the treatment regimen and take new decisions to improve the treatment outcome. Another application of analyzing public health databases with clinical pathway modelling is to perform this analysis periodically on a public health database to monitor the evolution of the risk of bad outcomes on an individual basis, thus adding more value to precision-medicine.

References

1. Cve prof. alexandre vranjac - manual de utilização do tb-web versão 1.6, http://www.saude.sp.gov.br/resources/cve-centro-de-vigilancia-epidemiologica/areas-de-vigilancia/tuberculose/manuais-tecnicos/dvtbc_tbweb_2008.pdf, [Online; accessed 07-February-2020]

2. Datanovia - determining the optimal number of clusters: 3 must know methods, https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods, [Online; accessed 07-February-2020]
3. Il-tb - sistema de informação para notificação das pessoas em tratamento de iltb, http://sitetb.saude.gov.br/iltb
4. Baker, K.e.a.: Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. International Journal of Medical Informatics (103), 32–41 (2017)
5. Bours, P.: Codes capable of correcting bursts of insertions and deletions. Proceedings of 1994 IEEE International Symposium on Information Theory p. 707–710 (1966). https://doi.org/10.1109/isit.1994.394907
6. Campbell, H.e.a.: Integrated care pathways. British Medical Journal (316), 133–137 (1998)
7. Caron, F.e.a.: A process mining-based investigation of adverse events in care processes. Health Information Management Journal (43), 16–25 (2014)
8. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal (Complex Systems) p. 1695 (11 2005)
9. Deneckere, S.e.a.: Care pathways lead to better teamwork: Results of a systematic review. social science and medicine. British Medical Journal (75), 264–268 (2012)
10. Elmasri, R.; Navathe, S.: Sistemas de Banco de Dados. Pearson-Addison-Wesley, São Paulo (2005)
11. Funkner, A.e.a.: Data-driven modeling of clinical pathways using electronic health records. Procedia Computer Science (2017)
12. Galesi, V.: Dados de tuberculose de estado de são paulo. revista de saúde pública. Revista de Saúde Pública (41) (2007)
13. Hunter, B.; Segrott, J.: Using a clinical pathway to support normal birth: Impact on practitioner roles and working practices. Birth (37), 227–236 (2010)
14. Johnson, S.: Hierarchical clustering schemes. Psychometrika **2**, 241–254 (1967)
15. MACIEL, E.: Efeitos adversos causados pelo novo esquema de tratamento da tuberculose preconizado pelo ministério da saúde do brasil. jornal brasileiro de pneumologia. Jornal Brasileiro de Pneumologia p. 232–238 (36 2010)
16. McQuitty, L.: Similarity analysis by reciprocal pairs for discrete and continuous data. Educational and Psychological Measurement **26**, 1695 (11 1966)
17. Sokal, R., M.C.: A statistical method for evaluating systematic relationships. The University of Kansas Science Bulletin p. 1409–1438 (1958)
18. Stevenson, M.e.a.: epir: Tools for the analysis of epidemiological data (2020), https://CRAN.R-project.org/package=epiR, [Online; accessed 7-February-2020]
19. Team, R.: Rstudio manual: Integrated development environment for r (2015), http://www.rstudio.com/
20. Van Der Aalst, W.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, London (2011)
21. Wickham, H.: stringr: Simple, consistent wrappers for common string operations (2020), https://CRAN.R-project.org/package=stringr, [Online; accessed 7-February-2020]
22. Williams, R.e.a.: Process mining in primary care: A literature review. Studies in Health Technology and Informatics (247), 37–380 (2018)